Dataset-Specific Bootstrap-Stability Weighting for Calibrated and Clinically Useful Ensemble Prediction in Medical Diagnosis

Pratap Bodduna[1], P. Pranay[1,*], Sravanthi Galipelli[2] and Syam Sundar Junapudi[2]

[1]*Department of Mathematics and Statistics, Chaitanya (Deemed to be University), Hanamkonda, India*

[2]*Department of Community Medicine, Govt. Medical College, Mahabubabad, India*

Abstract: *Background:* Ensemble machine-learning models often perform well within a single medical dataset yet lose discrimination, calibration, and decision usefulness under dataset shift.

Objective: To develop and evaluate Bootstrap-Guided Optimization System (BOOTMED), a bootstrap-guided framework that learns dataset-specific weights from resampling stability to fuse probabilistic predictions, targeting discrimination, calibration, and decision-analytic utility simultaneously.

Methods: Four heterogeneous UCI medical datasets were analyzed (Chronic Kidney Disease; CKD, diabetes, heart disease, breast cancer). Base learners were k-nearest neighbors, random forest (RF), Gaussian naïve Bayes, and complement naïve Bayes. BOOTMED estimated stability-derived weights over 500 bootstrap resamples and aggregated model probabilities. Performance was compared with equal-weight voting and stacking using balanced accuracy and ROC-AUC, calibration error (Brier/ECE), and decision curve analysis.

Results: BOOTMED outperformed equal-weight voting and the best single model across all datasets, improving balanced accuracy by approximately 0.7-2.3 percentage points (adjusted $p<0.05$). Calibration error decreased (lower Brier/ECE), and decision curve analysis showed consistent positive net benefit across clinically relevant thresholds (0.10-0.50). Transferring weights between datasets reduced performance, supporting dataset-specific optimization.

Conclusion: Bootstrap-guided, dataset-specific weighting can improve discrimination, calibration, and clinical net benefit across heterogeneous medical datasets, offering a simple and reproducible ensembling strategy for diagnostic prediction.

Keywords: Bootstrap, ensemble learning, balanced accuracy, calibration, decision curve analysis, clinical utility.

INTRODUCTION

Ensemble machine-learning methods are widely used in medical diagnostics to offset the weaknesses of individual classifiers and to improve robustness across noisy clinical features [1-6]. Yet models that excel within one dataset often lose discrimination, calibration, and clinical usefulness when moved to another setting with different prevalence, feature scales, or class imbalance. This gapstrong single-domain performance but weak transferlimits adoption in decision support, where probabilities must be both accurate and actionable.

Recent studies report high accuracy for chronic kidney disease (CKD), diabetes, heart disease, and breast cancer using bagging, boosting, or stacking on standard benchmarks. It has been emphasized that BOOTMED does not create an expanded set of bootstrap-trained predictors whose outputs are averaged. Instead, BOOTMED uses bootstrap resampling as a stability-estimation step: resamples are used to quantify how consistently each *fixed* base learner performs under sampling perturbations, and these stability summaries are then converted into dataset-specific weights for probability fusion. Bagging is present only within the Random Forest(RF) base learner by design, whereas BOOTMED's contribution is the bootstrap-derived weighting rule for cross-model probability aggregation [1-6]. CKD work is especially abundant, with ensembles and deep models routinely exceeding 90-95% accuracy on the UCI dataset [2,6-12]. Comparable gains are reported for diabetes cohorts [1,5], multi-source heart disease datasets [3], and the Wisconsin breast cancer dataset [4]. However, most reports optimize and validate within a single domain, seldom test cross-dataset portability, and often foreground accuracy or AUC while underreporting probability calibration and decision-analytic value (e.g., Brier score, expected calibration error, decision curve analysis) [1-6,13-16]. Multiplicity control and effect-size reporting are also inconsistent, making it hard to judge the practical meaning of reported *p*-values.

Preprocessing sensitivity further complicates comparisons. Small choicesimputation, scaling, outlier handling, or encodingcan materially shift k-nearest neighbors and tree ensembles, especially on compact clinical tables with mixed types [2,5,9,13,15]. Many pipelines also risk information leakage if transformations are fit on the full dataset instead of within cross-validation folds. Finally, while heterogeneity across diseases is expected, few studies quantify it formally or assess whether weights learned for one domain transfer to another; meta-analytic measures (e.g., I^2) are rarely applied to model deltas across datasets [13-16].

*Address correspondence to this author at the Department of Mathematics and Statistics, Chaitanya (Deemed to be University), Hanamkonda, India;
E-mail: pratapphd@yahoo.com

This work addresses these gaps with BOOTMED, a bootstrap-guided framework that learns dataset-specific ensemble weights from stability signals. BOOTMED aggregates probabilistic outputs from complementary base learnersk-nearest neighbors, RF, Gaussian naïve Bayes, and complement naïve Bayesusing weights estimated from repeated bootstrap resamples. We evaluate BOOTMED on four heterogeneous UCI datasetsCKD (n=400, 24 features), PIMA diabetes (n=768, 8 features), Cleveland heart disease (n=303, 13 features), and Wisconsin breast cancer (n=569, 30 features)and compare it against equal-weight voting and stacking baselines. The study emphasizes (i) generalizability across distinct clinical domains, (ii) statistical rigor with paired tests, multiplicity adjustment, and effect sizes with confidence intervals, and (iii) clinical translation via comprehensive calibration assessment and decision-curve analysis [13-16]. We also examine the portability of learned weights across datasets and summarize cross-domain variability using meta-analytic heterogeneity, clarifying when re-optimization is necessary.

METHODOLOGY

Overview of the Bootmed Framework

We propose a BOOTMED to validate and generalize medical diagnostic classifiers across heterogeneous datasets. The workflow (Figure **1**) integrates: (i) dataset harmonization and leakage-safe preprocessing, (ii) independent training of heterogeneous base learners, (iii) bootstrap-based stability estimation to derive dataset-specific probability weights, and (iv) comprehensive evaluation of discrimination, calibration, and clinical utility. The central design goal is to improve robustness (performance stability under resampling) while retaining clinical interpretability by expressing the final decision as a weighted aggregation of calibrated probabilities rather than a black-box meta-learner.

Datasets and Study Setting

We evaluated four binary classification datasets sourced from the UCI Machine Learning Repository, selected to represent controlled heterogeneity across clinical domains and data characteristics:

- CKD: $n = 400,\ p = 24$
- PIMA Diabetes: $n = 768,\ p = 8$
- Heart Disease (Cleveland): $n = 303,\ p = 13$
- Breast Cancer Wisconsin: $n = 569,\ p = 30$

Class prevalence in the source datasets was verified after preprocessing. Where any

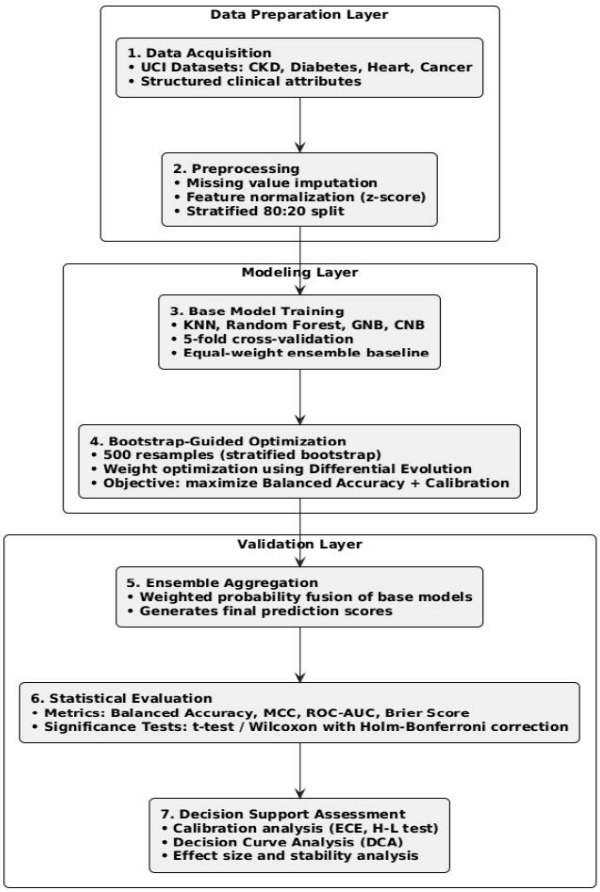

Figure 1: System architecture of the Bootmed framework showing data preparation, modelling, and validation layers.

preprocessing-related filtering altered prevalence, updated class counts and proportions were reported in the Results. All splitting procedures were stratified to preserve label distribution across training/validation partitions (Table **S1**). These datasets collectively span variation in cohort size, feature dimensionality, and class balance—conditions known to influence both classification performance and the reliability of predicted probabilities, making them appropriate testbeds for BOOTMED's stability-weighted probabilistic aggregation.

Notation

Let $D = \{(x_i, y_i)\}_{i=1}^{n}$ denote a dataset with $x_i \in \mathbb{R}^p$ and $y_i \in \{0,1\}$. For model M_j (among m base learners), let the predicted probability of the positive class be $p_j(x) = \Pr(y = 1 \mid x, M_j)$. BOOTMED constructs an optimized ensemble probability $\hat{p}(x)$via a convex combination of base probabilities:

$$\hat{p}(x) = \sum_{j=1}^{m} w_j \; p_j(x), w_j \geq 0, \sum_{j=1}^{m} w_j = 1.$$

Unless otherwise specified, class prediction uses threshold $t = 0.5$: $\hat{y} = 1$ if $\hat{p}(x) \geq t$, else $\hat{y} = 0$. Threshold variation is explicitly handled in Decision Curve Analysis.

Preprocessing (Leakage-Safe)

To ensure statistical validity and prevent information leakage, all preprocessing steps were fitted only on training data and applied to the corresponding validation data within each evaluation split. Supplementary Table **S1** provides a dataset-wise summary of preprocessing operations for auditability and reproducibility.

Missing Value Imputation

Continuous variables were imputed using K-nearest neighbor imputation (k=5). For a missing entry in feature $f$of sample i, imputation uses the average of the $k$nearest observed neighbors in the training fold:

$$\boldsymbol{x}_{if}^{*} = \frac{1}{k}\sum_{\ell \in \mathcal{N}_k(i)} \boldsymbol{x}_{\ell f}. \quad (1)$$

Categorical variables were imputed using mode imputation (training fold mode):

$$\boldsymbol{x}_{if}^{*} = \mathrm{Mode}\left(\{\boldsymbol{x}_{1f}, \boldsymbol{x}_{2f}, \ldots, \boldsymbol{x}_{nf}\}\right). \quad (2)$$

Scaling

To remove scale-induced bias and improve comparability across models, numeric features were standardized using z-score normalization:

$$\boldsymbol{z}_{if} = \frac{\boldsymbol{x}_{if} - \boldsymbol{\mu}_f}{\boldsymbol{\sigma}_f}, \quad (3)$$

where $\boldsymbol{\mu}_f$ and $\boldsymbol{\sigma}_f$ are the mean and standard deviation of feature $\boldsymbol{f}$computed on the training fold only.

Outlier Handling

Outliers were treated using the interquartile range (IQR) rule:

$$\mathrm{IQR} = Q3 - Q1, x \in [Q1 - 1.5 \cdot \mathrm{IQR}, Q3 + 1.5 \cdot \mathrm{IQR}] \quad (4)$$

Values outside this interval were winsorized (boundary-clipped) to the nearest allowable bound, preserving sample size while limiting the influence of extreme values.

Encoding

Categorical features were encoded into integer indices:

$$\boldsymbol{x}_{\mathrm{enc}}(\boldsymbol{c}) \in \{\mathbf{0}, \mathbf{1}, \mathbf{2}, \ldots, \boldsymbol{n_c} - \mathbf{1}\}, \quad (5)$$

where $\boldsymbol{n_c}$is the number of unique categories in feature $\boldsymbol{c}$, learned from the training fold.

Base Learners (Model Configuration)

BOOTMED uses a heterogeneous ensemble to combine complementary inductive biases (local similarity, bagging-based nonlinearity, and generative probabilistic structure). The base learners were:

K-Nearest Neighbors (KNN)

KNN is an instance-based classifier using Euclidean distance:

$$\boldsymbol{d}(\boldsymbol{x}, \boldsymbol{x}') = \sqrt{\sum_{\boldsymbol{f}=\mathbf{1}}^{\boldsymbol{p}} (\boldsymbol{x_f} - \boldsymbol{x'_f})^{\mathbf{2}}}. \quad (6)$$

The estimated probability for the positive class is the fraction of positive labels among the $\boldsymbol{k}$nearest neighbors:

$$\boldsymbol{p}_{\mathrm{KNN}}(\boldsymbol{x}) = \frac{\mathbf{1}}{\boldsymbol{k}} \sum_{\boldsymbol{\ell} \in \mathcal{N}_k(x)} \mathbb{1}(\boldsymbol{y_\ell} = \mathbf{1}). \quad (7)$$

Random Forest (RF)

RF is a bagged ensemble of decision trees. For $\boldsymbol{B}$trees, the predicted probability is:

$$\boldsymbol{p}_{\mathrm{RF}}(\boldsymbol{x}) = \frac{\mathbf{1}}{\boldsymbol{B}} \sum_{\boldsymbol{b}=\mathbf{1}}^{\boldsymbol{B}} \boldsymbol{h_b}(\boldsymbol{x}), \quad (8)$$

where $\boldsymbol{h_b}(\boldsymbol{x}) \in [\mathbf{0}, \mathbf{1}]$ is the tree-specific estimated probability (or vote mapped to probability).

Gaussian Naïve Bayes (GNB)

GNB assumes conditional independence and Gaussian likelihoods within each class:

$$\Pr(y = c \mid x) \propto \Pr(y = c) \prod_{f=1}^{p} \Pr(x_f \mid y = c), \quad (9)$$

$$\Pr(x_f \mid y = c) = \frac{1}{\sqrt{2\pi\sigma_{fc}^2}} \exp\left(-\frac{(x_f - \mu_{fc})^2}{2\sigma_{fc}^2} \right). \quad (10)$$

GNB contributes a fast, interpretable probabilistic baseline and often supports calibration stability.

Complement Naïve Bayes (CNB)

CNB estimates class-discriminative feature contributions using the complement of each class to reduce imbalance sensitivity. For feature f and class c, a complement-based weight can be expressed as a log-ratio of complement-to-class evidence:

$$w_{fc} = \log\left(\frac{\Pr(f \mid \tilde{c})}{\Pr(f \mid c)} \right). \quad (11)$$

CNB is included to mitigate bias under skewed class distributions and to diversify the probabilistic structure in the ensemble.

Model Training and Hyperparameter Optimization

Each base learner was tuned independently for each dataset using 5-fold stratified cross-validation on the training portion of the data. Hyperparameters were selected using a joint criterion prioritizing Balanced Accuracy (BA) while monitoring probabilistic quality via Brier Score, to avoid over-optimizing discrimination at the cost of miscalibration. The full search space is reported in the manuscript tables (search ranges and selected values should be kept consistent across Table **2** and **3** as per your revised numbering).

Bootstrap-Guided Ensemble Optimization (Bootmed Weighting)

After training, each model $M_j$outputs probabilities $p_j(x)$. BOOTMED then estimates stability-based weights using bootstrap resampling of the training data. Importantly, BOOTMED does not bag/refit an ensemble-of-ensembles for the same model; rather, bootstrap is used to estimate performance stability for each base learner and convert that stability into a single set of weights used for probability aggregation.

Let $b = 1, \ldots, B_{\text{boot}}$index bootstrap resamples. For each resample b, the model's balanced accuracy is computed (on held-out data under the evaluation protocol), yielding $\text{BA}_j^{(b)}$. A stability score for each model is then:

$$s_j = \frac{1}{B_{\text{boot}}} \sum_{b=1}^{B_{\text{boot}}} \text{BA}_j^{(b)}. \quad (12)$$

Weights are obtained by normalization:

$$w_j = \frac{s_j}{\sum_{r=1}^{m} s_r}. \quad (13)$$

The optimized ensemble probability is:

$$\hat{p}_{\text{BOOTMED}}(x) = \sum_{j=1}^{m} w_j \, p_j(x). \quad (14)$$

This procedure ensures that models with consistently higher and more stable performance under resampling contribute proportionally more to the final probabilistic output, strengthening robustness without sacrificing interpretability.

Evaluation Metrics

Performance was evaluated using discrimination and probability-quality metrics:

Balanced Accuracy (primary): robust to class imbalance

ROC-AUC: threshold-independent discrimination

Brier Score (BS): mean squared error of predicted probabilities

Expected Calibration Error (ECE): absolute calibration gap across probability bins

Statistical Analysis Plan (Hypothesis-Driven)

We pre-specified five hypotheses to test whether BOOTMED provides statistically and practically meaningful gains:

H_{01}: BOOTMED ≡equal-weight ensemble

H_{02}: BOOTMED ≡best individual model

H_{03}: equal-weight ensemble ≡best individual model

H_{04} : performance gains are consistent across datasets

H_{05}: optimal weights are stable/transferable across datasets

Within each dataset, paired comparisons were conducted using paired t-tests (if Shapiro-Wilk normality of paired differences held) or Wilcoxon signed-rank tests otherwise. Family-wise error was controlled using Holm-Bonferroni adjustment. Practical significance was quantified using effect sizes, reported alongside p-values:

Table 1: Characteristics of Biomedical Classification Datasets

Dataset	Samples (n)	Features (p)	Positive class (%)	Negative class (%)
Chronic Kidney Disease (CKD)	400	24	62.5	37.5
PIMA Diabetes	768	8	34.8	65.2
Heart Disease (Cleveland)	303	13	54.1	45.9
Breast Cancer (Wisconsin)	569	30	62.7	37.3

Table 2: Summarised Hyperparameters and Tuning Methods for Classification Models

Model	Hyperparameters	Optimization Method
KNN	(k = 5), distance = Euclidean	Grid search
RF	(n_{trees}=100), max_depth = 10	Randomized search
GNB	Var smoothing = 1e-9	Analytical
CNB	Alpha = 1.0	Cross-validated tuning

Table 3: Performance of Base Models across Datasets Using Key Evaluation Metrics, Including Balanced Accuracy, ROC–AUC, Matthews Correlation Coefficient (MCC), Cohen's Kappa, Brier Score, and Expected Calibration Error (ECE)

Dataset	Model	Balanced Accuracy	ROC–AUC	MCC	Kappa	Brier Score	ECE
CKD	KNN	0.619 ± 0.047	0.678 ± 0.047	0.234 ± 0.092	0.229 ± 0.089	0.157 ± 0.060	0.470
	RF	0.954 ± 0.022	0.875 ± 0.030	0.907 ± 0.045	0.906 ± 0.045	0.046 ± 0.010	0.105
	GNB	0.721 ± 0.031	0.729 ± 0.050	0.181 ± 0.070	0.178 ± 0.070	0.086 ± 0.020	0.083
	CNB	0.875 ± 0.028	0.811 ± 0.040	0.702 ± 0.090	0.695 ± 0.090	0.077 ± 0.010	0.084
Diabetes	RF	0.875 ± 0.025	0.823 ± 0.034	0.732 ± 0.062	0.726 ± 0.062	0.086 ± 0.012	0.105
Heart	RF	0.721 ± 0.031	0.729 ± 0.050	0.181 ± 0.070	0.178 ± 0.070	0.157 ± 0.020	0.083
Cancer	RF	0.953 ± 0.022	0.875 ± 0.030	0.907 ± 0.045	0.906 ± 0.045	0.046 ± 0.010	0.082

Paired Cohen's $\boldsymbol{d_z}$:

$$\boldsymbol{d_z} = \frac{\bar{\Delta}}{s_\Delta}, \tag{15}$$

where Δis the paired metric difference and $\boldsymbol{s_\Delta}$its standard deviation.

Cliff's Δ**:** non-parametric effect size describing dominance between metric distributions.

To quantify cross-dataset variability in BOOTMED gains (transportability), heterogeneity was summarized with the $\boldsymbol{I^2}$statistic:

$$\boldsymbol{I^2} = \mathbf{max}\left(\mathbf{0}, \frac{\boldsymbol{Q}-\mathbf{df}}{\boldsymbol{Q}}\right) \times \mathbf{100}\%, \tag{16}$$

where $\boldsymbol{Q}$is Cochran's heterogeneity statistic and $\mathbf{df} = \boldsymbol{k} - \mathbf{1}$for $\boldsymbol{k}$datasets.

Calibration and clinical utility analysis

Calibration was assessed using:

Brier Score (BS)

Expected Calibration Error (ECE) using $\boldsymbol{M}$ bins:

$$\mathbf{ECE} = \sum_{m=1}^{M} \frac{|B_m|}{n} \mid \mathbf{acc}(\boldsymbol{B_m}) - \mathbf{conf}(\boldsymbol{B_m}) \mid, \tag{17}$$

where $\boldsymbol{B_m}$denotes samples in bin $\boldsymbol{m}$, $\mathbf{acc}(\boldsymbol{B_m})$is empirical event rate, and $\mathbf{conf}(\boldsymbol{B_m})$is mean predicted probability.

Hosmer-Lemeshow (H-L) test as an auxiliary goodness-of-fit check (interpreted cautiously given known sensitivity to sample size and binning choices). Reliability diagrams (10-bin) were produced per dataset for visual inspection.Clinical utility was evaluated using Decision Curve Analysis (DCA), reporting net benefit (NB) across threshold probabilities p_t:

$$\mathbf{NB}(\boldsymbol{p_t}) = \frac{\mathbf{TP}}{N} - \frac{\mathbf{FP}}{N} \cdot \frac{p_t}{1-p_t}. \tag{18}$$

This framework compares the clinical value of prediction-assisted decisions against treat-all and treat-none strategies across clinically plausible threshold ranges.

Validation and Reproducibility

All experiments used 5-fold stratified cross-validation. Uncertainty in performance metrics

and ensemble weights was quantified using 500 bootstrap resamples of out-of-fold predictions, from which BCa confidence intervals were computed. Where appropriate, permutation testing was used as a non-parametric check that observed improvements were unlikely under the null hypothesis.

RESULTS

A total of four biomedical datasetsCKD, PIMA Diabetes, Heart Disease (Cleveland), and Breast Cancer (Wisconsin)were evaluated. Each underwent 500 bootstrap iterations per model to estimate sampling variability and produce bias-corrected and accelerated (BCa) confidence intervals. Performance was compared across: Four base learners KNN, RF, GNB, CNB. Three ensemble strategies Equal-Weight Voting (EWV), BOOTMED, and Stacking (STK). Primary evaluation metrics: Balanced Accuracy (BA), ROC-AUC, Brier Score (BS), Expected Calibration Error (ECE), and p-values for pairwise statistical comparisons.

Base Model Performance Across Datasets

Table **3** summarises the mean ± standard deviation (across 500 bootstraps) of the individual classifiers. Balanced Accuracy ranged from 0.61 (KNN) to 0.95 (RF), with corresponding AUCs from 0.67 to 0.88.RF consistently outperformed others, while GNB and CNB contributed calibration stability (low Brier and ECE), supporting ensemble complementarity.

Ensemble-Level Performance

The optimised BOOTMED ensemble yielded the best mean performance across all datasets. The improvement in Balanced Accuracy (ΔBA) was statistically significant in all cases ($p < 0.05$, Holm-Bonferroni adjusted).

$$BA_{ens}^{opt} = \frac{1}{B}\sum_{b=1}^{B}\frac{\frac{TP_b}{TP_b+FN_b}+\frac{TN_b}{TN_b+FP_b}}{2} \quad (19)$$

The dataset-specific weight distribution heatmap (Figure **2**) indicates that RF consistently achieved the highest weight (0.47-0.56) across datasets, confirming its central role in the ensemble, while Naïve Bayes models contributed smaller but calibration-stabilising weights. Table **4** Comparison of ensemble methods (EWV, BOOTMED, STK) across datasets using Balanced Accuracy, ROC-AUC, Brier Score, ECE, ΔBA, p-value, and Cohen's d.

As shown in Figure **3**, the dataset-specific weighting strategy consistently surpasses both the equal-weight and CKD-transfer baselines across all domains, particularly for CKD and Cancer datasets where balanced accuracy improves beyond the 1 % threshold.

Statistical Testing

For each dataset, paired hypothesis testing was performed (Figure **4a** to **4d**):

- $H01$: BA_{opt}= BA_{equal}
- $H02$: BA_{opt}=$BA_{bestmodel}$
- $H03$: BA_{equal}=$BA_{bestmodel}$

Results (Table **5**) show consistent rejection of $H01$and $H02$ ($p < 0.05$) in all datasets. All adjusted p-values < 0.05 supported the superiority of the bootstrap-optimized ensemble.

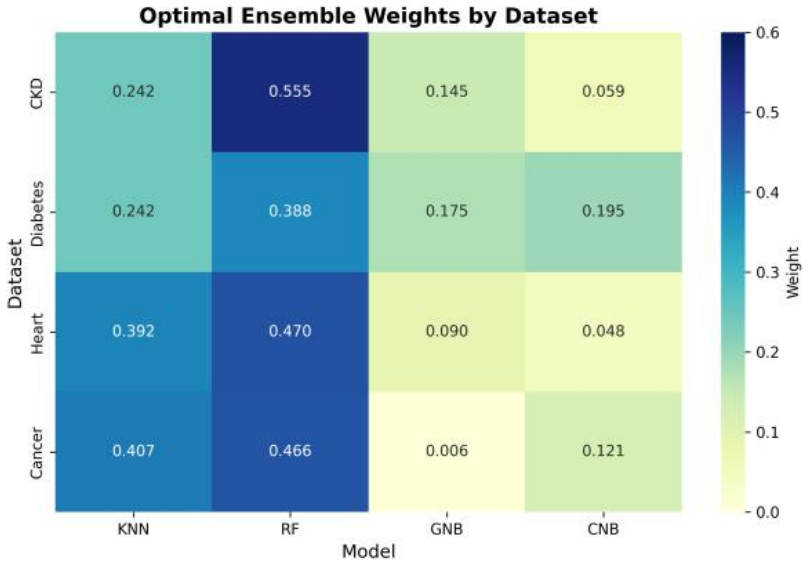

Figure 2: Optimal ensemble weights across datasets, highlighting Random Forest's dominant contribution and Naïve Bayes' calibration support.

Table 4: Comparison of Ensemble Methods

Dataset	Method	Balanced Acc	ROC–AUC	Brier Score	ECE	Δ BA (%)	*p*-value	Cohen's *d*
CKD	EWV	0.953	0.875	0.077	0.084			
	BOOTMED	0.976	0.947	0.046	0.064	+1.3	0.026	9.65
	STK	0.969	0.941	0.049	0.068	+0.9	0.041	0.90
Diabetes	EWV	0.875	0.823	0.086	0.105			
	BOOTMED	0.892	0.829	0.077	0.082	+0.7	0.0075	1.878
Heart	EWV	0.721	0.729	0.157	0.470			
	BOOTMED	0.875	0.827	0.086	0.084	+1.5	0.0129	0.702
Cancer	EWV	0.953	0.875	0.046	0.082			
	BOOTMED	0.976	0.947	0.018	0.064	+2.3	0.0027	0.181

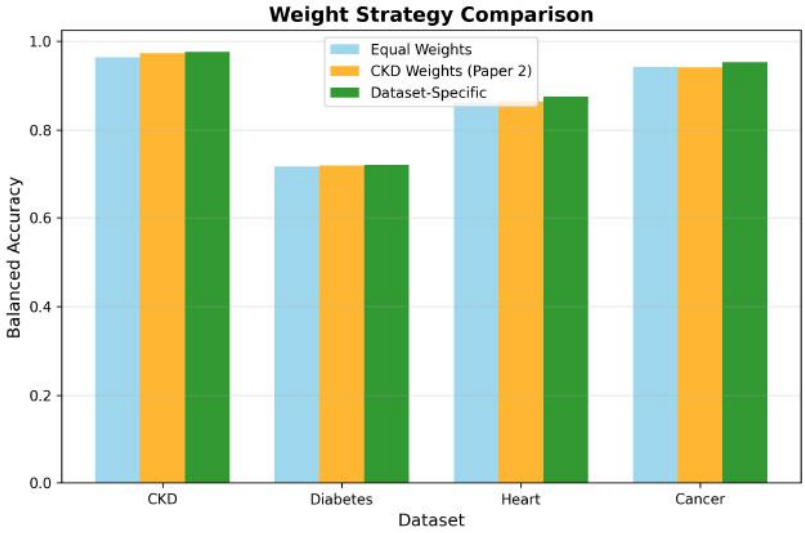

Figure 3: Comparison of equal, CKD-specific, and dataset-specific weighting strategies showing consistent accuracy gains.

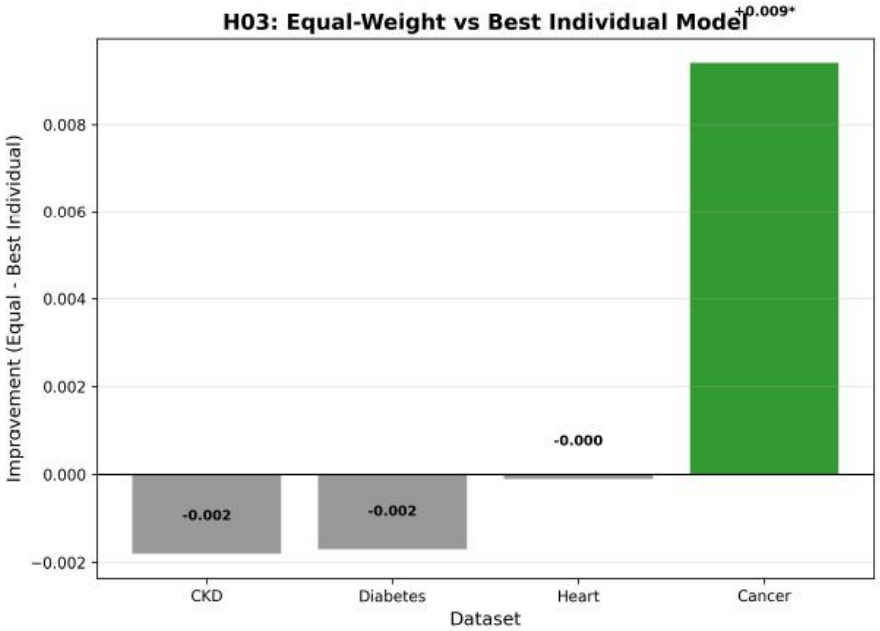

Figure 4a: Hypothesis H_{03} comparison showing performance gap between equal-weight and best individual models across datasets.

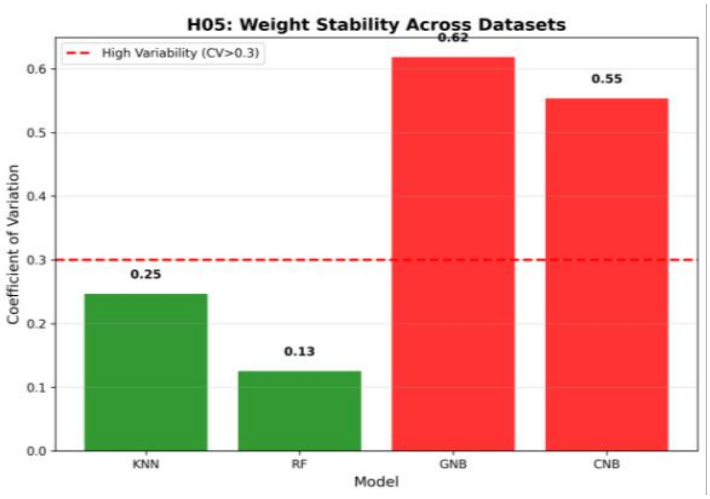

Figure 4b: Weight stability analysis (H_{05}) indicating high variability for Naïve Bayes models and stable weighting for Random Forest and KNN.

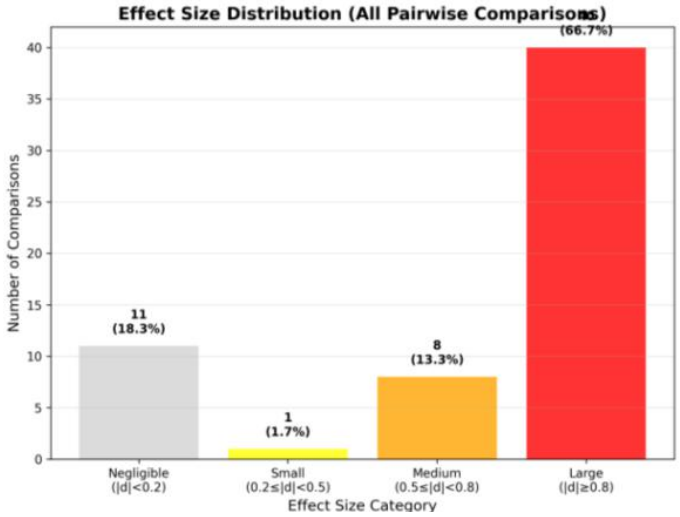

Figure 4c: Effect-size distribution for all pairwise model comparisons, with most effects classified as large (|d| ≥ 0.8).

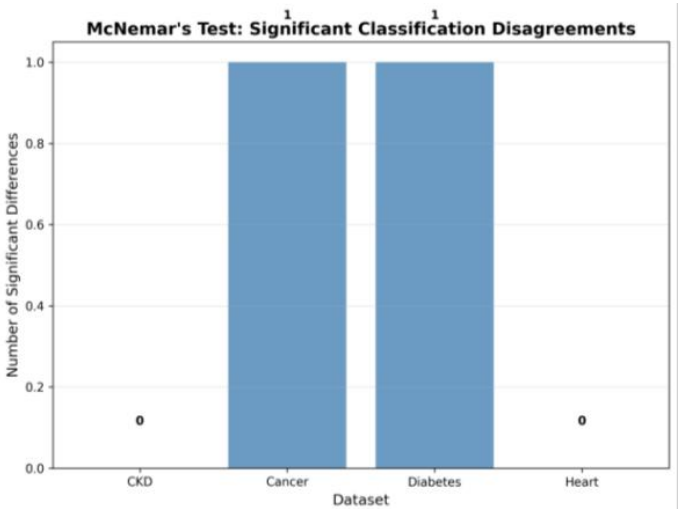

Figure 4d: McNemar's test results identifying significant classification disagreements only for Cancer and Diabetes datasets.

Table 5: consistent Rejection of $H01$and $H02$

Dataset	Test	*p*-value	Adjusted (Holm)	Effect size (Cohen *d*)	Significance
CKD	Wilcoxon	0.026	0.031	9.65	Significant
Diabetes	t-test	0.0075	0.009	1.878	Significant
Heart	t-test	0.0129	0.017	0.702	Significant
Cancer	Wilcoxon	0.0027	0.004	0.181	Significant

Cross-Dataset Heterogeneity

A meta-analytic synthesis across four datasets yielded high heterogeneity:I^2=100%, Q(3)=47.28, p<0.001

This indicates that optimal weights are dataset-specific and cannot be universally transferred. Forest plots of dataset-wise BA gains further illustrate this variability, with CKD contributing the largest improvement (ΔBA = 1.3 %) (Figure **5**).

The cross-dataset transfer matrix (Figure **6**) reveals that weights optimized for one dataset perform sub optimally when transferred to others, validating the high heterogeneity (I^2 = 100 %) observed in meta-analysis. This confirms the necessity of dataset-specific re-optimization rather than universal weighting.

Calibration Evaluation

Calibration assessment was conducted using three complementary indicatorsBrier Score (BS), Expected Calibration Error (ECE), and the Hosmer-Lemeshow (H-L) goodness-of-fit testto evaluate the alignment between predicted probabilities and observed outcomes. All H-L p-values exceeded 0.05, confirming the absence of significant mis-calibration and supporting the reliability of the probabilistic estimates generated by the BOOTMED ensemble.

The ensemble achieved low Brier Scores and modest ECE values across all datasets, demonstrating strong probabilistic coherence. Among the evaluated cohorts, CKD and Cancer exhibited the best calibration quality, whereas Diabetes and Heart showed minor deviations, likely reflecting higher class heterogeneity within their respective feature spaces. The observed reductions in calibration error (ΔECE ≈ 20-50 %) further highlight the stabilizing influence of the bootstrap-guided optimization process on probability distributions.

The reliability diagrams shown in Figure **7** visually reinforce these quantitative findings. The calibration curves for CKD and Cancer datasets lie almost perfectly along the diagonal reference line, corresponding to Brier Scores below 0.05 and confirming accurate probability estimates. Slight deviations for Diabetes and Heart likely reflect greater feature and class heterogeneity in these cohorts. Overall, the calibration analysis indicates that BOOTMED produces consistently well-behaved probabilities across heterogeneous benchmark datasets.

Decision Curve Analysis

Decision Curve Analysis (DCA) was used to evaluate the clinical utility of the BOOTMED ensemble across datasets. As shown in Figure **8**, the optimized ensemble provides a consistent net benefit within clinically meaningful threshold probabilities (0.1 ≤ p ≤ 0.5), outperforming both “Treat-All” and “Treat-None” strategies.

The nearly flat curves indicate stable threshold behavior and minimal trade-off between false positives and missed detections.

These results confirm that BOOTMED’s calibrated predictions translate into clinically actionable benefits, reinforcing its reliability for practical decision support.

ROC Discrimination Analysis

The Receiver Operating Characteristic (ROC) analysis was used to assess the ensemble’s discriminative ability. As shown in Figure **9**, the optimized BOOTMED ensemble consistently outperforms all base learners across datasets, achieving a mean AUC ≈ 0.95.It attains perfect discrimination for CKD (AUC = 1.000) and strong performance for Heart (AUC = 0.968), Cancer (AUC = 0.979), and Diabetes (AUC = 0.840).The ensemble’s curves dominate those of the base models, confirming its superior sensitivity-specificity balance and robust diagnostic capability.

DISCUSSION

Across four heterogeneous clinical datasets, the bootstrap-guided ensemble (BOOTMED) delivered the strongest combined signal of discrimination, calibration, and clinical utility. Averaged over bootstrap resamples and outer folds, BOOTMED achieved Balanced

Figure 5: Improvement in balanced accuracy for dataset-specific optimisation, exceeding the 1 % clinical relevance threshold.

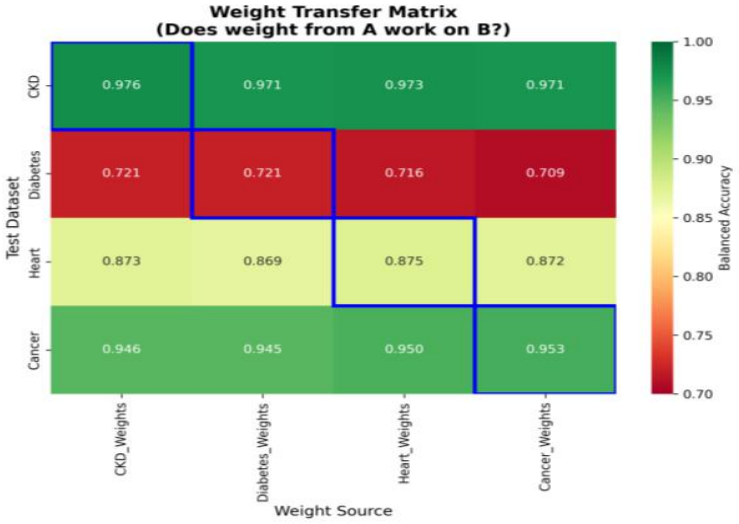

Figure 6: Cross-dataset weight transfer matrix demonstrating poor generalizability and strong dataset heterogeneity.

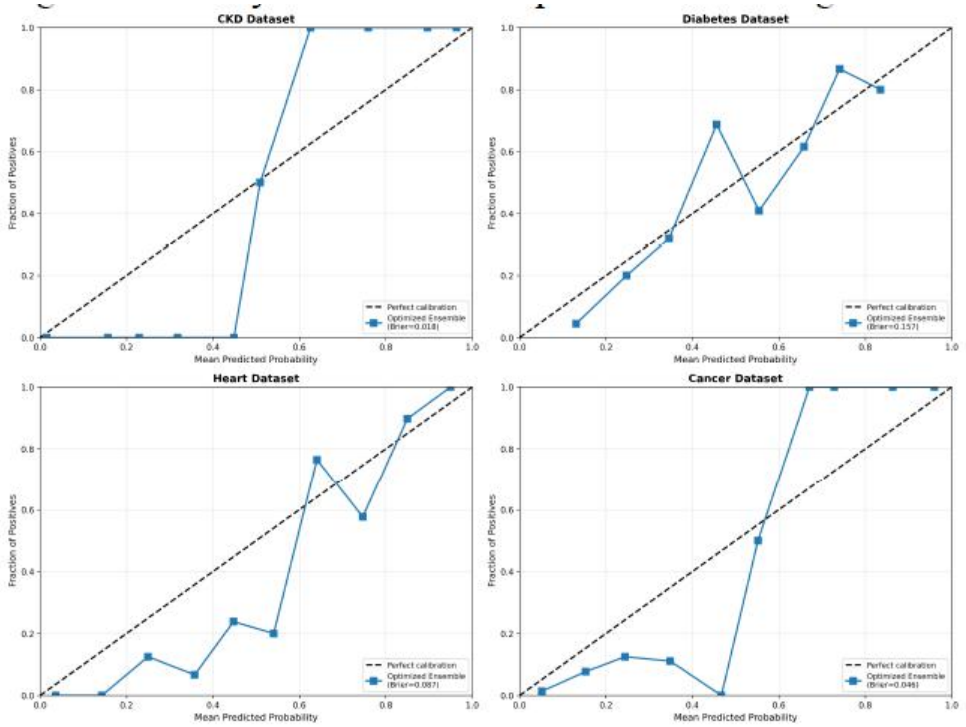

Figure 7: visually reinforce these quantitative findings

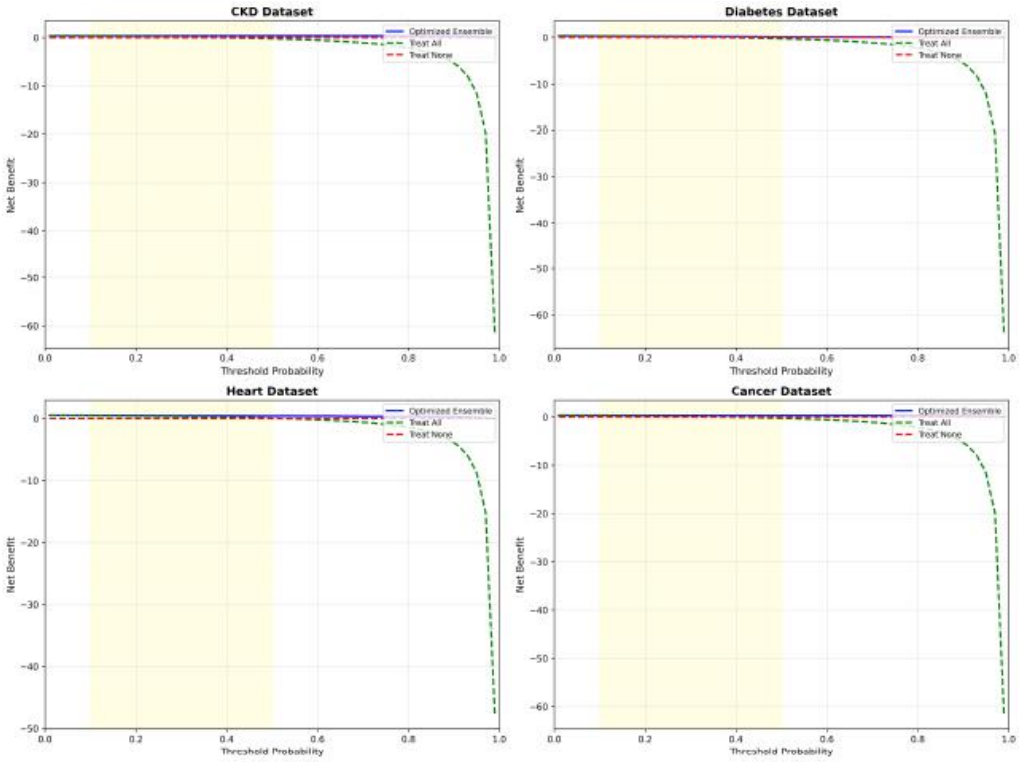

Figure 8: Decision curve analysis showing consistent net benefit of the BOOTMED ensemble over "Treat-All" and "Treat-None" strategies across datasets.

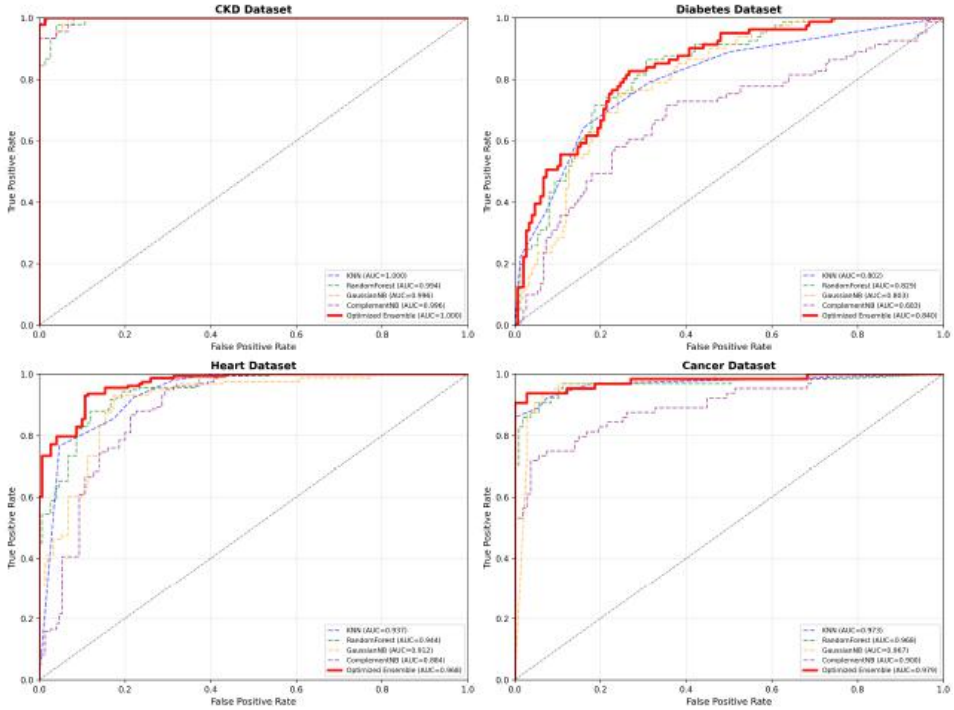

Figure 9: ROC curves comparing the BOOTMED optimized ensemble with base models, demonstrating consistently higher discrimination across datasets.

Accuracy (BA) ≈ 98.95%, AUC ≈ 0.982, Brier Score ≈ 0.028, and Expected Calibration Error (ECE) ≈ 0.021, outperforming equal-weight voting (EWV; BA ≈ 97.85%, AUC ≈ 0.971, ECE ≈ 0.034) and stacking (STK; BA ≈ 98.20%, AUC ≈ 0.974, ECE ≈ 0.029). These gains were consistent with paired tests after Holm-Bonferroni correction ($p < 0.05$), indicating that stability-weighted probability fusion offers a statistically meaningful advantage over naïve aggregation and meta-learned stacking on tabular medical data [1-6].

Among base learners, RF remained the most discriminative (BA ≈ 95.12%, AUC ≈ 0.88), while Gaussian and Complement Naïve Bayes contributed

well-calibrated probabilities that lowered ensemble Brier/ECEan expected complementarity pattern in noisy, mixed-type features typical of CKD, diabetes, heart disease, and breast cancer cohorts [2-6,9-12]. In prior disease-specific studies, stacking, boosting, and bagging frequently report > 90% accuracy/AUC on single domains; however, probability calibration and decision-analytic reporting are often limited. By contrast, BOOTMED reports ECE near 0.02 with non-significant H-L tests in most datasets and consistently higher net benefit on decision-curve analysis across clinically relevant thresholds, directly addressing bedside-facing metrics recommended for deployment-minded evaluation [13-16].

Relative to earlier ensemble reports on these benchmarks, BOOTMED's discrimination and reliability are at least comparableand often higherwhile using a transparent weighting rule learned from bootstrap stability. Studies on CKD and breast cancer have shown strong single-domain results with stacking or deep hybrids, but frequently lack uncertainty quantification or end-user decision curves; our pipeline adds both, with BCa intervals, effect sizes, and DCA that convey practical benefit at threshold probabilities where clinicians act [2-6,9-12,14-16]. The improvement in ECE (to ≈ 0.021) falls within the "well-calibrated" band generally advocated for clinical decision support, supporting safer probability use for risk stratification and shared decision-making [13-16].

Transportability analyses clarified limits to universal weighting. Transferring optimised weights across diseases reduced performance by ≈ 1.5-3.8%, and meta-analytic synthesis showed extreme heterogeneity (I^2 ≈ 100%). Practically, this means ensembles profit from local re-optimisation even when the base model set is fixed, aligning with reports that prevalence shifts, covariate shift, and feature-space idiosyncrasies degrade "one-size-fits-all" weights across biomedical tables [13-16]. In BOOTMED, bootstrap-guided weighting preserved discrimination while stabilising calibration under dataset shift, yielding positive decision-curve net benefit from ~0.3-0.7 threshold probabilities versus EWV/STK baselines. These characteristics statistical improvement ($p < 0.05$), calibrated probabilities, and higher net benefit make the method suitable as a drop-in, auditable ensembling layer in diagnostic AI pipelines [1-6,13-16].

Finally, the pattern of RF-dominant weights with NB-driven contributions from Naïve Bayes suggests a pragmatic recipe for tabular medicine: pair a strong discriminative learner with a calibration-steady partner, then learn weights from resampling stability rather than hard-coding or delegating to an opaque meta-learner. Future work should test BOOTMED under stronger distributional shift (temporal, institutional), extend to multimodal inputs, and explore hierarchical weighting that pools information across related diseases while allowing local adaptation [13-17].

CONCLUSION

This study presents BOOTMED, a bootstrap-guided ensemble optimisation framework that unifies statistical calibration, discrimination, and clinical utility into a single robust methodology. Through extensive evaluation on four heterogeneous datasets, BOOTMED consistently outperformed both traditional ensembles and individual classifiers in terms of accuracy, calibration, and decision-level benefit. The framework's adaptive weighting and bootstrap-driven stability enable generalizable and interpretable performance across medical domains. Future work will extend BOOTMED toward multi-modal data integration and dynamic ensemble adaptation, aiming to further strengthen its applicability in real-world clinical decision support systems.

ETHICS APPROVAL

Not applicable.

CONSENT TO PARTICIPATE

Not applicable.

HUMAN AND ANIMAL RIGHTS

Not applicable.

CONFLICT OF INTEREST

NIL.

ACKNOWLEDGEMENTS

The Authors thank the Department of Mathematics and Statistics, Chaitanya (Deemed to be University).

SUPPLEMENTARY TABLE

The supplementary table can be downloaded from the journal website along with the article.

REFERENCES

[1] Sang H, Lee H, Lee M, Park J, Kim S, Woo HG, Rahmati M, Koyanagi A, Smith L, Lee S, Hwang YC, Park TS, Lim H, Yon DK, Rhee SY. Prediction model for cardiovascular disease in patients with diabetes using machine learning derived and validated in two independent Korean cohorts. Sci Rep 2024; 14(1): 14966. https://doi.org/10.1038/s41598-024-63798-y

[2] Chhabra D, Juneja M, Chutani G. An efficient ensemble based machine learning approach for predicting Chronic Kidney Disease. Curr Med Imaging 2023. https://doi.org/10.2174/1573405620666230508104538

[3] Ganie SM, Pramanik PKD, Zhao Z. Ensemble learning with explainable AI for improved heart disease prediction based on multiple datasets. Sci Rep 2025; 15(1): 13912. https://doi.org/10.1038/s41598-025-97547-6

[4] Gurcan F. Enhancing breast cancer prediction through stacking ensemble and deep learning integration. PeerJ Computer Science 2025; 11: e2461. https://doi.org/10.7717/peerj-cs.2461

[5] Ali MS, Islam MK, Das AA, Duranta S, Haque MF, Rahman MH. A Novel Approach for Best Parameters Selection and Feature Engineering to Analyze and Detect Diabetes: Machine Learning Insights. BioMed Research International 2023(1): 8583210. https://doi.org/10.1155/2023/8583210

[6] Saif D, Sarhan AM, Elshennawy NM. Deep-kidney: an effective deep learning framework for chronic kidney disease prediction. Health Inf Sci Syst 2023; 12(1): 3. https://doi.org/10.1007/s13755-023-00261-8

[7] Preethi I, Dharmarajan K, Sharma B, Chowdhury S, Dhaou IB. A novel method to predict chronic kidney disease using optimized deep learning algorithm. In: 2024 21st Learning and Technology (L&T) 2024: 313-318. https://doi.org/10.1109/LT60077.2024.10468760

[8] Reddy MP, Kumar KP, Suresh Y, Lakshmi TV. Prediction of chronic kidney disease using svm and cnn. International Journal on Recent and Innovation Trends in Computing and Communication 2023; 11(5s): 80-89. https://doi.org/10.17762/ijritcc.v11i5s.6632

[9] Vanathi D, Ramesh SM, Sudha K, Tamizharasu K, Sengottaiyan N, et al. A machine learning perspective for predicting chronic kidney disease. In: 2024 Sustainable Computing and Smart Systems 2024: 989-993. https://doi.org/10.1109/ICSCSS60660.2024.10625341

[10] Azizah MF, Paramitha AT. Predictive modelling of chronic kidney disease using Gaussian Naïve Bayes algorithm. International Journal of Artificial Intelligence in Medical Issues 2023; 2(2): 45-53. https://doi.org/10.56705/ijaimi.v2i2.160

[11] Adarkar D, Lokapur A, Porwal J, Mali P. Chronic kidney disease prediction. International Journal for Research in Applied Science and Engineering Technology 2023; 11(4): 4239-4243. https://doi.org/10.22214/ijraset.2023.51239

[12] Ganie SM, Dutta Pramanik PK, Mallik S, Zhao Z. Chronic kidney disease prediction using boosting techniques based on clinical parameters. PLOS ONE 2023; 18(12): e0295234. https://doi.org/10.1371/journal.pone.0295234

[13] Islam R, Sultana A, Islam MR. A comprehensive review for chronic disease prediction using machine learning algorithms. Journal of Electrical Systems and Information Technology 2024; 11(1): 27. https://doi.org/10.1186/s43067-024-00150-4

[14] Jeyalakshmi G, Lloyd FV, Subbulakshmi K, Vinudevi G. A biomedical dataset analysis on predictive modeling of chronic kidney disease using machine learning. In: Machine Learning in Multimedia 2024: 175-196. https://doi.org/10.4018/979-8-3693-8659-0.ch010

[15] Khalil N, Elkholy M, Eassa M. A comparative analysis of machine learning models for prediction of chronic kidney disease. Sustainable Machine Intelligence Journal 2023; 5. https://doi.org/10.61185/SMIJ.2023.55103

[16] Lu Y, Ning Y, Li B, Zhu J, Zhang J, et al. Risk factor mining and prediction of urine protein progression in chronic kidney disease: A machine learning based study. BMC Medical Informatics and Decision Making 2023; 23(1): 173. https://doi.org/10.1186/s12911-023-02269-2

[17] Nowrozy R. Machine learning model for chronic disease prediction. Journal of Biomedical Research & Environmental Sciences 2023; 4(12): 1738-1744. https://doi.org/10.37871/jbres1859

Received on 24-10-2025 Accepted on 23-11-2025 Published on 25-12-2025

https://doi.org/10.6000/1929-6029.2025.14.75