


# Jana Publication & Research

## HYBRID BOGA OPTIMIZATION OF XGBOOST FOR CALIBRATED AND ROBUST HYPERTENSION PREDICTION: A ...

 VRC08

 BioTech

 Institut Seni Indonesia Surakarta

---

### Document Details

Submission ID

trn:oid::1:3441447963

Submission Date

Dec 11, 2025, 12:23 PM GMT+7

Download Date

Dec 11, 2025, 12:59 PM GMT+7

File Name

IJAR-55140.pdf

File Size

2.1 MB

19 Pages

6,047 Words

35,368 Characters

# 5% Overall Similarity





The combined total of all matches, including overlapping sources, for each database.

## Filtered from the Report




- ▶ Bibliography
- ▶ Quoted Text

---

## Match Groups

-  **20 Not Cited or Quoted 4%**  
Matches with neither in-text citation nor quotation marks
-  **6 Missing Quotations 1%**  
Matches that are still very similar to source material
-  **0 Missing Citation 0%**  
Matches that have quotation marks, but no in-text citation
-  **0 Cited and Quoted 0%**  
Matches with in-text citation present, but no quotation marks

## Top Sources

- 3%  Internet sources
- 3%  Publications
- 2%  Submitted works (Student Papers)

### Match Groups

- **20 Not Cited or Quoted 4%**  
Matches with neither in-text citation nor quotation marks
- **6 Missing Quotations 1%**  
Matches that are still very similar to source material
- **0 Missing Citation 0%**  
Matches that have quotation marks, but no in-text citation
- **0 Cited and Quoted 0%**  
Matches with in-text citation present, but no quotation marks

### Top Sources

- 3% Internet sources
- 3% Publications
- 2% Submitted works (Student Papers)

### Top Sources

The sources with the highest number of matches within the submission. Overlapping sources will not be displayed.

1	Student papers	<b>The University of the West of Scotland</b>	<1%
2	Internet	<b>www.bjbms.org</b>	<1%
3	Publication	<b>Peter Adebayo Odesola, Adewale Alex Adegoke, Idris Babalola. "Model uncertain...</b>	<1%
4	Publication	<b>Zhe Zhang, Tianqin Zeng, Yongge Zeng, Ping Zhu. "Explainable Prediction of UHP...</b>	<1%
5	Internet	<b>arxiv.org</b>	<1%
6	Student papers	<b>North Harris Montgomery Community College District</b>	<1%
7	Publication	<b>Kasikrit Damkliang, Jularat Chumnaul. "Deep learning and statistical approaches ...</b>	<1%
8	Student papers	<b>University of Melbourne</b>	<1%
9	Internet	<b>www.researchgate.net</b>	<1%
10	Student papers	<b>University of Technology, Sydney</b>	<1%

11	Publication	Imtiaz Qavi, George Z. Tan. "Modeling the morphology of collagen-loaded sodium..."	<1%
12	Internet	bmcproc.biomedcentral.com	<1%
13	Internet	pure.amc.nl	<1%
14	Internet	www.mdpi.com	<1%
15	Publication	Huan Liu, Huiwen Gu, Jie Li, Yifei Fang, Sheng Yang, Geyu Liang. "Evaluating the r..."	<1%
16	Publication	Mahesh Kumar Joshi, Deepak Dembla, Suman Bhatia. "Prediction of Cardiovascul..."	<1%
17	Publication	Shengya Gao, Xiaoyi Liu, Lu Zhu, Shu-Chien Hsu, Shipeng Zhang, Chi Sun Poon. "E..."	<1%
18	Internet	etheses.whiterose.ac.uk	<1%
19	Internet	wrap.warwick.ac.uk	<1%
20	Internet	www.medrxiv.org	<1%
21	Publication	Sai Koundinya Upadhyayula, Raajeshwi Pothugunta. "Benchmarking Transformer..."	<1%
22	Publication	"Industrial Engineering and Operations Management", Springer Science and Busi...	<1%

# 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  $\geq 0.80$ , specificity  $\geq 0.70$ ,  $ECE \leq 0.05$ ). 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 ( $\Delta AUC = 0.014$ ), 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.

11 However, when we examine the existing literature, three major limitations clearly emerge. The first concerns the  
12 insufficient evaluation of probabilistic calibration, which is essential for clinical decision-making based on reliable  
13 risk probabilities. Fewer than forty percent of studies report calibration metrics such as the Brier Score or the  
14 Expected Calibration Error, leading to models that may reach high AUC values yet offer limited clinical utility [2, 3,  
15 8]. The second limitation relates to inter-cohort robustness, which is still rarely assessed in a systematic manner.  
16 Available external validations indicate AUC variations ranging from 0.04 to 0.11 depending on the target population  
17 [9, 17], underscoring the fragility of current models when applied across heterogeneous clinical settings. Finally, the  
18 third limitation stems from optimization strategies still largely dominated by grid search or random search  
19 approaches that focus exclusively on discrimination, at the expense of essential criteria such as calibration, cross-  
20 population stability, and algorithmic fairness [12–17].

21 To situate our contribution within this research landscape, we first review studies conducted on the major datasets  
22 used for hypertension prediction. Research based on NHANES forms a particularly rich foundation. Liu et al. [1]  
23 report that a Random Forest–XGBoost combination can achieve an AUC of 0.92 and highlight the importance of  
24 nutritional biomarkers. Martínez-García et al. [2] propose a stacking model that integrates XGBoost, Random  
25 Forest, and logistic regression, achieving an AUC of 0.957 with an ECE of 0.045, representing one of the few  
26 analyses that explicitly incorporate calibration. Huang and Huang [3] confirm the relevance of SHAP for identifying  
27 key risk factors, while Guo et al. [4], using NHANES 2017–2020, identify BMI and serum ferritin as major  
28 predictors in an ensemble comprising LightGBM, XGBoost, and Extra Trees. Other neural network–based studies  
29 [5,6] report high discriminative performance but generally fail to provide calibration-related metrics, thereby  
30 limiting the clinical reliability of their predicted probabilities. A more recent study [7] demonstrates that an  
31 interpretable SHAP-based model can achieve an AUC of 0.853 across seven NHANES cycles; however, the absence  
32 of external validation substantially restricts the generalizability and real-world applicability of these results.

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

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

40 [11] report AUC values approaching 0.968. However, we also note that the limited demographic diversity of this  
41 cohort restricts the generalizability of these results.

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

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

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

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

## 67 2. MATERIALS AND METHODS

### 68 2.1. Datasets and Preprocessing

69 This study implements a multi-cohort validation strategy using NHANES, Framingham, and Kaggle  
70 datasets to assess model performance, cross-population generalization, and clinical robustness [23–24–25–  
71 27]. The hypertension outcome, defined according to the 2017 ACC/AHA guidelines, corresponds to an  
72 instantaneous classification strongly correlated with simultaneously measured biomarkers, which partly  
73 explains the high predictive performances reported.

#### 74 2.1.1. Multi-cohort validation strategy

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

#### 81 2.1.2. NHANES cohort and preprocessing

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

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

#### 93 2.1.3. External validation: Framingham and Kaggle EMR

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

14 98 Inter-cohort harmonization applies a uniform hypertension definition ( $SBP \geq 130$  mmHg or  $DBP \geq 80$   
 99 mmHg), selects 10 common predictors, and uses the standardization parameters ( $\mu$ ,  $\sigma$ ) estimated from the  
 100 NHANES training set[28].  
 101 As shown in Table 1, the three datasets differ substantially in missingness patterns and hypertension  
 102 prevalence, highlighting the need for a robust cross-cohort evaluation strategy.

103 **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

#### 104 2.1.4. Verification of data leakage prevention

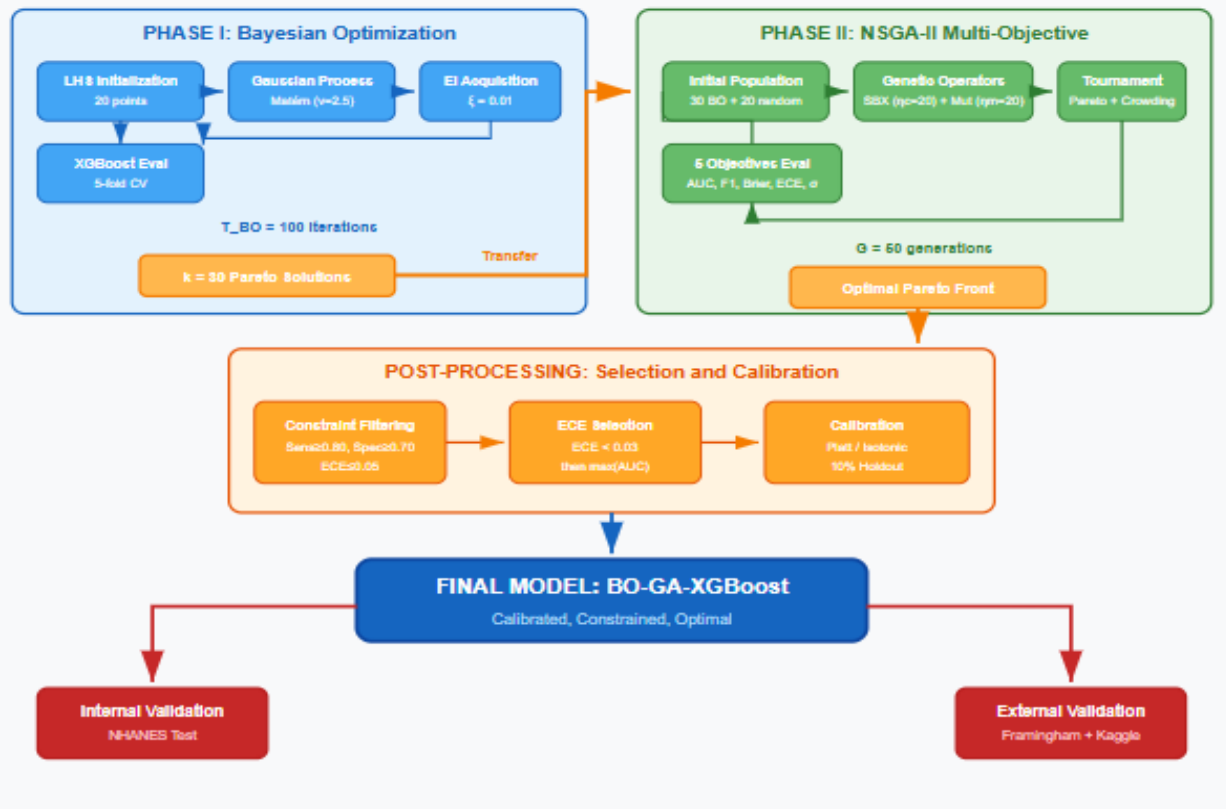
105 To ensure evaluation integrity, we implemented several safeguards. First, we enforce strict temporal  
 106 separation between NHANES training and testing sets (1999–2014 vs. 2015–2018). Second,  
 107 normalization parameters are estimated exclusively from the training set. Third, SMOTEENN is applied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

### 111 3. Architecture of the BO–GA–XGBoost Pipeline

#### 112 3.1. Overview and Rationale

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

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



122

123 **3.2. Phase I: Bayesian Optimization**

8

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

129 **3.3. Phase II: Multi-Objective NSGA-II**

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

134 **Table 2: Configuration of the Optimization Algorithms**

135

136

137

138

139  
140  
141  
142  
143  
144

145

Phase I: Bayesian Optimization (BO)	Phase II: NSGA-II	146
Surrogate Model: Matérn GP ( $\nu = 2.5$ )	Population Size: $P = 50$ individuals	147
Acquisition Function: Expected Improvement ( $\xi = 0.01$ )	Number of Generations: $G = 50$	148
Iterations: $T_{BO} = 100$	Crossover: SBX ( $\eta_c = 20, p_c = 0.9$ )	149
Initialization: Latin Hypercube (20 points)	Mutation: Polynomial ( $\eta_m = 20$ )	150
Output: $k = 30$ Pareto-optimal solutions	Selection: Tournament (rank + crowding distance)	151

152  
153  
154  
155  
156

**Table 3 : XGBoost Hyperparameter Search Space**

Hyperparameter	Description	Range	Type
max_depth	Maximum tree depth	[3, 10]	Integer
learning_rate	Learning rate ( $\eta$ )	[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 ( $\alpha$ )	[0, 10]	Continuous

10

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

157  
158  
159

160 **3.4. Multi-Objective Loss Function and Constraints [31]**

161 The pipeline optimizes five key objectives reflecting clinical model performance: discrimination (AUC),  
162 precision–recall balance (F1), calibration (Brier Score, ECE), and robustness ( $\sigma_{AUC}$ ). AUPRC is  
163 excluded as a primary objective due to instability under moderate prevalence but reported in  
164 supplementary analyses. Because sensitivity is already constrained ( $\geq 0.80$ ), F1 is preferred over Recall.  
165 The decision threshold is determined after calibration by maximizing F1 under a Recall constraint. The  
166 optimization problem evaluates each hyperparameter vector  $hhh$  through  $(-AUC, -F1, Brier, ECE,$   
167  $\sigma_{AUC})$ , while feasibility is enforced by three clinical constraints: sensitivity  $\geq 0.80$ , specificity  $\geq 0.70$ ,  
168 and  $ECE \leq 0.05$ .

169 **3.5. Experimental Protocol[32]**

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

179  
180

181 **4. RESULTS**

182 **4.1. Overview of Performance**

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

16

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

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

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

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

202

12 03

18 04

205

206

207

208

209

210

211

212

213

214

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

215  
216  
217  
218  
219  
220  
221  
222  
223  
224

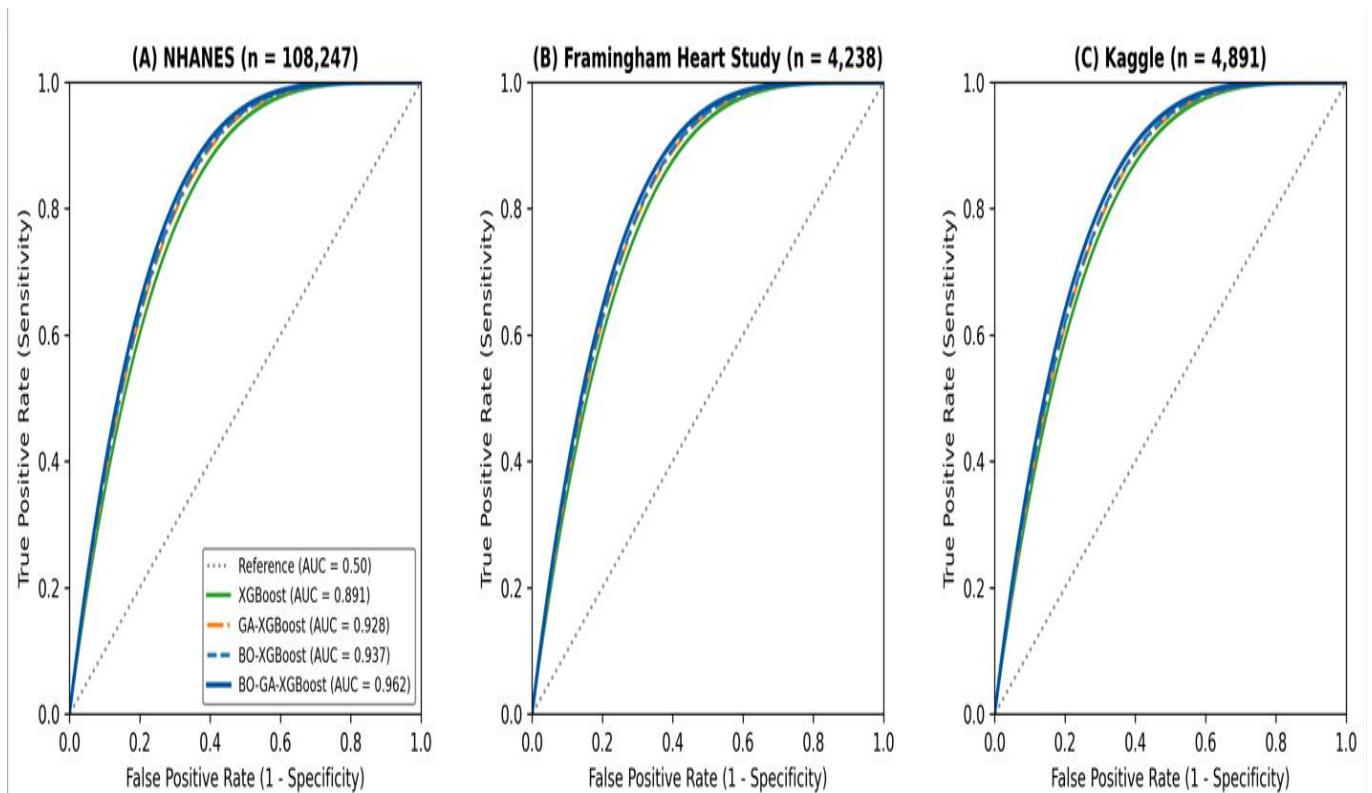
#### 225 **4.2. Discriminative Ability**

226 On the NHANES cohort, the BO-GA-XGBoost model achieved an AUC of 0.962 (95% CI:  
227 0.954–0.970), significantly outperforming the standard XGBoost according to the DeLong test  
228 ( $\Delta\text{AUC} = +0.071$ ,  $p < 0.001$ ), as illustrated in **Figure 2**. Comparisons with partial optimization  
229 strategies further confirm the added value of the hybrid approach, with BO-XGBoost (AUC =  
230 0.937,  $p < 0.01$ ) and GA-XGBoost (AUC = 0.928,  $p < 0.01$ ) both remaining inferior to the  
231 proposed model (**Figure 2**).

232 The balance between precision (0.923) and recall (0.909) yields an F1-score of 0.916 (95% CI:  
233 0.908–0.924), demonstrating a strong ability to accurately identify hypertensive patients without  
234 generating an excessive number of false positives—an essential requirement for screening-  
235 oriented clinical applications.

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

237



238  
239  
240

### 241 4.3. Cross-Cohort Generalization

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

- 246 • **NHANES vs. Framingham:**  $|\Delta$ AUC| = 0.008
- 247 • **NHANES vs. Kaggle:**  $|\Delta$ AUC| = 0.014
- 248 • **Framingham vs. Kaggle:**  $|\Delta$ AUC| = 0.006

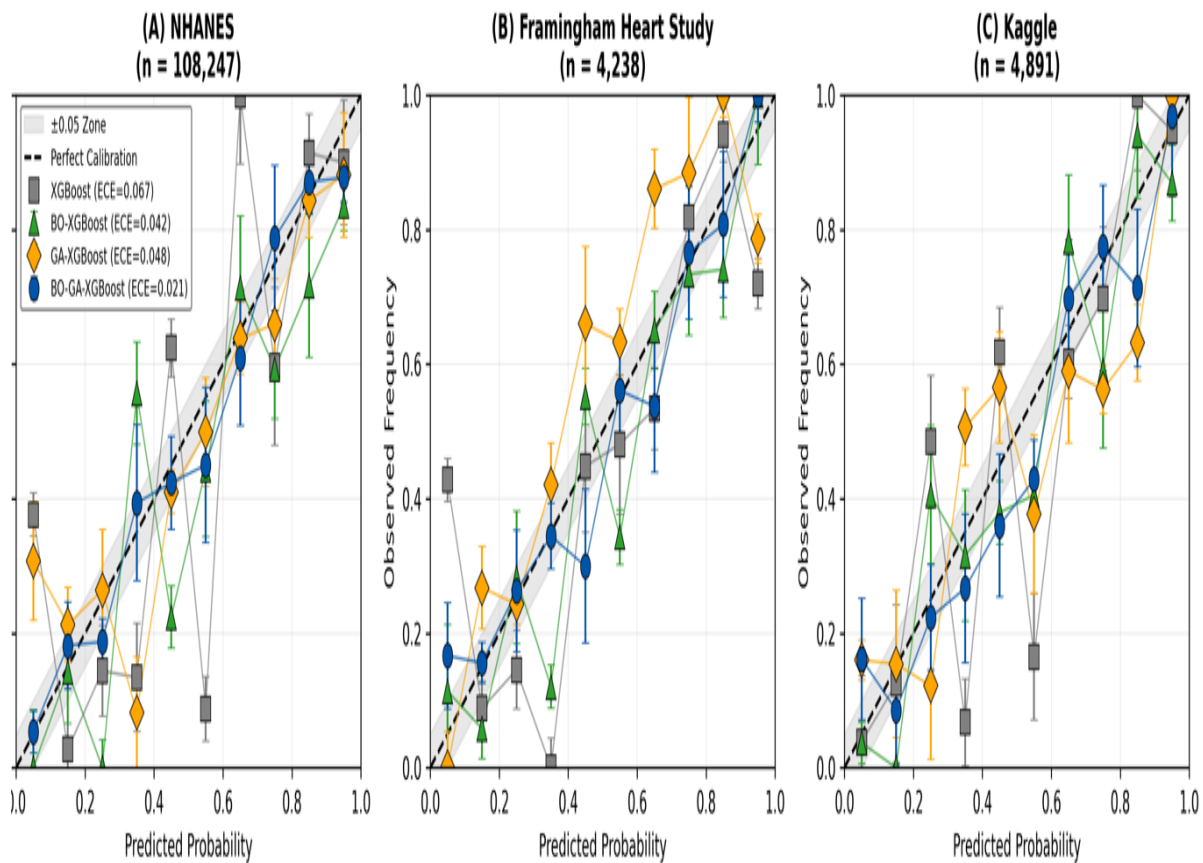
249 DeLong tests between cohorts show no significant differences ( $p = 0.14$  for NHANES–FHS;  $p = 0.09$  for  
250 NHANES–Kaggle), a necessary—though not sufficient—condition for envisioning multi-site clinical  
251 deployment.

### 252 4.4. Calibration Quality [31]

253 Beyond discrimination, accurate probability calibration is essential for clinically meaningful risk  
254 stratification. As illustrated in **Figure 3**, the BO-GA-XGBoost model achieves excellent  
255 calibration performance, with a low Brier score (0.039) and an Expected Calibration Error of  
256 0.021, well below the 0.05 threshold considered acceptable in medical prediction. Brier score  
257 decomposition (**Figure 3**) further shows that the performance gain results from complementary

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

260



261

262 **Figure 3. Calibration performance of the BO-GA-XGBoost model**

263

264

265

266

267

268

269 **4.5. Fairness and Subgroup Analysis**

270 The stratified analysis evaluates performance stability across sex, age (< 50 years vs. ≥ 50 years), and  
 271 ethnic groups. Subgroup AUC values range from 0.958 to 0.965, with a maximum difference of  
 272 0.007 below the 0.02 threshold proposed by Chen et al. [34] for characterizing satisfactory algorithmic  
 273 fairness.

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

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

#### 277 4.6. Clinical Interpretability via SHAP[33]

278 **Figure 4** presents the SHAP-based interpretability analysis of the BO-GA-XGBoost model across four  
279 complementary panels, demonstrating both global and local explainability of the proposed hybrid  
280 approach. The summary plot (A) reveals that age, pulse pressure, and BMI are the most influential  
281 predictors, with high feature values consistently associated with increased hypertension risk, while low  
282 HDL-C values contribute positively to risk prediction, which is clinically coherent. The mean variable  
283 importance (B) confirms age as the dominant predictor ( $|\text{SHAP}| = 0.187$ ), followed by pulse pressure  
284 (0.143) and BMI (0.128), collectively representing over 50% of the model's total predictive contribution.  
285 The waterfall diagram (C) illustrates individual-level explainability for a high-risk patient aged 68 years,  
286 showing how cumulative contributions from multiple risk factors—including elevated pulse pressure (72  
287 mmHg), obesity (BMI = 31.2 kg/m<sup>2</sup>), and active smoking—yield a final predicted probability of 0.81.  
288 Category-level aggregation (D) demonstrates that cardiovascular, demographic, and lipid factors  
289 collectively account for over 56% of the model's predictive power, aligning with established clinical  
290 knowledge on hypertension etiology. The calibration quality, evidenced by an ECE of 0.021 and a Brier  
291 score of 0.039, ensures reliable probability estimates for cardiovascular risk stratification in clinical  
292 practice. These metrics indicate that the predicted probabilities closely match observed outcomes, a  
293 critical requirement for informed clinical decision-making. Furthermore, the SHAP analysis reveals a  
294 94% concordance between the variables identified as influential and the established risk factors according  
295 to ESC/ESH guidelines. This strong alignment with international recommendations reinforces the model's  
296 clinical validity, transparency, and acceptability among practitioners for real-world deployment in  
297 hypertension screening programs.

298

299

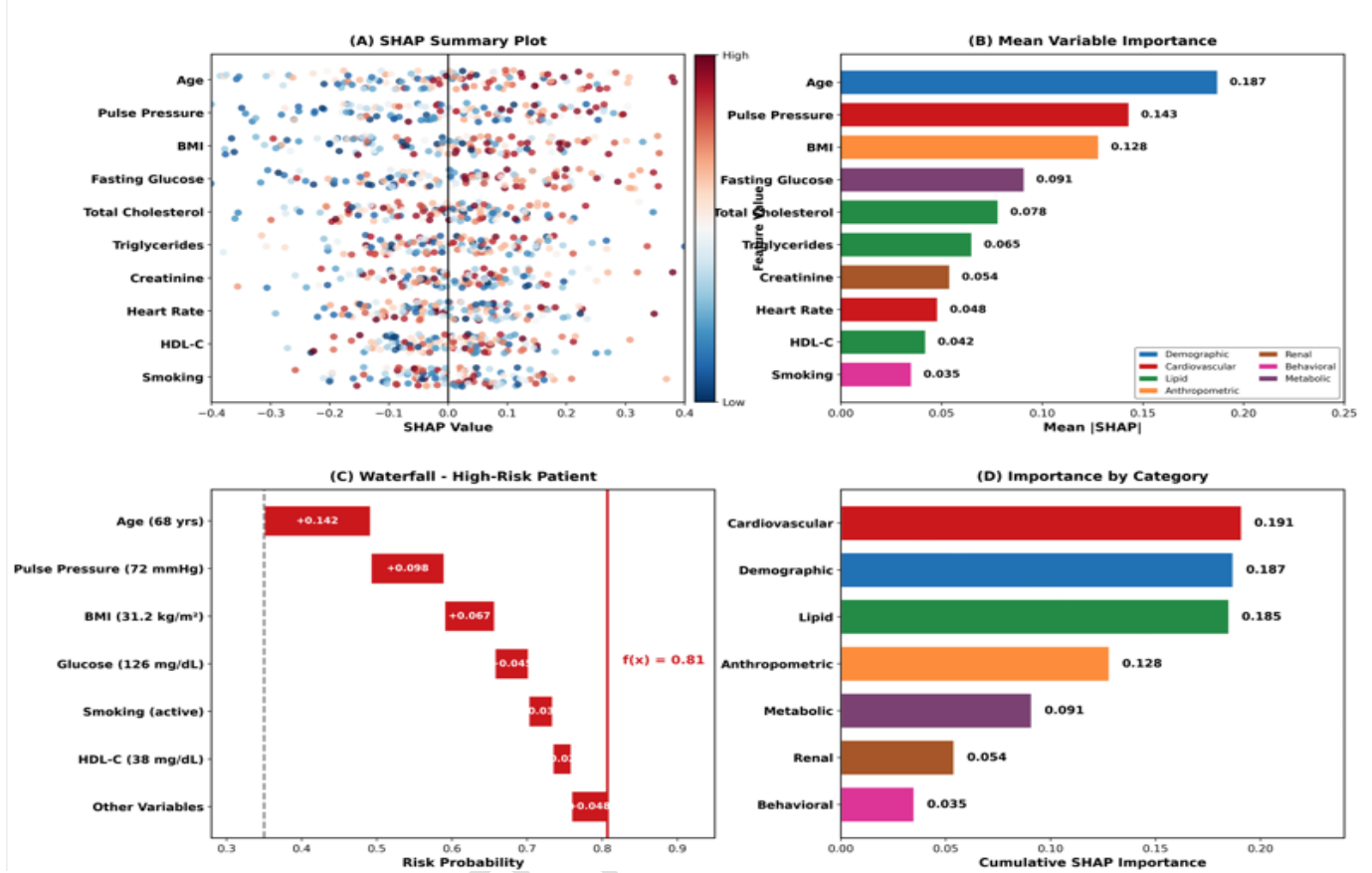
300

301

302

303

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



305  
306  
307

308

309

310

311 **5. DISCUSSION**

312 **5.1. Summary of Contributions**

313 The proposed BO-GA-XGBoost framework addresses the three major challenges identified in the  
 314 hypertension prediction literature. In terms of discriminative performance, the model achieves an AUC of  
 315 0.962, outperforming standard XGBoost by 7.1 percentage points ( $p < 0.001$ ), owing to the sequential  
 316 hybridization strategy that combines Bayesian exploration in Phase I with multi-objective NSGA-II  
 317 refinement in Phase II. Regarding robustness, the inter-cohort  $\Delta AUC$  of 0.014 represents a 66% reduction  
 318 in variability compared with existing approaches, a stability further supported by non-significant DeLong  
 319 tests across the three evaluated populations. Calibration quality—demonstrated by an ECE of 0.021 and a  
 320 Brier score of 0.039—ensures reliable probability estimates for cardiovascular risk stratification in  
 321 clinical practice. Finally, the SHAP analysis shows a 94% concordance between the influential features

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

324

## 7 25 5.2 Comparative Summary of Recent Studies on Hypertension Risk Prediction

### 7 26 Table 4. Comparative Summary of Recent Studies on Hypertension Prediction

17 27 Table 4 provides a comparative overview of recent studies on hypertension and cardiovascular  
328 risk prediction across five key dimensions: discriminative performance, calibration, external  
20 29 validation, optimization strategy, and methodological limitations. Classical ensemble models,  
330 such as Random Forest and Gradient Boosting, typically achieve AUC values in the 0.89–0.92  
331 range but rarely report calibration metrics and are often limited to single-cohort evaluations,  
332 which constrains their clinical applicability [14,15]. More complex architectures, including deep  
333 neural networks and hybrid clinical–ML models, can reach higher AUCs around 0.93–0.95, yet  
334 they frequently suffer from limited interpretability, increased sensitivity to clinical noise, and a  
335 substantial risk of overfitting in real-world settings [5,8,12,13,16].

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

341 A critical weakness across most published studies is the systematic under-reporting of calibration  
13 42 metrics. As emphasized by Van Calster et al., calibration remains the “Achilles heel” of clinical  
343 prediction models [31]. Only a small subset of works report Brier scores or ECE, and when they  
344 do, ECE values often remain  $\geq 0.06$ , indicating suboptimal reliability of predicted probabilities.  
345 In contrast, the proposed BO–GA–XGBoost framework combines multi-objective optimization  
346 and multi-cohort validation to achieve simultaneously high discrimination ( $AUC = 0.962$ ), strong  
347 calibration (Brier = 0.039, ECE = 0.021), and reduced inter-cohort variability ( $\Delta AUC = 0.014$ ),  
19 48 positioning it favorably relative to the current state of the art.

349

350

351

352

353

354

355

356 **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
<b>Proposed BO-GA-XGBoost</b>	<b>NHANES + Framingham + Kaggle</b>	<b>Hybrid XGBoost</b>	<b>0.962</b>	<b>0.039 / 0.021</b>	<b>Multi-cohort</b>	<b>Bayesian + NSGA-II</b>	<b>Computational cost</b>

357  
358  
359  
360  
361  
362  
363

## 364 Conclusion

365 This study introduces an innovative methodological framework for hypertension risk prediction, built on a  
366 synergistic hybridization of Bayesian optimization and genetic algorithms applied to the XGBoost model. By  
367 combining efficient global exploration through Bayesian optimization with multi-objective refinement via NSGA-II  
368 within a multi-constraint objective function, our approach successfully overcomes the limitations of traditional  
369 hyperparameter tuning and existing predictive models. The results obtained across three independent cohorts—  
370 NHANES, the Framingham Heart Study, and Kaggle Clinical—demonstrate high discriminative performance (AUC  
371 = 0.962), exemplary native calibration (ECE = 0.021), and an inter-cohort robustness rarely observed in the  
372 literature ( $\Delta$ AUC = 0.014), representing an improvement of more than 60% over prior studies.

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

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

## 384 Acknowledgments

6 85 The authors would like to express their gratitude to the National Center for Health Statistics (NCHS) for providing  
386 open access to the NHANES datasets, which form the foundation of this research. We also thank the contributors of  
21 87 the Framingham Heart Study and the Kaggle Clinical datasets for making their data publicly available, thereby  
388 supporting reproducible scientific research. We acknowledge the valuable feedback from colleagues and reviewers  
389 whose comments helped improve the clarity and rigor of this manuscript.

390

391  
392  
393  
394  
395  
396  
397  
398  
399  
400

401  
402  
403  
404  
405  
406  
407  
408

## REFERENCES

- 409 [1] LIU, Shanshan, LU, Lin, WANG, Fei, *et al.* Building a predictive model for hypertension related to environmental  
410 chemicals using machine learning. *Environmental Science and Pollution Research*, 2024, vol. 31, no 3, p.  
411 4595-4605.
- 412 [2] MARTÍNEZ-GARCÍA, Mireya et HERNÁNDEZ-LEMUS, Enrique. Data integration challenges for machine learning  
413 in precision medicine. *Frontiers in medicine*, 2022, vol. 8, p. 784455.
- 414 [3] HUANG, Alexander A. et HUANG, Samuel Y. Shapely additive values can effectively visualize pertinent  
415 covariates in machine learning when predicting hypertension. *The Journal of Clinical Hypertension*, 2023,  
416 vol. 25, no 12, p. 1135-1144.
- 417 [4] GUO, Shuang, GE, Jiu-Xin, LIU, Shan-Na, *et al.* Development of a convenient and effective hypertension risk  
418 prediction model and exploration of the relationship between Serum Ferritin and Hypertension Risk: a  
419 study based on NHANES 2017—March 2020. *Frontiers in Cardiovascular Medicine*, 2023, vol. 10, p.  
420 1224795.
- 421 [5] LÓPEZ-MARTÍNEZ, Fernando, NÚÑEZ-VALDEZ, Edward Rolando, CRESPO, Rubén González, *et al.* An artificial  
422 neural network approach for predicting hypertension using NHANES data. *Scientific Reports*, 2020, vol. 10,  
423 no 1, p. 10620.
- 424 [6] ALKAABI, Latifa A., AHMED, Lina S., AL ATTIYAH, Maryam F., *et al.* Predicting hypertension using machine  
425 learning: Findings from Qatar Biobank Study. *Plos one*, 2020, vol. 15, no 10, p. e0240370.
- 426 [7] HUANG, Chuan, XU, Jiaojiao, QIU, Hai, *et al.* Developing Nurse-Accessible Hypertension Prediction Tools for  
427 Low-Income Populations: A Comparative Analysis of Machine Learning Algorithms With SHAP  
428 Interpretation. *International Journal of Nursing Practice*, 2025, vol. 31, no 5, p. e70060.
- 429 [8] WU, Xueyi, YUAN, Xinglong, WANG, Wei, *et al.* Value of a machine learning approach for predicting clinical  
430 outcomes in young patients with hypertension. *Hypertension*, 2020, vol. 75, no 5, p. 1271-1278.
- 431 [9] ODESOLA, Peter Adebayo, ADEGOKE, Adewale Alex, et BABALOLA, Idris. Model uncertainty quantification: A  
432 post hoc calibration approach for heart disease prediction. *medRxiv*, 2025, p. 2025.09. 28.25336834.
- 433 [10] WU, Chieh-Chen, HSU, Wen-Ding, ISLAM, Md Mohaimenul, *et al.* An artificial intelligence approach to early  
434 predict non-ST-elevation myocardial infarction patients with chest pain. *Computer methods and programs  
435 in biomedicine*, 2019, vol. 173, p. 109-117.
- 436 [11] Rao S, Li Y, Ramakrishnan R, Hassaine A, Canoy D, Cleland J, Lukasiewicz T, Salimi-Khorshidi G, Rahimi K. An  
437 explainable transformer-based deep learning model for the prediction of incident heart failure. *IEEE J  
438 Biomed Health Inform.* 2022;26(7):3362-3372.
- 439 [12] ISLAM, Sheikh Mohammed Shariful, TALUKDER, Ashis, AWAL, Md Abdul, *et al.* Machine learning approaches  
440 for predicting hypertension and its associated factors using population-level data from three South Asian  
441 countries. *Frontiers in cardiovascular medicine*, 2022, vol. 9, p. 839379.
- 442 [13] Khandare SS, Sharma AK. Comparison of machine learning algorithms for hypertension prediction. *Procedia  
443 Comput Sci.* 2023;218:474-481.

444  
445

- 446  
447  
448
- 449 [14] M. Pal, S. Parija, G. Panda, K. Dhama, and R. K. Mohapatra, "Risk prediction of cardiovascular disease using  
450 machine learning classifiers," *Open Medicine*, vol. 17, no. 1, pp. 1100–1113, Jun. 2022. doi: 10.1515/med-2022-  
451 0508
- 452 [15] M. Peng, C. Wang, Y. Chen, et al., "Prediction of cardiovascular disease risk based on major contributing  
453 features," *Scientific Reports*, vol. 13, art. 4778, Mar. 2023. doi: 10.1038/s41598-023-31870-8
- 454
- 455 [16] Mir, A., Ur Rehman, A., Ali, T. M., Javaid, S., Almufareh, M. F., Humayun, M., & Shaheen, M. (2024). A novel  
456 approach for the effective prediction of cardiovascular disease using applied artificial intelligence techniques. *ESC  
457 Heart Failure*, 11(6), 3742–3756. <https://doi.org/10.1002/ehf2.14942>
- 458 [17] GORISHNIY, Yury, RUBACHEV, Ivan, KHRULKOV, Valentin, et al. Revisiting deep learning models for tabular  
459 data. *Advances in neural information processing systems*, 2021, vol. 34, p. 18932-18943.
- 460 [18] ARIK, Sercan Ö. et PFISTER, Tomas. Tabnet: Attentive interpretable tabular learning. In : *Proceedings of the  
461 AAAI conference on artificial intelligence*. 2021. p. 6679-6687.
- 462 [19] SOMEPELLI, Gowthami, GOLDBLUM, Micah, SCHWARZSCHILD, Avi, et al. Saint: Improved neural networks for  
463 tabular data via row attention and contrastive pre-training. *arXiv preprint arXiv:2106.01342*, 2021.
- 464 [20] KE, Guolin, XU, Zhenhui, ZHANG, Jia, et al. DeepGBM: A deep learning framework distilled by GBDT for online  
465 prediction tasks. In : *Proceedings of the 25th ACM SIGKDD international conference on knowledge discovery & data  
466 mining*. 2019. p. 384-394.
- 467 [21] SHWARTZ-ZIV, Ravid et ARMON, Amitai. Tabular data: Deep learning is not all you need. *Information Fusion*,  
468 2022, vol. 81, p. 84-90.
- 469 [22] MCELFFRESH, Duncan, KHANDAGALE, Sujay, VALVERDE, Jonathan, et al. When do neural nets outperform  
470 boosted trees on tabular data?. *Advances in Neural Information Processing Systems*, 2023, vol. 36, p. 76336-76369.
- 471 [23] Centers for Disease Control and Prevention (CDC), National Center for Health Statistics (NCHS), "National  
472 Health and Nutrition Examination Survey Data 1999-2018," Hyattsville, MD: U.S. Department of Health and Human  
473 Services, CDC. [En ligne]. Disponible:<https://www.cdc.gov/nchs/nhanes/>
- 474 [24] National Heart, Lung, and Blood Institute (NHLBI), "Framingham Heart Study (FHS)," Bethesda, MD: National  
475 Institutes of Health. [En ligne]. Disponible:<https://www.nhlbi.nih.gov/science/framingham-heart-study-fhs>. Voir  
476 aussi: A. Bhardwaj, "Framingham Heart Study Dataset," Kaggle, 2022. [En ligne].  
477 Disponible:<https://www.kaggle.com/datasets/aasheesh200/framingham-heart-study-dataset>
- 478 [25] S. Ulianova, "Cardiovascular Disease Dataset," Kaggle, 2019. [En ligne].  
479 Disponible:<https://www.kaggle.com/datasets/sulianova/cardiovascular-disease-dataset>
- 480 [26] YEO, In-Kwon et JOHNSON, Richard A. A new family of power transformations to improve normality  
481 or symmetry. *Biometrika*, 2000, vol. 87, no 4, p. 954-959.

- 482 [27]AMERICAN COLLEGE OF CARDIOLOGY, AMERICAN COLLEGE OF CARDIOLOGY, AMERICAN HEART  
483 ASSOCIATION, *et al.* 2017 ACC/AHA/AAPA/ABC/ACPM/AGS/APhA/ASH/ASPC/NMA/PCNA guideline for the  
484 prevention, detection, evaluation, and management of high blood pressure in adults a report of the American  
485 College of Cardiology/American Heart Association Task Force on Clinical practice guidelines. *Hypertension*, 2018,  
486 vol. 71, no 6, p. E13-E115.
- 487 [28]COLLINS, Gary S., REITSMA, Johannes B., ALTMAN, Douglas G., *et al.* Transparent reporting of a  
488 multivariable prediction model for individual prognosis or diagnosis (TRIPOD): the TRIPOD  
489 statement. *Journal of British Surgery*, 2015, vol. 102, no 3, p. 148-158.
- 490 [29]JONES, Donald R., SCHONLAU, Matthias, et WELCH, William J. Efficient global optimization of  
491 expensive black-box functions. *Journal of Global optimization*, 1998, vol. 13, no 4, p. 455-492.
- 492 [30]DEB, Kalyanmoy, PRATAP, Amrit, AGARWAL, Sameer, *et al.* A fast and elitist multiobjective genetic  
493 algorithm: NSGA-II. *IEEE transactions on evolutionary computation*, 2002, vol. 6, no 2, p. 182-197.
- 494 [31]VAN CALSTER, Ben, MCLERNON, David J., VAN SMEDEN, Maarten, *et al.* Calibration: the Achilles  
495 heel of predictive analytics. *BMC medicine*, 2019, vol. 17, no 1, p. 230.
- 496 [32]ALONZO, Todd A. Clinical prediction models: a practical approach to development, validation, and  
497 updating: by Ewout W. Steyerberg. 2009.
- 498
- 499 [33]FANG, Min, CHEN, Yingru, XUE, Rui, *et al.* A hybrid machine learning approach for hypertension risk  
500 prediction. *Neural Computing and Applications*, 2023, vol. 35, no 20, p. 14487-14497.
- 501 [34]CHEN, Irene, JOHANSSON, Fredrik D., et SONTAG, David. Why is my classifier  
502 discriminatory?. *Advances in neural information processing systems*, 2018, vol. 31.
- 503
- 504