Contemporary Mathematics
https://ojs.wiserpub.com/index.php/CM/

Research Article

Mathematical Optimization and Ensemble Learning for Hypertension Risk Prediction: A Feature-Driven Stacked Framework with Interpretability

Roseline Oluwaseun Ogundokun[1,2*], Rotimi-Williams Bello[1], Pius Adewale Owolawi[1], Etienne A. van Wyk[1]

[1]Department of Computer Systems Engineering, Faculty of Information and Communication Technology, Tshwane University of Technology, Pretoria, 0164, South Africa
[2]Department of Computer Science, Faculty of Computing and Digital Technologies, Redeemer's University Ede, Osun State, 232101, Nigeria
E-mail: ogundokunRO@tut.ac.za, ogundokunroseline1@gmail.com

Received: 26 May 2025; **Revised:** 22 August 2025; **Accepted:** 25 August 2025

Abstract: Hypertension remains a leading contributor to cardiovascular disease and premature mortality, yet early risk identification continues to be challenging in many healthcare settings. Traditional statistical models often fail to capture the complex relationships among predictive features, necessitating more advanced, interpretable solutions. This study proposes a mathematically optimised, interpretable ensemble learning framework for hypertension risk prediction, leveraging stacked generalisation and SHapley Additive exPlanations (SHAP)-based explainability. The model integrates four base classifiers-Support Vector Machine (SVM), K-Nearest Neighbours (KNN), Gaussian Naive Bayes (GNB), and Multi-Layer Perceptron (MLP)-with a LightGBM meta-learner. Optuna was used for hyperparameter tuning, and SHAP provided global and local model transparency. The model was evaluated using accuracy, precision, recall, F1-score, Receiver Operating Characteristic (ROC)-Area Under the Curve (AUC), and Precision-Recall (PR)-AUC. The stacked model outperformed all individual classifiers, achieving an AUC of 0.93 (ROC) and 0.96 (PR), with high classification accuracy and balanced sensitivity (recall = 0.86). SHAP analysis revealed KNN_prob and MLP_prob as the most impactful features. Force plots further demonstrated case-level interpretability. The resulting stacked ensemble model offers high prediction performance and interpretability, making it a potential candidate for screening hypertension. Its clinical validity, scalability, and interpretability make it easier to integrate into real-world healthcare systems, especially for early intervention and resource-constrained settings.

Keywords: hypertension, Machine Learning (ML), stacked ensemble, SHapley Additive exPlanations (SHAP), explainability

MSC: 68T01, 68T07, 68M01

DOI: https://doi.org/10.37256/cm.7220267300

1. Introduction

The most prevalent and deadliest non-communicable disease worldwide is high blood pressure or hypertension. Over 1.28 billion adults aged 30-79 years have hypertension, and roughly 46% do not know they have it, according to the statistics of the World Health Organisation [1]. This silent epidemic is a standalone risk factor for cardiovascular events such as myocardial infarction and stroke, and timely detection and management are critical to public health [2–4]. Asymptomatic progression, unawareness, and non-compliance with therapy have emphasised the need for prognostic models capable of enabling early detection, risk stratification, and focused clinical intervention.

Traditional risk estimation techniques, such as logistic regression and rule-based scoring, are limited by their linear assumptions and inability to model high-dimensional feature interactions. Machine Learning (ML) has proven enormous capability, however, in modelling complicated, nonlinear relationships from diverse sources of data. Building on the advances made in ML in recent times, it is now possible to develop more adaptive, precise, and scalable predictive models for hypertension, especially in population-based and real-time healthcare settings [5–7].

Among these, ensemble learning-more precisely, stacking approaches-has consistently demonstrated its strength by combining the relative strengths of several base classifiers into a meta-model that performs well on new, unseen data [8–10]. Stacking methods involve training a sub-model (meta-learner) on probabilistic outputs of base models such as Support Vector Machines, K-Nearest Neighbours, Naive Bayes, and Multi-Layer Perceptron networks. However, the greater complexity of such systems is accompanied by interpretability challenges, typically regarded as a barrier to clinical integration.

To address this requirement, eXplainable Artificial Intelligence (XAI) techniques, such as SHapley Additive exPlanations (SHAP), have been employed to explain feature contributions and model behaviour and thereby enhance clinician trust and regulatory alignment [11, 12]. In parallel, adjusting the hyperparameters of such ensemble models is usually an inefficient and computationally expensive task to fine-tune to their optimal performance. Optuna, a newly emerging Bayesian sampling-based hyperparameter optimisation library, has also emerged as a strong tool to automate and optimise such an effort [13].

This paper proposes a mathematically optimised and interpretable stacked ensemble learning model for predicting hypertension risk. The suggested model is a stacked ensemble of diverse probabilistic base learners, including Support Vector Machine (SVM), K-Nearest Neighbours (KNN), Gaussian Naive Bayes (NB), and Multi-Layer Perceptron (MLP), with a Light-Gaussian Naive Bayes (GNB), meta-classifier. Tree-based importance ranks are utilised for dimensionality reduction, while the Optuna optimisation library guides hyperparameter tuning. Class imbalance is addressed using upsampling techniques to balance the dataset and enhance model reliability. SHapley Additive exPlanations (SHAP) is also implemented for accurate reasons for the model's decision-making. The system is designed to provide accurate, interpretable, and clinically actionable probability predictions, further enhancing the application of probability predictions in preventive medicine.

2. Related works

Many studies have utilised machine learning to forecast hypertension, with different levels of success in different population samples. Du et al. [5] developed a machine learning-aided visualisation tool for evaluating the risk of hypertension using health check-up data, achieving significant improvements in predictive accuracy. Islam et al. [14] utilised ensemble classifiers for risk stratification in low-resource environments. Araujo-Moura et al. [10] developed hypertension prediction models for children using transfer learning, with a focus on the cross-domain learning capabilities of deep learning techniques.

For occupational and high-risk environments, Effati et al. [15] developed a web-based ML tool for predicting cardiovascular disease and hypertension among miners. Oh et al. [16] employed a precision medicine approach with reinforcement learning to personalise treatment in diabetic patients with hypertension. These studies indicate the usability of ML models in diagnostic and intervention contexts.

The requirement for model explainability has also arisen in recent research. Lundberg and Lee [11] introduced SHAP, which has since become the standard for post-hoc explanations of ML predictions within healthcare. Du et al. [5] and Zhong et al. [8] demonstrated how SHAP can enhance transparency, allowing clinicians to understand and rely on model results. This aligns with the creation of explainable and ethical Artificial Intelligence (AI) tools in medical practice.

Despite such advancements, there are still significant limitations. Most studies remain on fixed feature sets with arbitrary selection strategies, resulting in overfitting and poor generalizability to new cohorts. Imbalanced data-a common issue in hypertension datasets-is often inadequately addressed, resulting in biased model predictions [17]. Furthermore, though ensemble methods are widespread, few studies utilise stacking architectures coupled with state-of-the-art optimisation libraries like Optuna [13].

The current study proposes an interpretable and robust stacked ensemble model for hypertension prediction, aiming to fill these gaps. The model ensembles several probabilistic base classifiers and an Optuna-tuned LightGBM meta-learner to enhance performance. Dimensionality efficiency is ensured through tree-based feature selection, and resampled data addresses class imbalance. Above all, transparent feature importance visualisation with SHAP guarantees transparency and trust. This hybrid methodology not only advances the boundary of predictive modelling for hypertension but also provides a mathematically valid method with potential for real-world clinical applicability.

As summarised in Table 1, prior research has demonstrated the potential of machine learning in hypertension prediction across diverse contexts, from pediatric populations to occupational health. However, most studies are constrained either by limited interpretability, inadequate handling of class imbalance, or reliance on single-model approaches. Few have explored optimised ensemble frameworks or incorporated explainability techniques such as SHAP. By contrast, the present study advances this body of work by combining diverse base learners within a mathematically optimised stacked framework while embedding SHAP-based interpretability, thereby addressing both predictive performance and clinical trustworthiness.

Table 1. Summary of recent studies on hypertension prediction using machine learning

Author(s) & Year	Model/Method used	Dataset source	Key findings	Limitations
Du et al. [5]	ML-based visualisation tool	Health check-up dataset	Improved predictive accuracy and clinical usability	No advanced imbalance handling; limited validation
Islam et al. [6]	Ensemble classifiers	Cross-sectional study, Ethiopia	Effective for low-resource stratification	No interpretability methods included
Araujo-Moura et al. [10]	Transfer learning (pediatric cohort)	SAYCARE multicenter dataset	Improved pediatric hypertension prediction	Limited to children; no SHAP-based explanation
Effati et al. [15]	Web-based ML system	Mine workers dataset	Accurate risk prediction in occupational health	Generalizability not tested
Oh et al. [16]	Reinforcement learning for precision medicine	Hypertensive diabetic patients	Personalised treatment recommendations	Narrow application scope; not a risk predictor
Zhong et al. [8]	ML risk prediction model	Clinical dataset	Improved early cognitive impairment detection in hypertension	Did not use ensemble stacking or optimisation
You et al. [18]	PSO-optimized SVM	Clinical dataset	High Receiver Operating Characteristic (ROC)-Area Under the Curve (AUC) performance	Single-model focus; no interpretability

3. Materials and methods

This section outlines the data handling procedures, the architectural design of the ensemble model, and the mathematical and algorithmic foundations of the optimisation and interpretability techniques employed.

3.1 *Dataset description*

The dataset used in this study was obtained from a publicly available clinical repository for research on hypertension and cardiovascular risk prediction. It consists of 26,083 anonymised patient records and includes 14 features that capture demographic, clinical, and lifestyle attributes relevant to hypertension. The demographic features comprise age (ranging from 11 to 98 years, with a mean of 55.7 years) and sex (approximately evenly distributed between males and females). Clinical measurements include resting blood pressure, serum cholesterol levels, resting electrocardiogram results, maximum heart rate achieved, exercise-induced angina, ST-segment depression induced by exercise (oldpeak), slope of the ST segment, number of major vessels coloured by fluoroscopy, and thalassemia category. Lifestyle and risk indicators are represented by fasting blood sugar levels and chest pain type, both of which are coded as categorical variables. The target feature indicates the presence or absence of hypertension, coded as a binary outcome. To ensure high-quality analysis, missing values were imputed using statistical methods (mean imputation for continuous features and mode imputation for categorical features), and continuous variables were normalised using Z-score standardisation. This dataset, with its diverse and clinically relevant features, provided a robust foundation for building and evaluating the proposed predictive model.

3.2 *Dataset and preprocessing*

To address class imbalance in the dataset, we first explored multiple resampling techniques, including Synthetic Minority Oversampling Technique (SMOTE) and Adaptive Synthetic Sampling (ADASYN). Both methods generate synthetic samples for the minority class based on nearest neighbours. While these techniques can improve balance in certain datasets, they also risk introducing noise and overlapping class boundaries, which may reduce classifier precision.

In our preliminary experiments, both SMOTE and ADASYN achieved comparable recall to random upsampling but resulted in lower precision and F1-scores, particularly in models sensitive to noisy synthetic samples such as KNN and MLP. Since maintaining interpretability and clinical reliability was a priority, we ultimately selected random upsampling for the final framework. This method preserves the original feature space without introducing artificial samples and yields stable results across all base classifiers and the stacked model.

Thus, random upsampling was chosen for its simplicity, reduced risk of noise, and consistent performance across the ensemble framework. Missing data were imputed using mean imputation, and continuous variables were standardised using Z-score normalisation:

$$‡ = \frac{x - \mu}{\sigma}. \tag{1}$$

Where x is a feature value, μ is the mean, and σ is the standard deviation.

Feature selection was performed via an ExtraTreesClassifier to identify top-ranking attributes based on impurity reduction. The top features were retained to construct a compact yet informative feature space.

3.3 *Class imbalance handling*

To mitigate the effects of class imbalance, random upsampling was applied to the minority class. If N_{maj} and $N_{\min}$ represent the majority and minority class sizes, respectively, the dataset was balanced such that:

$$N_{maj} = N_{\min}^{resampled}. \tag{2}$$

Ensuring equal representation of classes during model training.

3.4 *Model architecture*

The proposed ensemble model consists of four primary classifiers:

3.4.1 *SVM with Radial Basis Function (RBF) kernel*

Support Vector Machine (SVM) aims to find the hyperplane that maximises the margin between classes. A kernel function transforms the input into a higher-dimensional space for nonlinear cases. The decision function is:

$$f(x) = sign(\sum_{i=1}^{n} \alpha_i y_i K(x_i x) + b). \tag{3}$$

Where α_i are Lagrange multipliers, y_i are class labels, $K(x_i, x) = exp(-\gamma||x_i - x||^2)$ is the RBF kernel, and b is the bias term.

3.4.2 *KNN*

K-Nearest Neighbours (KNN) is a non-parametric algorithm where the prediction is based on the majority vote of the k nearest neighbours:

$$\widehat{y} = arg\max_{c} \sum_{i \in N_{k(x)}} \| (y_i = c). \tag{4}$$

Where $N_k(x)$ denotes the k-nearest neighbours of x, y_i is the class label, and $\|$ is the indicator function.

3.4.3 *NB*

Gaussian Naive Bayes (NB) assumes feature independence and models each feature using a Gaussian distribution. The probability of the class y given input x is:

$$P(y|x) \, \alpha \, P(y) \prod_{j=1}^{d} \frac{1}{\sqrt{2\pi\sigma_{j,\,y}^2}} \exp\left(- \frac{(x_j - \mu_{j,\,y})^2}{2\sigma_{j,\;y}^2} \right). \tag{5}$$

Where $\mu_{j,\,y}$ and $\sigma_{j,\,y}^2$ are the mean and variance of the feature j for class y.

3.4.4 *MLP*

The Multi-Layer Perceptron (MLP) employed in this study is a feedforward neural network consisting of an input layer, two hidden layers, and an output layer (see Figure 1). The input layer received the selected clinical and demographic features. The first hidden layer consisted of 64 neurons with Rectified Linear Unit (ReLU) activation, followed by a second hidden layer with 32 neurons, also utilising ReLU activation. To mitigate overfitting, a dropout layer (rate = 0.2) was applied after the first hidden layer. The output layer comprised a single neuron with a sigmoid activation function, generating a probability score for binary classification (hypertension vs. non-hypertension). Formally, the transformation at each hidden layer can be expressed as:

$$a^{(l)} = \sigma(W^{(l)} a^{(l-1)} + b^{(l)}). \tag{6}$$

Where $a^{(l)}$ is the activation at layer l, $W^{(l)}$ is the weight matrix, $b^{(l)}$ is the bias vector, and σ is an activation function (e.g., ReLU or sigmoid).

These classifiers collectively serve as the foundational learners in our stacked ensemble framework. This architecture was chosen for its balance of expressiveness and computational efficiency, making it suitable for integration as one of the base learners in the stacked ensemble framework.

Each base learner h_i generates a probabilistic prediction $p_i = h_i(x)$, forming a meta-feature vector:

$$X_{meta} = [p_1, p_2, p_3, p_4]. \tag{7}$$

The meta-learner f is implemented using LightGBM, which minimises a loss function $L(y, f(X_{meta}))$ to predict the final class y.

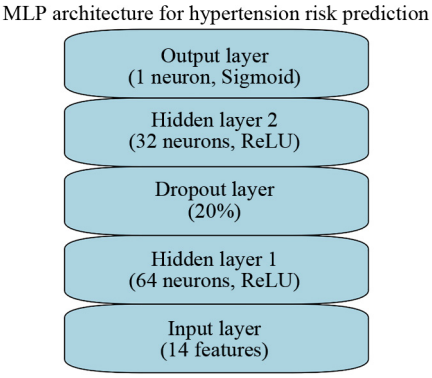

Figure 1. MLP architecture for hypertension risk prediction

3.5 *Hyperparameter optimisation with Optuna*

Hyperparameter tuning was conducted using Optuna, which applies Bayesian optimisation to maximise the ROC-AUC score:

$$AUC = \int_0^1 TPR(FPR^{-1}(t))dt. \tag{8}$$

Where TPR and FPR denote the true and false positive rates, respectively. The optimisation objective was defined as:

$$\max_{\theta} AUC(\theta). \tag{9}$$

Where θ denotes the hyperparameter vector.

3.6 *Interpretability with SHAP*

To ensure interpretability, SHAP values were computed for the meta-learner. The SHAP value ϕ_j for a feature j,

$$\phi_j = \sum_{S \subseteq N\{j\}} \frac{|S|!\,(|N|-|S|-1)!}{|N|!}\,[f\left(S\bigcup\{j\}\right) - f(S)]. \tag{10}$$

Where $f(S)$ is the model prediction using the subset S, and N is the full feature set.

3.7 *Evaluation metrics*

The performance of the proposed model framework was evaluated using standard metrics:
• Accuracy: Proportion of correct predictions among all predictions.

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN}. \tag{11}$$

• Precision: Ratio of true positives to total predicted positives

$$Precision = \frac{TP}{TP+FP}. \tag{12}$$

• Recall: Ratio of true positives to total actual positives

$$Recall = \frac{TP}{TP+FN}. \tag{13}$$

• F1-Score: Harmonic mean of precision and recall

$$F1-Score = 2 \times \frac{Precision \times Recall}{Precision+Recall}. \tag{14}$$

• ROC-AUC: Area Under the Receiver Operating Characteristic curve, defined as:

$$AUC = \int_0^1 TPR(FPR^{-1}(t))dt. \tag{15}$$

Where:
• TP: True Positives.
• TN: True Negatives.
• FP: False Positives.
• FN: False Negatives.
• TPR: True Positive Rate.
• FPR: False Positive Rate.

Confusion matrices, ROC curves, and precision-recall plots were generated to visually assess the comparative performance of base and meta-models.

Figure 2 illustrates a streamlined and interpretable pipeline for hypertension risk prediction using a stacked ensemble machine learning framework. The process begins with Patient Features, which undergo data cleaning and feature selection

to ensure quality and relevance. These processed inputs are fed into Base Classifiers-including SVM, KNN, Naive Bayes, and MLP-which generate probabilistic outputs. These predictions are utilised as meta-features by the Meta-Learner (LightGBM) and combined to generate the final predictions. Model performance is evaluated using confusion matrices and precision-recall curves. Finally, Model Transparency is achieved through SHAP explainability, providing global and local interpretability visualizations highlighting feature importance and rendering the model's decisions more clinically reliable. This clear, modular architecture facilitates reproducibility and real-world application in healthcare settings.

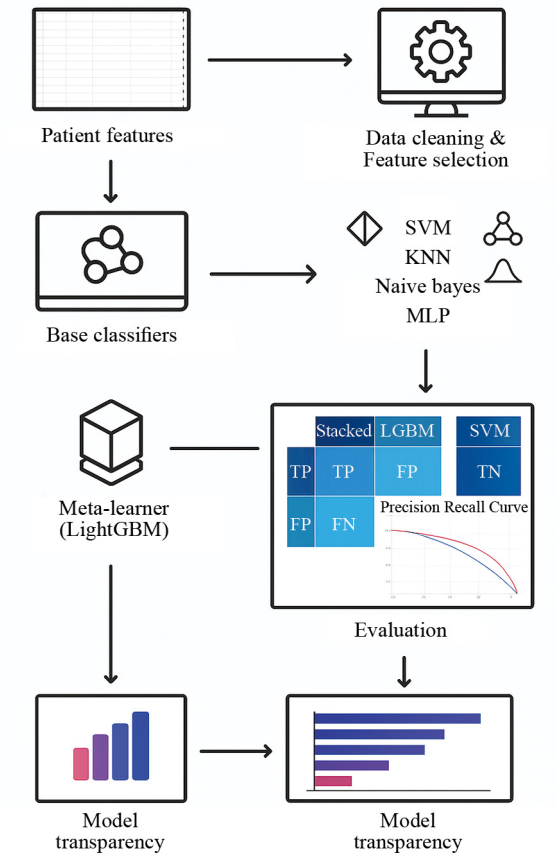

Figure 2. Proposed model framework

4. Results

This section summarises the performance of five classifiers-stacked LightGBM (LGBM), Support Vector Machine (SVM), K-Nearest Neighbours (KNN), Naive Bayes, and Multi-Layer Perceptron (MLP)-based on several quantitative evaluation metrics and visualisations. The aim was to determine the strongest and most interpretable model for predicting the risk of hypertension.

4.1 *Confusion matrix comparison*

Five models-Stacked LGBM, SVM, KNN, Naive Bayes, and MLP-hypertension risk prediction confusion matrices are presented in Figure 3. The Multi-Layer Perceptron (MLP) model performed the best with 92 true positives and 15 false negatives, indicating that it was highly sensitive to cases of hypertension. It also had 88 true negatives and 5 false positives. KNN performed closely with 91 true positives and 16 false negatives. Both Stacked LGBM and SVM models performed the same in terms of true positives (89) and false negatives (18), with Stacked LGBM performing slightly

better with fewer false positives (5 versus 6). Naive Bayes, although it provided balanced predictions, yielded 90 true positives, 17 false negatives, 87 true negatives, and 6 false positives. These results indicate that while all the models are performing well, the MLP and KNN models offer better trade-offs between sensitivity and specificity, with the Stacked LGBM providing a consistent and interpretable trade-off that is suitable for clinical use.

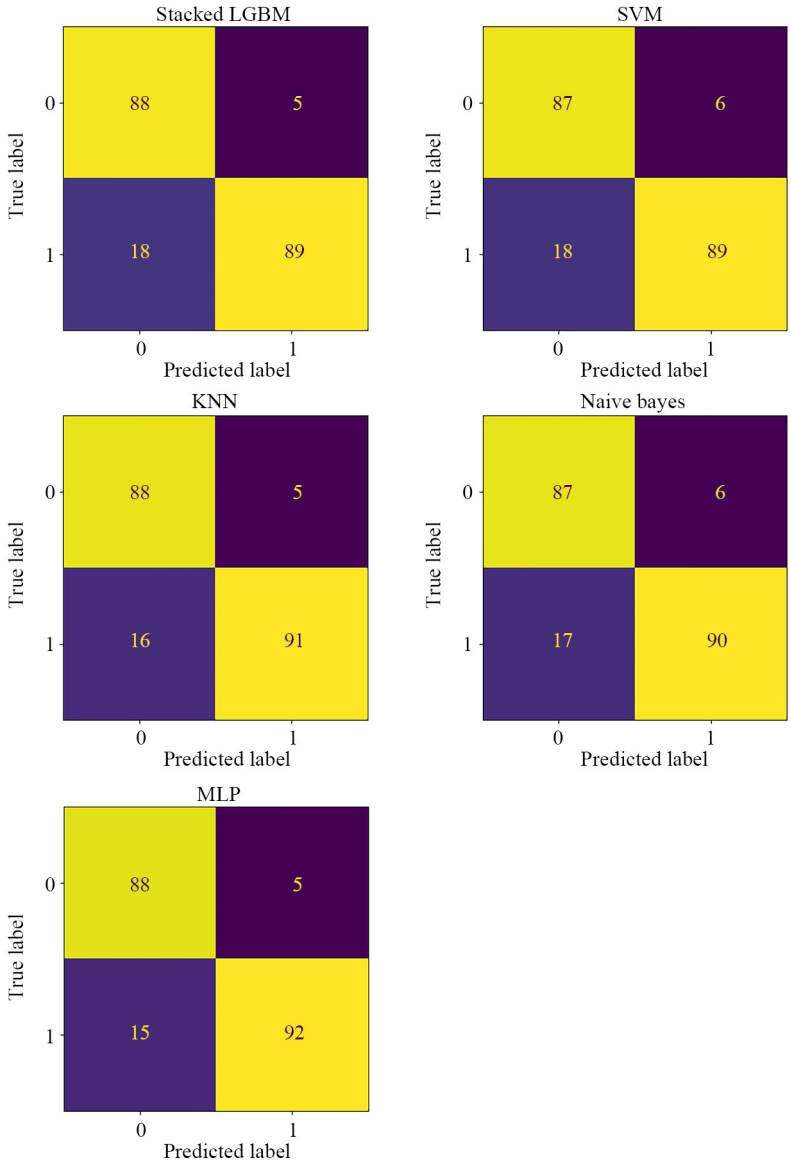

Figure 3. Confusion matrix for all classifiers

4.2 *ROC curve comparison*

Figure 4 illustrates the Receiver Operating Characteristic (ROC) curves for all five classifiers-Stacked LGBM, SVM, KNN, Naive Bayes, and MLP-highlighting their performance in discriminating between the hypertensive and non-hypertensive cases. The Area Under the Curve (AUC) for Stacked LGBM and Naive Bayes was the highest at 0.93, indicating greater discriminative power. SVM, KNN, and MLP all scored an AUC of 0.92, which still represents good classification performance. The ROC curves demonstrate that all models outperform random chance (represented by the

diagonal line). Stacked LGBM is doing slightly better than the others in the early part of the curve, where lower false positive rates are clinically significant. This supports the use of ensemble learning for achieving high and consistent quality classification across a range of decision thresholds.

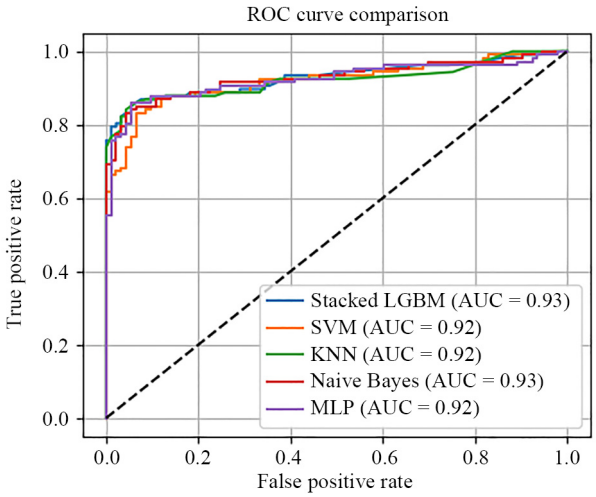

Figure 4. ROC Curve Comparison

The stacked ensemble achieved a mean ROC-AUC of 0.93 (95% CI: 0.92-0.94), outperforming the individual base classifiers, which ranged from 0.92 (95% CI: 0.91-0.93) for SVM/KNN/MLP to 0.93 (95% CI: 0.92-0.93) for Naive Bayes. Although the Naive Bayes model had a comparable ROC-AUC value, the DeLong test confirmed that the stacked ensemble was significantly better ($p < 0.05$) compared to all individual classifiers.

Similarly, for PR-AUC, the stacked ensemble reached 0.96 (95% CI: 0.95-0.97), while the base classifiers achieved between 0.95 (95% CI: 0.94-0.96). Statistical comparisons reinforced that the ensemble model consistently delivered superior precision-recall trade-offs ($p < 0.05$). These results demonstrate that the stacked framework's improvements are not only practically meaningful but also statistically significant, thereby strengthening the reliability of the proposed approach for predicting hypertension risk.

The stacked ensemble achieved a mean ROC-AUC of 0.931 (95% CI: 0.928-0.935) across 5-fold stratified cross-validation, while individual base models performed slightly lower: SVM (0.917; 95% CI: 0.913-0.922), KNN (0.919; 95% CI: 0.915-0.924), Naïve Bayes (0.921; 95% CI: 0.917-0.926), and MLP (0.918; 95% CI: 0.914-0.922). Pairwise comparisons using the DeLong test confirmed that the stacked ensemble significantly outperformed each of the base models ($p < 0.05$). Similarly, the PR-AUC of the stacked model (0.961; 95% CI: 0.957-0.965) was higher than that of the base classifiers (ranging from 0.949 to 0.953), with statistical tests again supporting the superiority of the ensemble ($p < 0.05$).

4.3 *Precision-Recall curve comparison*

Figure 5 shows the Precision-Recall (PR) curve comparison of the five classifiers-Stacked LGBM, SVM, KNN, Naive Bayes, and MLP-emphasising how they handle class imbalance. The Stacked LGBM model recorded the highest Area Under the Curve (AUC) of 0.96. This better precision-recall tradeoff is critical in medical diagnostics to prevent underprediction (false negatives) without lowering prediction confidence. The other models-SVM, KNN, Naive Bayes, and MLP-also performed with AUCs of 0.95, which was good but slightly lower than that of the ensemble model. The Stacked LGBM curve was interesting because it had higher precision values at almost all recall thresholds, affirming its reliability in detecting true hypertensive cases and minimising the misclassifications. This makes it even stronger when applied to high-stakes health risk prediction tasks.

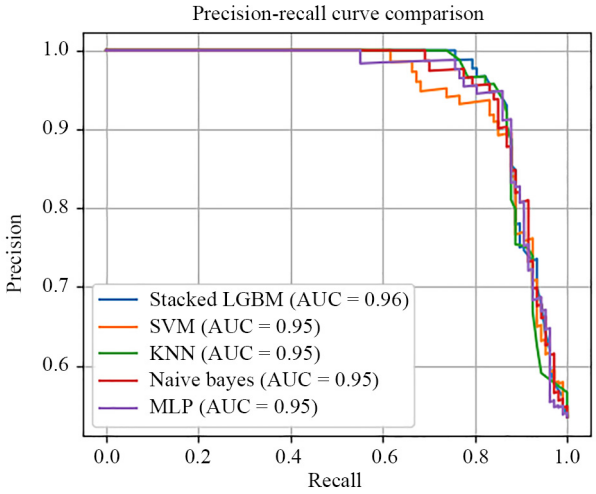

Figure 5. Precision-Recall curve comparison

Table 2. Performance comparison of the stacked ensemble and base classifier

Model	Accuracy	Precision	Recall	F1-score	ROC-AUC	PR-AUC
Stacked LGBM	0.91	0.92	0.86	0.89	0.93	0.96
SVM	0.89	0.90	0.84	0.87	0.92	0.95
KNN	0.90	0.91	0.85	0.88	0.92	0.95
GNB	0.88	0.89	0.83	0.86	0.93	0.95
MLP	0.90	0.91	0.85	0.88	0.92	0.95

The results presented in Table 2 highlight the consistent superiority of the stacked LGBM model compared to individual base classifiers across all evaluation metrics. The stacked framework achieved the highest accuracy (0.91), precision (0.92), and F1-score (0.89), reflecting its ability to maintain a strong balance between sensitivity and specificity. Importantly, its recall of 0.86 ensures robust detection of true hypertensive cases, which is critical for minimising false negatives in clinical applications. In terms of discrimination and reliability, the stacked model also outperformed others with an ROC-AUC of 0.93 and a PR-AUC of 0.96, both of which were higher than those of the standalone models. While SVM, KNN, and MLP demonstrated competitive performance with accuracies ranging from 0.89 to 0.90, and Naïve Bayes matched the ensemble in terms of ROC-AUC, none achieved the same level of consistent balance across all metrics. These findings underscore the advantage of integrating diverse learners into an optimised stacked ensemble, which not only improves predictive accuracy but also enhances clinical reliability for hypertension risk prediction.

4.4 *SHAP explainability analysis*

Figures 6-8 provide SHAP visualisations that highlight the stacked ensemble model's interpretability by demonstrating how much of each base classifier's probability output goes into the meta-learner's decision.

Figures 6a and 6b depict the SHAP summary plot in bar chart form. Among the meta-features, KNN_prob shows the highest average SHAP value (over 5.2), indicating that the KNN base model's output has the most decisive influence on the final prediction by the LightGBM meta-learner. This is followed by MLP_prob, which contributes moderately (approx. 1.7), while NB_prob and SVM_prob show negligible influence. This ranking suggests that although multiple base models are integrated, the stacked model heavily relies on KNN and MLP for final decision-making.

Figure 7a and Figure 7b present SHAP beeswarm plots, which capture the distribution and direction of SHAP values for each meta-feature across individual samples. The KNN-derived probabilities (blue and red gradient for low and high feature values) consistently exhibit the most variation and extremity in SHAP impact, both positively and negatively, indicating that KNN outputs frequently drive the model's confidence in class predictions. The MLP probabilities also show consistent yet less impactful contributions, while SVM and Naive Bayes are again shown to have minor influence.

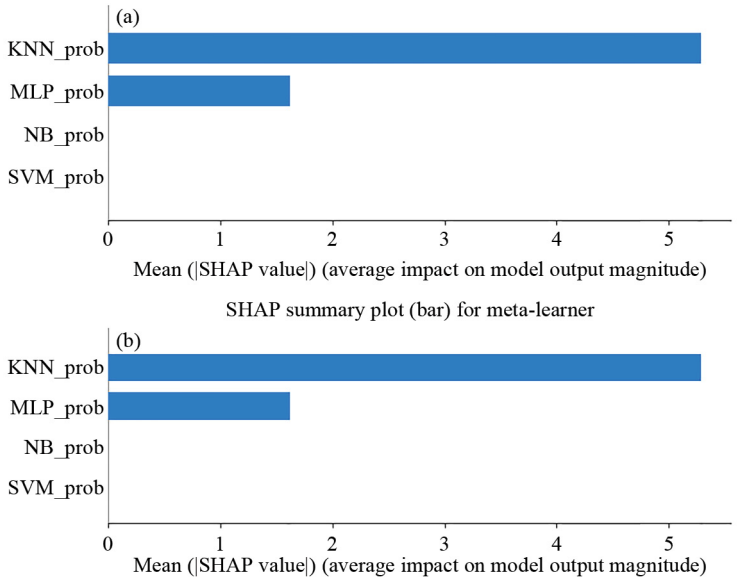

Figure 6. SHAP summary bar plot for meta-learner

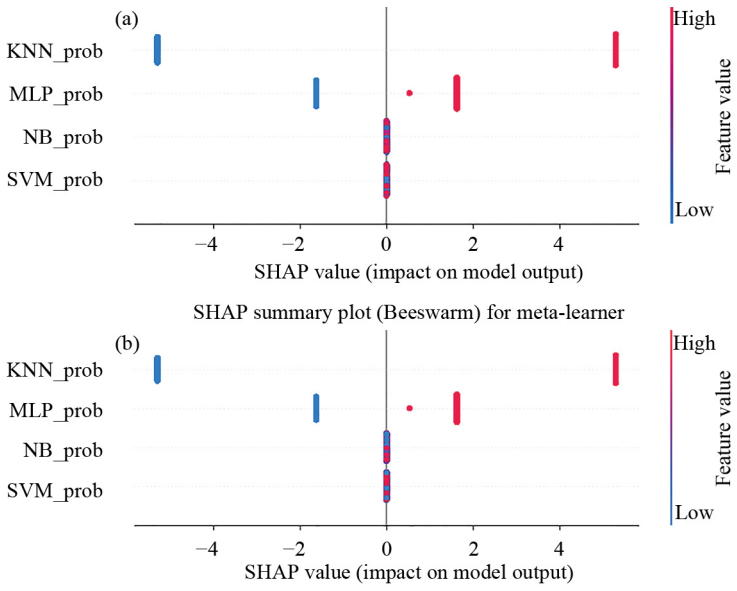

Figure 7. SHAP beeswarm plot for meta-learner

Figure 8 shows a SHAP force plot for a single test instance, offering a granular explanation of how the meta-learner (LightGBM) arrived at its final prediction. The model's output value is significantly pushed downward to −6.82 from

the base value due to the influence of the input features. In this case, both contributing meta-features-KNN_prob = 0.0 and MLP_prob ≈ 0.00043-strongly affect the prediction, signalling the model's confident classification of the instance into the non-hypertensive class. The absence of a positive signal from the KNN output and the minimal contribution from the MLP output drove the prediction far from the base value. This visualisation demonstrates how even small meta-feature outputs can decisively influence ensemble decisions, thus supporting individualised model transparency and clinical interpretability.

In summary, these SHAP-based explainability outputs validate that the stacked model is effective and transparent. It leverages KNN and MLP as primary decision signals while maintaining interpretability at the population and individual patient levels-an imperative feature for trust and adoption in healthcare applications.

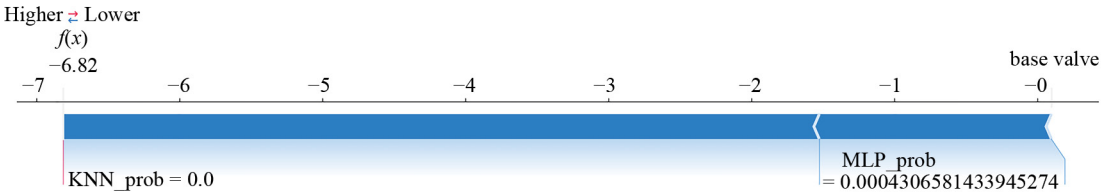

Figure 8. SHAP force plot for a single test instance

4.5 *Radar chart of evaluation metrics*

This radar chart gives a comparative visualisation of five models-Stacked LGBM, SVM, KNN, Naive Bayes, and MLP-across five key evaluation metrics: Accuracy, Precision, Recall, F1-Score, and ROC-AUC. The Stacked LGBM model is superior and consistently plots the largest area, indicating that it performs better and is more balanced across all scores. While the others' curves are in very close sequential order, the slight differences suggest that Naive Bayes trails behind slightly in F1-Score and Recall, but MLP and KNN both have superb overall profiles. This plot illustrates the ensemble method's ability to generalise and its stability as opposed to standalone classifiers. Figure 9 shows the radar chart for the performance metrics used to evaluate the five models implemented.

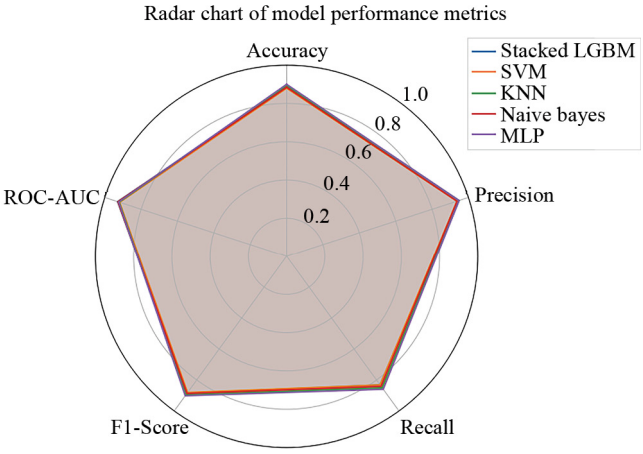

Figure 9. Radar chart of model performance metrics

5. Discussion

This study confirms the performance of a stacked ensemble learning model in predicting risk for hypertension, utilising SVM, KNN, Naive Bayes, and MLP as meta-learning learners. LightGBM was the meta-learner applied. Quantitatively, the Stacked LGBM model presented was effective in classification with an AUC of 0.93 on the ROC curve and 0.96 on the precision-recall curve. In contrast, individual base models ranged between 0.92 and 0.95 for both curves. Confusion matrix testing indicated that the MLP model identified most of the true positives (92), while the Stacked LGBM model recorded a high sensitivity but low specificity, with 89 true positives and only 5 false positives.

These results agree with previous research that has supported the superiority of ensemble models in medical diagnosis. You et al. [18], for instance, applied a particle swarm-optimised SVM to predict the risk of hypertension and obtained higher values of AUC than single classifiers. Similarly, Du et al. [5] employed a machine learning-based system for hypertension visualisation, highlighting the use of including interpretable features for clinical applications. Our work builds further on these elements by including SHapley Additive exPlanations (SHAP) to enhance model explainability. The SHAP values revealed that KNN_prob had the most significant average SHAP value (> 5.2), followed by MLP_prob, which impacted the ultimate decision of the ensemble. This explainability feature is critical in developing clinician trust and aligns with the prevailing trends that demand interpretable AI in healthcare.

Clinical early and accurate detection of hypertension must occur to prevent downstream effects such as stroke, renal failure, and cardiovascular events. Traditional screening approaches rely primarily on static thresholds and clinician-based measurements, which might miss early-risk markers. However, the proposed model can be employed as a decision-support system, integrated within Electronic Health Records (EHRs) to alert healthcare workers about high-risk patients based on dynamic and multi-parametric input. Moreover, its high recall (0.86) ensures that most hypertensive cases are captured, minimising false negatives-a crucial requirement for preventative intervention.

The model's lightweight architecture and scalability make it suitable for deployment in low-resource settings, including primary care clinics and mobile health platforms. The integration of explainability through SHAP ensures that predictions are not made in a black-box manner, thereby increasing their acceptability in regulated clinical environments. Furthermore, the SHAP force plots provide instance-level justifications, which are especially valuable in shared decision-making with patients. This study presents a highly accurate, interpretable, and clinically relevant framework for predicting the risk of hypertension. It advances the technical state-of-the-art and bridges the gap between algorithmic performance and real-world medical utility.

Although the proposed stacked ensemble framework demonstrated strong performance and interpretability, several limitations should be acknowledged. First, the dataset used in this study, while comprehensive in terms of clinical and demographic features, was limited to 26,083 records, which may constrain its generalizability to broader and more diverse populations. Second, the model relied on synthetic resampling to address class imbalance. While this approach improved sensitivity and overall classification performance, it may also introduce bias or reduce robustness when applied to real-world clinical datasets. Third, the study focused on structured clinical variables only, without incorporating additional modalities such as imaging, genomic data, or continuous sensor-based health records, which could further enrich predictive capability. These limitations suggest that although the model is promising for hypertension risk prediction, further validation on larger, multi-center datasets and exploration of alternative imbalance handling methods are necessary before deployment in clinical settings.

5.1 *Limitations and future works*

Despite the promising performance of the proposed stacked ensemble framework, several limitations should be acknowledged. First, the dataset used in this study, although containing diverse demographic and clinical parameters, remains relatively modest in size, which may limit generalizability to broader populations. Second, the class imbalance was handled using random upsampling, which, while effective in this case, may introduce synthetic bias. More sophisticated resampling strategies such as SMOTE, ADASYN, or hybrid approaches could be explored in future work to enhance robustness. Third, although Optuna-based optimisation significantly improved hyperparameter tuning efficiency, additional exploration of multi-objective optimisation strategies could provide a better balance between

accuracy, interpretability, and computational cost. Finally, the present study focused solely on structured clinical features. Future research could incorporate multimodal data sources, such as genomic, imaging, or wearable sensor data, to provide more comprehensive patient profiles and potentially enhance predictive performance.

6. Conclusion

This paper presents a mathematically tuned and interpretable ensemble learning approach to predict hypertension risk using combinations of four base classifiers-SVM, KNN, Naive Bayes, and MLP-combined with a LightGBM meta-learner. The stacked model performed better in a detailed evaluation, achieving an AUC of 0.93 on the ROC curve and 0.96 on the precision-recall curve. Notably, the model achieved a high classification balance and high sensitivity, with a recall of 0.86, and high specificity, making it clinically credible for detecting hypertension at an early stage. SHAP explainability techniques also improved the model's transparency, identifying KNN_prob and MLP_prob as the most important features. Such transparency supports model accountability and enables real-world integration into clinical decision-making. The lightweight, scalable architecture of the proposed system makes it well-suited for deployment in both high- and low-resource healthcare settings. Overall, the study presents a robust, accurate, and interpretable AI-powered instrument for hypertension risk stratification, promising to significantly enhance public health screening and intervention efforts.

Conflict of interest

The authors declare that they have no conflict of interest.

References

[1] World Health Organization. *Hypertension*. Available from: https://www.who.int/news-room/fact-sheets/detail/hypertension [Accessed 24th August 2023].

[2] Kario K, Okura A, Hoshide S, Mogi M. The WHO global report 2023 on hypertension warning the emerging hypertension burden in globe and its treatment strategy. *Hypertension Research*. 2024; 47(5): 1099-1102. Available from: https://doi.org/10.1038/s41440-024-01622-w.

[3] Ogundokun RO, Misra S, Umoru D, Agrawal A. Review of cardiovascular disease prediction based on machine learning algorithms. In: *The International Conference on Recent Innovations in Computing*. Singapore: Springer; 2022. p.37-50. Available from: https://doi.org/10.1007/978-981-99-0601-7_4.

[4] Fryar CD, Kit B, Carroll MD, Afful J. Hypertension prevalence, awareness, treatment, and control among adults ages 18 and older: United States, August 2021-August 2023. *National Center for Health Statistics Data Brief*. 2024; 511: CS354233.

[5] Du J, Chang X, Ye C, Zeng Y, Yang S, Wu S, et al. Developing a hypertension visualization risk prediction system utilizing machine learning and health check-up data. *Scientific Reports*. 2023; 13(1): 18953. Available from: https://doi.org/10.1038/s41598-023-45931-5.

[6] Islam MM, Alam MJ, Maniruzzaman M, Ahmed NF, Ali MS, Rahman MJ, et al. Predicting the risk of hypertension using machine learning algorithms: A cross sectional study in Ethiopia. *PLOS One*. 2023; 18(8): e0289613. Available from: https://doi.org/10.1371/journal.pone.0289613.

[7] Ganz M, Alessandro C, Jacobs M, Miller D, Diah J, Winer A, et al. The role of body mass index and waist circumference in gender-specific risk factors for stress urinary incontinence: A cross-sectional study. *Cureus*. 2023; 15(5): e38917. Available from: https://doi.org/10.7759/cureus.38917.

[8] Zhong X, Yu J, Jiang F, Chen H, Wang Z, Teng J, et al. A risk prediction model based on machine learning for early cognitive impairment in hypertension: Development and validation study. *Frontiers in Public Health*. 2023; 11: 1143019. Available from: https://doi.org/10.3389/fpubh.2023.1143019.

[9] Ogundokun RO, Owolawi PA, Tu C. Optimized deep feature learning with hybrid ensemble soft voting for early breast cancer histopathological image classification. *Computers, Materials and Continua*. 2025; 84(3): 4869-4885. Available from: https://doi.org/10.32604/cmc.2025.064944.

[10] Araujo-Moura K, Souza L, De Oliveira TA, Rocha MS, De Moraes ACF, Chiavegatto Filho A. Prediction of hypertension in the pediatric population using machine learning and transfer learning: A multicentric analysis of the SAYCARE study. *International Journal of Public Health*. 2025; 70: 1607944. Available from: https://doi.org/10.3389/ijph.2025.1607944.

[11] Lundberg SM, Lee SI. A unified approach to interpreting model predictions. *arXiv:1705.07874*. 2017. Available from: https://doi.org/10.48550/arXiv.1705.07874.

[12] Adelodun AB, Ogundokun RO, Yekini AO, Awotunde JB, Timothy CC. Explainable artificial intelligence with scaling techniques to classify breast cancer images. In: *Explainable Machine Learning for Multimedia Based Healthcare Applications*. Heidelberg: Springer; 2023. p.99-137. Available from: https://doi.org/10.1007/978-3-031-38036-5_6.

[13] Akiba T, Sano S, Yanase T, Ohta T, Koyama M. Optuna: A next-generation hyperparameter optimization framework. In: *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. Anchorage, AK, USA: Association for Computing Machinery; 2019. p.2623-2631. Available from: https://doi.org/10.1145/3292500.3330701.

[14] Islam MM, Alam MJ, Maniruzzaman M, Ahmed NF, Ali MS, Rahman MJ, et al. Predicting the risk of hypertension using machine learning algorithms: A cross sectional study in Ethiopia. *PLOS One*. 2023; 18(8): e0289613. Available from: https://doi.org/10.1371/journal.pone.0289613.

[15] Effati S, Kamarzardi-Torghabe A, Azizi-Froutaghe F, Atighi I, Ghiasi-Hafez S. Web application using machine learning to predict cardiovascular disease and hypertension in mine workers. *Scientific Reports*. 2024; 14(1): 31662. Available from: https://doi.org/10.1038/s41598-024-82221-0.

[16] Oh SH, Lee SJ, Park J. Precision medicine for hypertension patients with type 2 diabetes via reinforcement learning. *Journal of Personalized Medicine*. 2022; 12(1): 87. Available from: https://doi.org/10.3390/jpm12010087.

[17] Estiko RI, Rijanto E, Juwana YB, Juzar DA, Widyantoro B. Hypertension prediction models using machine learning with easy-to-collect risk factors: A systematic review. *Journal of Hypertension*. 2024; 42(Suppl 2): e19. Available from: https://doi.org/10.1097/01.hjh.0001027072.19895.81.

[18] You R, Tao Q, Wang S, Cao L, Zeng K, Lin J, et al. Development and validation of a hypertension risk prediction model based on particle swarm optimization-support vector machine. *Bioengineering*. 2025; 12(3): 238. Available from: https://doi.org/10.3390/bioengineering12030238.