Gray Langurs Optimized Hybrid Learning for Heart Disease Detection

Abstract

Heart disease remains a leading cause of death worldwide, and screening from routine clinical attributes can support earlier treatment. Existing detection models for the Cleveland heart disease records often report accuracy from single data partitions, use the complete attribute set and rarely compare competing methods under identical folds. To address these issues, this study proposes ShPCN-Net, an optimization-based feature selection and hybrid learning framework. Clinical records are cleaned and normalized with a Gaussian function, after which the Gray Langurs Optimizer selects a compact attribute subset through an accuracy and compactness fitness. A ShuffleNet-V2 branch with squeeze-and-excitation recalibration and a regression branch then process the selected attributes, and their outputs are combined through a first-order Taylor-series fusion and classified by a multi-scale parallel convolutional network. The framework was evaluated on the 303 Cleveland records with stratified 10-fold cross-validation and compared with three published baselines and two ablation variants under the same folds. ShPCN-Net retained 5.7 of the 13 attributes on average and achieved an accuracy of 81.15%, a Matthews correlation coefficient of 0.630 and a true positive rate of 77.03%. In comparison, the baselines reached accuracies between 79.86% and 82.83%, and paired tests found no significant difference between ShPCN-Net and any baseline (all $p > 0.21$). Shapley additive explanations identified the number of major vessels, thallium scan result, ST slope and chest pain type as the dominant attributes. These findings indicate that compact optimized attribute subsets preserve detection performance, while larger cohorts are needed to establish an accuracy advantage.

Keywords: Heart Disease Detection; Feature Selection; Gray Langurs Optimizer; Hybrid Deep Learning; Explainable Artificial Intelligence

1. Introduction

Cardiovascular diseases remain the leading cause of morbidity and mortality worldwide, and heart disease accounts for a large share of this burden [1, 2]. Early identification of patients at risk allows clinicians to begin preventive treatment before irreversible cardiac damage occurs [3, 4]. Routine clinical examinations already record attributes such as age, blood pressure, serum cholesterol, chest pain type and exercise test results, which makes these records a practical basis for automated screening [5, 6]. Machine learning models can learn the relationship between such attributes and the presence of disease, thereby supporting clinical decisions in settings where specialist cardiologists are scarce [7, 8]. Consequently, the design of accurate and transparent heart disease detection models has become an active research area.

However, the diagnosis of heart disease from clinical attributes is complicated by the heterogeneity of the recorded variables and by the limited size of the available cohorts [9, 10]. The Cleveland heart disease records, which remain the most widely used public benchmark for this task, contain only 303 patients described by 13 numerical and categorical attributes [7, 11]. Such data sets combine continuous measurements with ordinal codes, and several attributes carry redundant or weakly related information [4, 12]. In addition, delayed or incorrect diagnosis has severe consequences, since a missed positive case may delay treatment of a patient with obstructive coronary disease [13, 14]. These characteristics make reliable feature selection, careful evaluation and interpretable predictions essential for clinical heart disease detection.

Researchers have developed a broad range of detection methods, including support vector machines, random forests, gradient boosting and ensemble learners [3–6, 12]. Deep and hybrid architectures, such

as convolutional networks combined with bidirectional recurrent layers, autoencoders with densely connected networks and ensembles of convolutional and recurrent models, have also been reported [9, 10, 15, 16]. However, these methods exhibit important limitations. Several studies report accuracy from a single hold-out partition or even from training data, which makes the estimates sensitive to the particular split [5, 7, 9]. Many pipelines use all available attributes, and comparisons with competing methods are frequently drawn from different publications with incompatible protocols [2, 17]. Consequently, there is a need for a detection framework that selects a compact attribute subset, evaluates every competing method under identical folds and explains its decisions.

To address these challenges, this study proposes ShPCN-Net, an optimization-based feature selection and hybrid learning framework for heart disease detection. Its first stage cleans and normalizes the clinical records and then applies the Gray Langurs Optimizer [18] as a wrapper feature selector that balances classification accuracy against subset size [19]. The selected attributes are processed by a lightweight ShuffleNet-V2 branch with squeeze-and-excitation recalibration [20] and by a regression branch, and a first-order Taylor-series rule fuses the two representations. A multi-scale parallel convolutional network [21] then maps the fused representation to the final disease probability. Stratified cross-validation, data-fraction analysis, paired statistical tests and Shapley additive explanations are used to evaluate the framework [1, 8].

The main contributions of this study are as follows:

1) A clinical preprocessing and Gray Langurs Optimizer wrapper pipeline is designed to select compact attribute subsets from the Cleveland records using an accuracy and compactness fitness.
2) ShPCN-Net is proposed as a hybrid network that combines ShuffleNet-V2 units, squeeze-and-excitation recalibration, a regression branch with Taylor-series fusion and multi-scale parallel convolution for tabular clinical data.
3) Three published baselines and two ablation variants are re-implemented and compared with the proposed framework under identical stratified folds, data fractions and paired statistical tests.
4) Shapley additive explanations are applied to the trained network to identify the clinical attributes that drive individual predictions.

The rest of this paper is organized as follows. Section 2 reviews related studies and states the research gaps that motivate this work. Section 3 presents the proposed ShPCN-Net methodology, including preprocessing, Gray Langurs Optimizer feature selection and the hybrid classification network. The experimental procedure and the results of the proposed framework and the competing methods are reported in Section 4. Section 5 discusses the findings in relation to previous studies and outlines their limitations. Finally, Section 6 concludes the paper and identifies directions for future research.

2. Motivation

Heart disease detection from routine clinical attributes is valuable only when the reported performance reflects how a model behaves on unseen patients and when clinicians can understand which attributes drive a decision. Published studies on this task differ widely in their data partitions, attribute sets and explanation methods, which makes it difficult to judge whether a new architecture offers a genuine improvement. This section reviews the relevant literature and summarizes the research gaps that motivate the proposed framework.

2.1 Literature Survey

Several studies have applied conventional machine learning classifiers to clinical heart disease records. Nashif et al. [3] compared several classifiers in the WEKA platform on two open-access databases, and a support vector machine reached 97.53% accuracy with 97.50% sensitivity and 94.94% specificity

under 10-fold cross-validation. However, the study focused on classifier accuracy and a cloud-based monitoring system, without explaining which attributes determined individual predictions. Chang et al. [5] developed a Python-based detection application that combined logistic regression with a random forest classifier and reported approximately 83% accuracy. Because this value was obtained over the training data, it does not indicate how the model generalizes to unseen patients. Nagavelli et al. [6] examined four machine learning approaches, including a weighted naive Bayes model, support vector machines combined with XGBoost and an XGBoost model with density-based outlier removal and SMOTE-ENN balancing, where XGBoost reached 95.9% accuracy. Nevertheless, the analysis did not report feature-level explanations of the predictions. Ketu and Mishra [14] investigated eight classifiers on an imbalanced electrocardiogram-based arrhythmia dataset and showed that SMOTE balancing improved both overall and minority-class accuracy. However, the analysis was limited to conventional classifiers and did not consider deep or hybrid architectures. Although these studies establish useful classifier baselines, they rarely combine held-out evaluation, compact attribute subsets and explanation within one framework.

Feature selection and hyperparameter optimization have been introduced to improve these conventional pipelines. Javeed et al. [7] employed a random search algorithm for feature selection with a grid-search-optimized random forest on the Cleveland records, obtaining 93.33% accuracy with only seven attributes, which was 3.3% higher than a conventional random forest. However, this result was derived from a single 70/30 hold-out partition, so the estimate may change with the particular split. Abdellatif et al. [2] combined SMOTE balancing, six classifiers and hyperparameter optimization, and an Extra Trees model tuned with hyperband achieved 99.2% and 98.52% detection accuracy on two public datasets together with 95.73% severity classification on the Cleveland records. Nevertheless, the model was built and tested with all attributes, without a subset selection stage that could reduce the clinical input burden. Tama et al. [4] designed a two-tier ensemble that stacked random forest, gradient boosting and XGBoost learners and used particle swarm optimization to select attributes for four datasets, including Cleveland, under 10-fold cross-validation with a two-step statistical test. However, the stacked ensemble-of-ensembles structure did not provide feature-level explanations of individual decisions. These studies show that search-based selection can reduce the attribute set, yet recent multi-group bio-inspired optimizers have not been examined for this task.

Deep, hybrid and explainable architectures form the third group of related studies. Ramesh and Lakshmanna [13] introduced an optimal scrutiny boosted graph convolutional LSTM tuned by the Eurygaster optimization algorithm, together with a fuzzy inference system, for early detection and prevention of coronary heart disease in diabetic patients. The authors summarized the performance as exceeding 98% accuracy in most studies; however, the framework targeted diabetes-related complications and was not assessed under a single cross-validated protocol on a public clinical benchmark. Khan et al. [15] presented EnsCVDD-Net, an ensemble of LeNet and a gated recurrent unit, and BlCVDD-Net, which blends LeNet, a gated recurrent unit and a multilayer perceptron, with adaptive synthetic sampling, point-biserial feature selection and Shapley additive explanations. BlCVDD-Net reached 91% accuracy under 10-fold cross-validation, but the reported execution times of 247 s for BlCVDD-Net and 777 s for EnsCVDD-Net indicate a considerable computational cost. Shorewala [17] compared bagging, boosting and stacking with K-nearest neighbors, logistic classification and naive Bayes on a 70,000-record cardiovascular dataset, where a stacked model achieved 75.1% accuracy. However, bagging improved accuracy by only 1.96% on average over the corresponding base learners, which indicates limited gains from ensembling alone. Although these architectures increase model capacity, few of them pair a compact optimized attribute subset with explanation and a statistically tested comparison under identical folds.

2.2 Research Gap

The reviewed studies reveal the following research gaps, which define the scope of the proposed framework:

- Several Cleveland studies report accuracy from a single hold-out split or from training data, which makes the published estimates sensitive to the chosen partition [5, 7, 9].
- Most detection pipelines use the complete attribute set or simple filter ranking, and recent multi-group metaheuristics have not been applied to clinical attribute selection [2, 18].
- Lightweight channel-attention networks and multi-scale parallel convolution have rarely been combined for tabular clinical records, leaving their joint value for small cohorts unexamined [20, 21].
- Many studies lack feature-level explanations of their predictions, which limits clinical trust in the learned decision rules [4, 6].
- Comparisons with competing methods are commonly drawn from different publications without paired statistical tests on identical folds, so reported improvements cannot be verified [1, 17].

Table 1 maps each gap to the component of the proposed framework that addresses it and to the evidence reported in Section 4. The mapping distinguishes gaps that the evaluation resolves from gaps that remain only partially addressed, since the hybrid architecture did not improve accuracy over the strongest baselines.

Table 1. Mapping of research gaps to framework components and supporting evidence.

Research gap	Framework component	Evidence in this study	Status
Single-split or training-set estimates	Stratified 10-fold protocol and 50–90% data-fraction sweep	Table 5 and Table 6; Fig. 18 to Fig. 22	Addressed
Complete attribute sets	Gray Langurs Optimizer wrapper selection	5.7 of 13 attributes retained on average (Fig. 8)	Addressed
Unexamined hybrid lightweight design	ShPCN-Net with SE recalibration, Taylor fusion and PCNN	81.15% accuracy, not significantly different from baselines	Partially addressed
Missing explanations	Shapley additive explanations	Fig. 26 and Fig. 27	Addressed
Untested comparisons	Paired t-test and Wilcoxon test on identical folds	Table 7; Fig. 28	Addressed

3. Proposed ShPCN-Net Methodology

This section presents the proposed ShPCN-Net framework for heart disease detection from clinical records. Fig. 1 illustrates the overall workflow, which consists of four stages: clinical data acquisition, record preprocessing, Gray Langurs Optimizer (GLO) feature selection and hybrid classification. Each input record contains 13 clinical attributes, and the preprocessing stage removes duplicate records, imputes missing values and applies Gaussian-function normalization. Next, the GLO searches for a compact attribute subset that preserves classification accuracy. The selected attributes enter the ShPCN-Net classifier, in which a ShuffleNet-V2 branch with squeeze-and-excitation (SE) recalibration and a regression branch produce two complementary representations. A first-order Taylor-series fusion combines these representations, and a multi-scale parallel convolutional neural network (PCNN) produces the final probabilities for the Normal and Disease classes. Every stage is fitted on training records only, and the held-out records are used solely for evaluation.

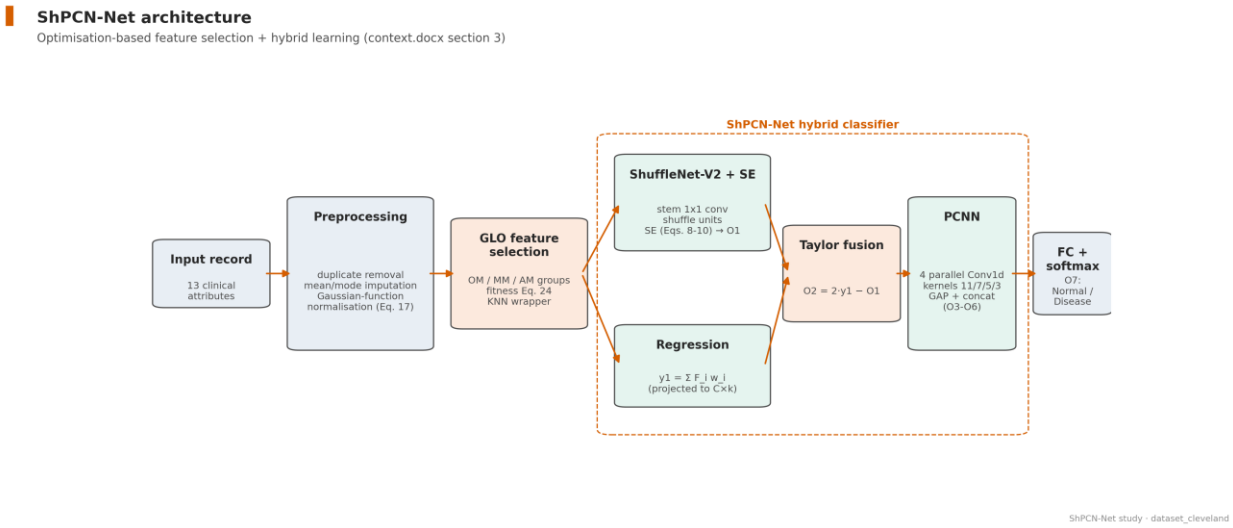


Fig. 1. Block diagram of the proposed ShPCN-Net framework, from the input record through preprocessing, GLO feature selection and the hybrid classifier.

The framework name reflects its three learning components: the ShuffleNet-V2 branch (Sh), the parallel convolutional network (PCN) and the network-level fusion that joins them. This design pursues two goals that arise from the characteristics of the Cleveland records. First, the attribute set is reduced to the attributes that carry diagnostic information, because redundant inputs increase the risk of overfitting a 303-record cohort. Second, nonlinear interactions among the selected attributes are learned at several receptive-field sizes while keeping the number of trainable parameters small. Each stage is described below in execution order.

3.1 Cleveland Clinical Data Acquisition and Problem Formulation

The framework operates on the Cleveland heart disease records from the UCI Machine Learning Repository [11]. Each record describes one patient by 13 attributes. The demographic and resting attributes are age, sex, chest pain type (cp), resting blood pressure (trestbps), serum cholesterol (chol), fasting blood sugar (fbs) and resting electrocardiographic result (restecg). Exercise and imaging attributes comprise maximum heart rate (thalach), exercise-induced angina (exang), ST depression (oldpeak), slope of the peak exercise ST segment (slope), number of major vessels coloured by fluoroscopy (ca) and thallium scan result (thal). A binary target variable indicates the absence (Normal) or presence (Disease) of heart disease. Let the dataset be denoted as follows:

$$D = \{d_j = (x_j, y_j)\},\ x_j \in \mathbb{R}^K,\ y_j \in \{0,1\},\ j = 1, \ldots, N \tag{1}$$

where x_j denotes the attribute vector of the j-th patient, y_j indicates the diagnostic label, $K = 13$ is the number of attributes and $N = 303$ is the number of records. The detection task is formulated as a binary classification problem, where the objective is to learn a mapping f: $\mathrm{R}^{\mathrm{K}} \rightarrow [0, 1]$ that estimates the probability of disease from a selected subset of the attributes. The five continuous attributes (age, trestbps, chol, thalach and oldpeak) and the eight categorical attributes (sex, cp, fbs, restecg, exang, slope, ca and thal) are treated differently during preprocessing, as described next.

3.2 Clinical Record Preprocessing

Clinical records frequently contain repeated entries, missing measurements and attributes with very different numerical ranges. The preprocessing stage addresses these issues in three steps, namely duplicate removal, missing-value imputation and Gaussian-function normalization. The resulting cleaned matrix is denoted M_v after imputation and N_d after normalization, following the stage notation in Fig. 1.

3.2.1 Duplicate Removal and Mean–Mode Imputation

Duplicate records are removed once at dataset level, before any partition is formed. This order guarantees that an identical patient record cannot appear simultaneously in a training fold and in the corresponding test fold, which would otherwise inflate the evaluation. Missing values are then replaced by the mean of a continuous attribute or by the mode of a categorical attribute, following the imputation rule adopted by Anwer et al. [22]:

$$x_{filled} = \{ \mu_x, \ if\ x\ is\ numerical; \quad mode(x), \ if\ x\ is\ categorical \tag{2}$$

where μ_x represents the mean of attribute x and mode(x) is its most frequent value. The imputation statistics are computed on the training portion of each split and then applied to the held-out records. Mean imputation preserves the central tendency of continuous measurements, while mode imputation keeps categorical codes within their valid set, which avoids fractional codes that have no clinical meaning.

3.2.2 Gaussian-Function Normalization of Continuous Attributes

The continuous attributes differ widely in scale; for example, serum cholesterol is measured in milligrams per decilitre, whereas ST depression is measured in millimetres. The framework therefore applies the Gaussian probability density function as a normalization transform, as investigated by Kim et al. [23]. The normalized value of a continuous attribute is calculated as follows:

$$x' = \frac{1}{\sigma_x\sqrt{2\pi}}\ exp\left(-\frac{(x-\mu_x)^2}{2{\sigma_x}^2}\right) \tag{3}$$

where μ_x and σ_x denote the mean and standard deviation of attribute x on the training records. Because the peak value $1/(\sigma_x\sqrt{(2\pi)})$ depends on the spread of each attribute, an attribute with a large standard deviation such as cholesterol would otherwise span only a narrow range near zero. The framework therefore rescales the output by its constant peak, as in Eq. (4):

$$\tilde{x} = \sigma_x\sqrt{2\pi} \cdot x' = exp\left(-\frac{(x-\mu_x)^2}{2{\sigma_x}^2}\right) \tag{4}$$

where $\tilde{x}$ lies in the interval (0, 1] for every continuous attribute. This rescaling preserves the shape of the Gaussian transform while giving the continuous attributes the same numerical range as the categorical codes, which are min–max scaled to [0, 1]. The transform maps values close to the training mean to values near one and atypical values towards zero, so it emphasizes how unusual a measurement is relative to the cohort. However, the transform is symmetric about the mean, and values equally far above and below the mean receive the same normalized value. The selected subsets reported in Section 4 indicate that the optimizer favours categorical attributes, which are unaffected by this symmetry.

3.3 Gray Langurs Optimizer-Based Feature Subset Selection

The feature selection stage searches for the attribute subset F_s that maximizes classification accuracy while minimizing the number of retained attributes. An exhaustive search is feasible in principle for 13 attributes, but it becomes impractical inside repeated cross-validation, and wrapper metaheuristics offer a flexible alternative that can be applied unchanged to larger attribute sets. For this purpose, the framework adopts the GLO proposed by Barshandeh et al. [18], a multi-group bio-inspired algorithm that models the social organization of gray langurs into one-male, multi-male and all-male groups. This optimizer is selected for its explicit balance between exploration and exploitation, which is controlled by an adaptive scaling factor, Levy-flight steps, migration between groups and an autonomous roaming phase.

3.3.1 Solution Encoding and Subset Decoding

Each langur represents a candidate subset through a continuous position vector $X = (x_1, \ldots, x_K)$ in the search space $[lb, ub]^K = [0, 1]^K$. The position is decoded into a binary selection mask as follows:

$$s_i = \{ 1,\ if\ x_i > 0.5;\ \ 0,\ otherwise \},\ \ i = 1, \ldots, K \tag{5}$$

where $s_i = 1$ indicates that the i-th attribute is retained. The decoded subset is stored as the list of indices of the retained attributes, for example (6, 8, 4, 1, 2) for a five-attribute subset. If no position exceeds the threshold, the attribute with the largest position value is retained, so every candidate contains at least one attribute. This continuous-to-binary encoding allows the original GLO update rules to operate without modification.

3.3.2 Accuracy–Compactness Fitness Function

Every candidate subset is scored by the fitness function employed by Qiu et al. [19] for simultaneous feature selection, which the GLO minimizes:

$$Fitness = \alpha\,(1 - Accuracy) + (1 - \alpha)\,\frac{R}{K} \tag{6}$$

where Accuracy is the classification accuracy obtained with the candidate subset, R is the number of selected attributes, K is the total number of attributes and $\alpha = 0.99$ weighs accuracy against subset size. The accuracy term is computed by a K-nearest neighbour classifier with five neighbours under five-fold stratified cross-validation on the training records of the current split. This wrapper classifier is chosen because it is sensitive to irrelevant attributes, so its accuracy decreases when uninformative attributes are retained. In addition, the classifier is fast enough to evaluate thousands of candidates within each outer fold. Fitness values are memoized for each distinct mask, and the pairwise distances are precomputed per attribute, which reduces a complete GLO run to a few seconds on a laptop processor. With a population of 30 langurs and 100 iterations, each run proposes up to 6,000 candidate positions, because every langur is updated once by its group rule and once by the autonomy phase in each iteration. Many of these positions decode to identical masks, so the number of distinct wrapper evaluations is considerably smaller than the number of proposals. The memoization therefore keeps the cost of the selection stage nearly independent of the iteration budget once the search concentrates on a small region of the subset space.

3.3.3 Multi-Group Position Updates of the Gray Langurs Optimizer

The population of N_p langurs is divided into a one-male group, a multi-male group and an all-male group of equal initial size. Within each group, the langurs are ranked by fitness, and the best member becomes the alpha. Each alpha follows one of two behaviours chosen with equal probability. The first behaviour is roaming around the current position, which is expressed as follows:

$$X_\alpha(t + 1) = X_\alpha(t) + SR,\ \ SR = \alpha_r \cdot \beta \cdot L \tag{7}$$

where t is the current iteration, α_r is a random number in $(-5, 5)$, β controls the search range and L is a Levy-based step size. The search range and the step size are defined as follows:

$$\beta = ASF \cdot max(\,1 - \frac{t}{T}, 0.4\,) \cdot \frac{ub - lb}{30},\ \ L = \frac{1}{|\sigma|^{1/1.5}},\ \ ASF = 0.1 + 0.9\,(\,1 - \frac{t}{T}\,)^2 \tag{8}$$

where T is the maximum number of iterations, σ is a vector drawn from the standard normal distribution and ASF is the adaptive scaling factor, which decreases from 1.0 to 0.1 as the search progresses. The second behaviour moves the alpha relative to the best position found so far, as in Eq. (9):

$$X_\alpha(t + 1) = X_\alpha(t) + C_1\,(X_\alpha(t) - X_{GB}(t)),\ \ C_1 = (\,10 - \frac{t}{T}\,)\,\mu \tag{9}$$

where X_{GB} denotes the global best position and μ is a vector of random numbers in (−1, 1). Females follow their leader, which is the alpha in the one-male group and a mate chosen by roulette-wheel selection among the alpha and the males in the multi-male group:

$$X_f(t+1) = X_f(t) + C_2\,(X_{lead}(t) - X_f(t)),\quad C_2 = 0.8\,\sigma \cdot ASF \tag{10}$$

where X_f and X_{lead} represent the positions of a female and her leader. The number of females equals 40% of the one-male group, while the multi-male group contains 30% males and 40% females. Children inherit a combination of the alpha and a randomly selected mother, perturbed by a decreasing mutation term:

$$X_{ch}(t+1) = C_f\,X_\alpha(t) + C_m\,X_m(t) + C_3,\ C_f = \sigma \cdot ASF,\ C_m = 1 - C_f,\ C_3 = 0.05\,\sigma \cdot ASF\,(ub - lb) \tag{11}$$

where X_m is the position of the mother and C_f, C_m and C_3 are the paternal, maternal and mutation coefficients. Males in the multi-male and all-male groups move relative to the global best position with a large random perturbation, as given by Eq. (12):

$$X_M(t+1) = X_M(t) + \theta\,(X_M(t) - X_{GB}(t)) + C_4\,\sigma\,(ub - lb),\ \theta = max(0.4, 0.9\,(1 - \frac{t}{T})),\ C_4 = 1 + \sigma \tag{12}$$

where X_M denotes the position of a male and θ is a coefficient that decreases linearly to 0.4. The large perturbation of the males maintains population diversity, whereas the females and children concentrate the search around the leaders. This division of roles is the main reason for selecting the GLO, because it provides exploration and exploitation simultaneously within one population.

3.3.4 Dynamism, Autonomy and Greedy Acceptance

Two further mechanisms complete each iteration. The dynamism mechanism moves each langur to another group with probability ω = 0.05, provided that its current group retains at least φ = 5 members, so the group sizes adapt while the total population remains constant. The autonomy phase lets every langur roam within a radius that shrinks over time:

$$X_i(t+1) = X_i(t) \odot step,\ step = scale \cdot L \cdot 0.001,\ scale = r_2\,(ub - lb)\,e^{-20(t/T)^2} \tag{13}$$

where ⊙ denotes element-wise multiplication and r_2 is a uniform random number in (0, 1). After every update, positions are clipped to [0, 1], and a new position replaces the previous one only if it lowers the fitness of that langur. This greedy acceptance prevents the large exploratory moves from degrading the population, which is important because several update rules in Eq. (9) to Eq. (13) produce steps that exceed the search range. Algorithm 1 summarizes the complete feature selection procedure.

Algorithm 1. GLO-based attribute subset selection used inside each training split.

Input: normalized training records N_d, labels y, population N_p = 30, iterations T = 100, α = 0.99
Output: selected attribute subset F_s and convergence history
1: Initialize N_p positions uniformly in [0, 1]^K; split them into OM, MM and AM groups
2: Evaluate the fitness of every position with Eq. (6) using the KNN wrapper
3: for t = 1 to T do
4: Compute ASF; identify X_GB and the alpha of each group
5: Update each alpha with Eq. (7) or Eq. (9)
6: Update females with Eq. (10), children with Eq. (11) and males with Eq. (12)

7: Clip positions to [0, 1]; accept each new position only if its fitness decreases
8: Apply autonomy with Eq. (13) to every langur with greedy acceptance
9: Migrate each langur with probability ω = 0.05 if its group keeps ϕ = 5 members
10: Record the best fitness, mean fitness and subset size
11: end for
12: Decode X_GB with Eq. (5) and return F_s

3.4 ShPCN-Net Hybrid Classification Network

The ShPCN-Net classifier receives the k attributes selected by the GLO and estimates the probability of disease. Fig. 2 depicts its architecture, which contains four modules: a pointwise stem, a ShuffleNet-V2 branch with SE recalibration, a regression branch fused with the convolutional representation through a Taylor-series rule and a multi-scale PCNN followed by a classification head. In the selected configuration with five input attributes, the network contains 38,314 trainable parameters with five input attributes, which keeps the model small relative to the deep baselines described in Section 3.6. Each module is described below.

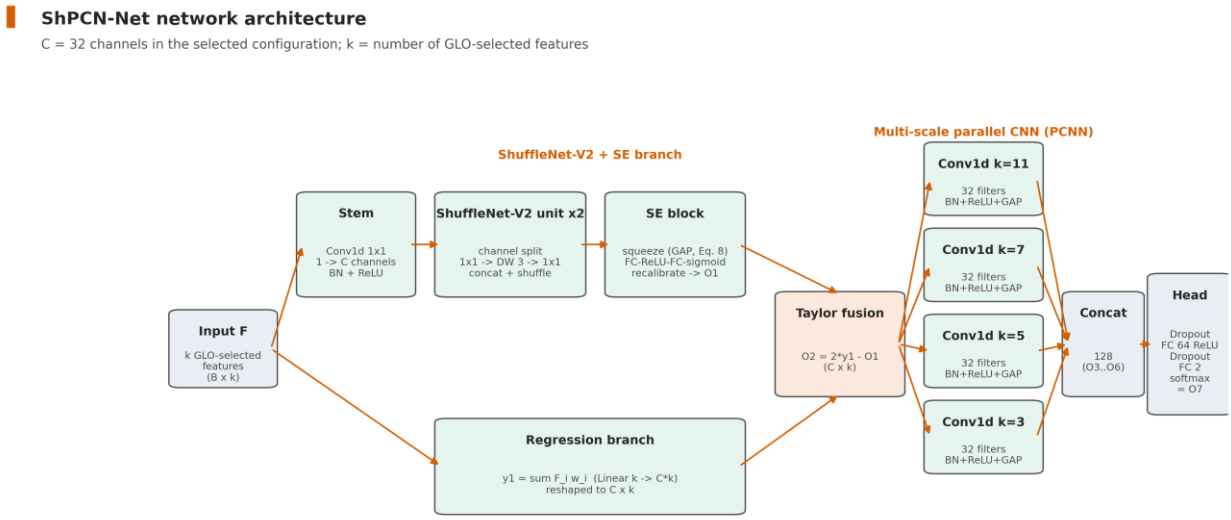


Fig. 2. Architecture of the ShPCN-Net classifier with the ShuffleNet-V2 and SE branch, the regression branch, the Taylor-series fusion and the multi-scale PCNN.

3.4.1 Tabular-to-Sequence Representation and Pointwise Stem

Convolutional architectures such as ShuffleNet-V2 and the multi-scale PCNN were originally designed for images [20, 21]. The framework adapts them to tabular records by treating the k selected attributes as a one-dimensional sequence of length k with a single channel, and all convolutions are implemented as one-dimensional operations with same padding. A pointwise convolution then lifts the single input channel to C feature channels, as expressed in Eq. (14):

$$H_0 = ReLU(BN(W_s * F)) \tag{14}$$

where $F \in \mathrm{R}^{1\times k}$ is the selected attribute sequence, W_s denotes the pointwise convolution kernel, BN represents batch normalization and $H_0 \in \mathrm{R}^{C\times k}$ is the stem output. The selected configuration uses C = 32 channels. This representation allows each channel to learn a different weighting of the same attribute, which later convolutions combine across neighbouring attributes.

3.4.2 ShuffleNet-V2 Units with Squeeze-and-Excitation Recalibration

The stem output passes through two ShuffleNet-V2 basic units, following the description by Liu et al. [20]. Each unit splits the channels into two halves, leaves the first half unchanged and processes the second half by a pointwise convolution, a depthwise convolution with kernel size three and a second pointwise convolution. The two halves are concatenated and then shuffled across groups, which can be written as follows:

$$U = Shuffle([\, H_1 \parallel B(H_2) \,]) \tag{15}$$

where H_1 and H_2 denote the two channel halves, B(·) represents the convolutional branch, ‖ is channel concatenation and Shuffle(·) interleaves the channels of the two halves. Channel splitting keeps equal input and output widths, which minimizes memory access cost, while channel shuffling allows information to flow between the two halves [20]. The SE module then recalibrates the channels of U. The squeeze operation compresses each channel into a descriptor, as in Eq. (16):

$$z_c = \frac{1}{k} \sum_{j=1}^{k} u_c(j) \tag{16}$$

where $u_c(j)$ is the value of channel c at attribute position j. The one-dimensional descriptor replaces the spatial average over $H \times W$ positions in the original image formulation [20]. The excitation and recalibration steps are calculated as follows:

$$s = \sigma(W_2\, \delta(W_1\, z)), \quad O_1 = s \odot U \tag{17}$$

where W_1 and W_2 are the weights of two fully connected layers with a reduction ratio of four, δ represents the ReLU activation, σ is the sigmoid function and $O_1 \in \mathrm{R}^{C\times k}$ is the recalibrated representation. The SE module assigns a weight between zero and one to every channel, allowing the network to emphasize channels that respond to informative attribute combinations. This attention mechanism adds only a small number of parameters, which suits the limited size of the training set.

3.4.3 Regression Branch and Taylor-Series Fusion

The regression branch provides a direct linear view of the selected attributes, which complements the nonlinear convolutional representation. The branch computes a weighted sum of the attributes and projects it to the shape of O_1, as follows:

$$y_1 = \sum_{i=1}^{k} F_i\, w_i, \quad Y_1 = reshape(W_r\, F + b_r) \in \mathbb{R}^{C\times k} \tag{18}$$

where w_i are learnable weights, $W_r \in \mathrm{R}^{Ck\times k}$ and b_r denote the projection weights and bias, and Y_1 is the projected regression output. The projection generalizes the scalar weighted sum y_1 so that every channel and position receives its own linear combination of the attributes. The fusion of Y_1 and O_1 is derived from a first-order Taylor expansion of a sequence $x(t)$, as in Eq. (19):

$$x(t+1) = x(t) + \frac{x'}{1!}(t), \quad x'(t) = \frac{x(t) - x(t-k)}{k} \tag{19}$$

where $x'(t)$ is approximated by a backward difference with step k. Setting $k = 1$ and substituting the derivative gives $x(t + 1) = 2x(t) − x(t − 1)$. The framework assigns the regression output to the current state, $x(t) = Y_1$, and the convolutional output to the previous state, $x(t − 1) = O_1$. The fused representation is therefore obtained as follows:

$$O_2 = 2\,Y_1 - O_1 = 2\,reshape(W_r\, F + b_r) - \sigma(W_2\, \delta(W_1\, z)) \odot U \tag{20}$$

where $O_2 \in \mathrm{R}^{C\times k}$ is the fused representation. This Taylor-series fusion is the original element of ShPCN-Net. Its rationale is to extrapolate from the attention-weighted convolutional state towards the linear state, so the network can emphasize the difference between the two views rather than simply adding them. Because both terms are learned jointly, the network can also reduce the contribution of either

branch if it does not improve the training objective. The ablation analysis in Section 4 evaluates whether this fusion improves performance over the individual branches.

3.4.4 Multi-Scale Parallel Convolution and Classification Head

The fused representation is processed by a multi-scale PCNN that follows the parallel structure introduced by Li et al. [21] for face identification. Four parallel convolutional branches with kernel sizes of 11, 7, 5 and 3 analyse O_2 at different receptive-field sizes, as given by Eq. (21):

$$O_{j+2} = GAP(\, ReLU(\, BN(\, f_{convj}(\, O_2 \,)\,)\,)\,), \quad j = 1, 2, 3, 4 \tag{21}$$

where $f_{\mathrm{conv,j}}$ denotes the convolution of the j-th branch with 32 filters, GAP represents global average pooling over the attribute positions and O_3 to O_6 are the pooled branch outputs. With only five to eight selected attributes, the kernel of size 11 spans the entire sequence, whereas the kernel of size 3 captures interactions among neighbouring attributes. The concatenated outputs are classified as follows:

$$O_7 = softmax(\, W_b \, ReLU(\, W_a \, [\, O_3 \parallel O_4 \parallel O_5 \parallel O_6 \,]\,)\,) \tag{22}$$

where W_{a} maps the 128 concatenated features to 64 hidden units, W_{b} maps them to the two classes and dropout with a rate of 0.4 is applied before both layers. The output O_7 contains the probabilities of the Normal and Disease classes, and the predicted label is the class with the larger probability. The multi-scale design is selected because the relevant interactions among clinical attributes are not known in advance, so several kernel sizes are evaluated in parallel rather than fixed by assumption.

3.5 Network Training and Configuration Selection

The network parameters are optimized with the Adam optimizer by minimizing the cross-entropy loss, which is defined as follows:

$$\mathcal{L} = -\frac{1}{B} \sum_{n=1}^{B} \sum_{c=1}^{2} y_{n,c} \, log \, p_{n,c} \tag{23}$$

where B is the mini-batch size, $y_{\mathrm{n,c}}$ indicates whether record n belongs to class c and $p_{\mathrm{n,c}}$ denotes the predicted probability. A stratified validation subset containing 15% of each training portion is separated for early stopping. The learning rate is halved when the validation loss does not improve for 10 epochs, and training stops after 25 epochs without improvement, after which the weights with the lowest validation loss are restored. The maximum number of epochs is 150. These rules prevent the network from fitting the small training folds for longer than necessary, which can be observed in the learning curves reported in Section 4.

The configuration of each method is selected by a bounded random search. An 80/20 stratified partition of the dataset is formed, and a validation set containing 20% of the 80% development portion is carved out, while the remaining 20% hold-out portion is never used for selection. Six candidate configurations per method, including the default configuration, are trained on the development records and scored on the validation records, and the candidate with the highest validation accuracy is chosen, with ties broken by the Matthews correlation coefficient. This rule is applied to every method, including the baselines, so that no method is disadvantaged by the selection procedure. The search space of ShPCN-Net covers the learning rate, dropout rate, channel width, number of ShuffleNet-V2 units, weight decay and batch size, and the chosen values are reported in Section 4.3. Because this search shares records with the later evaluation folds, the resulting estimates may be slightly optimistic, and the same effect applies equally to every method.

3.6 Baseline and Ablation Networks Under a Common Protocol

Three published methods that were evaluated on heart disease records serve as baselines. Fig. 3 illustrates the two deep baseline architectures. The HCBiLSTM model of Shrivastava et al. [9] applies

min–max scaling and Extra Trees importance ranking to select attributes. A one-dimensional convolution with 64 filters, max pooling and a bidirectional LSTM then process the selected attributes, and the forward and backward hidden states are summed. Because the original study does not report layer sizes or training settings, the selected configuration uses eight attributes, 32 LSTM units and a 32-unit dense layer. The Dense Autoencoder model of Alghamdi et al. [10] standardizes the attributes and encodes them with a dense autoencoder. Its 13-value reconstruction feeds a DenseNet with an initial convolution of 64 filters, dense blocks of six and twelve layers with a growth rate of 12, global average pooling and a 512-unit dense layer.

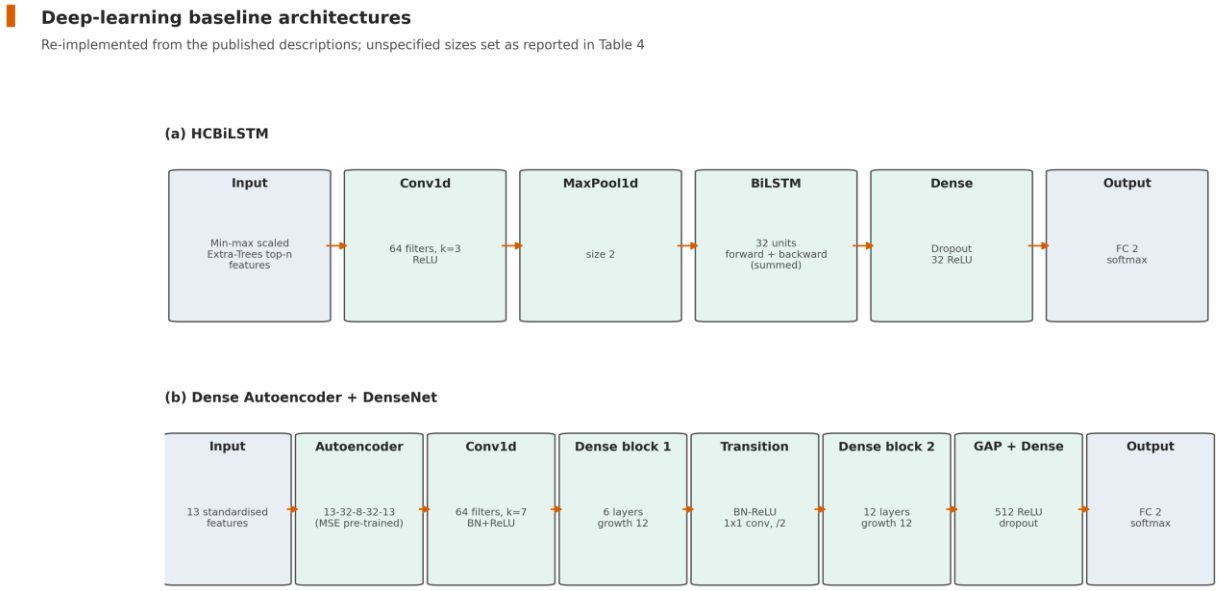


Fig. 3. Re-implemented deep baseline architectures: (a) HCBiLSTM and (b) Dense Autoencoder with DenseNet.

The XGBoost baseline follows the pipeline of Alharbi [12], which removes outliers with a z-score threshold of three and then trains a gradient-boosted tree ensemble. Outlier removal is applied to the training records only, so no test record is discarded. Whereas the original study used a combined dataset of 1,190 records, this study re-runs the pipeline on the Cleveland records under the common protocol. Two ablation variants isolate the contribution of the ShPCN-Net components. In particular, the ShuffleNet variant retains the stem, ShuffleNet-V2 units and SE module and classifies O_1 directly after global average pooling, whereas the PCNN variant passes the stem output directly to the multi-scale PCNN. Both variants use exactly the same preprocessing, GLO subset and training rules as ShPCN-Net, so any difference in performance reflects the network design alone.

All six methods are evaluated on identical stratified folds with a fixed random seed. Every method applies its own preprocessing inside each training fold, and the GLO is executed once per training fold and shared by ShPCN-Net and its two ablation variants. This common protocol allows the per-fold results to be paired, which is a prerequisite for the statistical tests reported in Section 4.6. Within each fold, the imputation and normalization statistics are first estimated on the training records, and the GLO is run on the normalized training records to obtain the subset. ShPCN-Net is then trained on the selected attributes with early stopping on a validation subset of the training records, and the held-out records are transformed with the stored statistics before prediction. Finally, the predicted probabilities and confusion counts of the held-out records are stored, so that both per-fold metrics and pooled out-of-fold predictions can be computed. No statistic, subset or stopping decision is derived from the held-out records at any stage.

3.7 Explainability Analysis Using Shapley Additive Explanations

Clinical acceptance of an automated detector requires an explanation of how each attribute influences a prediction. The framework applies Shapley additive explanations (SHAP), as employed by Shah et al. [8] for cardiovascular risk prediction, to the trained ShPCN-Net. Because the hybrid network is not a tree model, the model-agnostic kernel explainer is used with a background set summarized by 20 k-means centroids of the training records. SHAP values of the selected attributes are estimated for 60 records, comprising the 31 held-out test records of the reference split and 29 randomly chosen training records, with 200 coalition samples per record. The SHAP value of attribute *i* for a record is the Shapley value of a cooperative game in which the attributes act as players, as defined in Eq. (24):

$$\varphi_i = \sum_{S \subseteq F \setminus \{i\}} \frac{|S|!\,(|F| - |S| - 1)!}{|F|!} \left[f_{S \cup \{i\}}(x_{S \cup \{i\}}) - f_S(x_S) \right] \tag{24}$$

where F is the set of selected attributes, S is a subset that excludes attribute i and f_S denotes the model output when only the attributes in S are known, with the remaining attributes marginalized over the background set. The kernel explainer approximates these values by a weighted linear regression over sampled coalitions. A positive SHAP value indicates that the attribute increases the predicted probability of disease, whereas a negative value indicates a decrease. The SHAP values of a record sum to the difference between its prediction and the average prediction over the background set.

Two complementary analyses characterize the data and the learned representation. Kernel density estimation compares the distributions of key continuous attributes between the two classes, following the presentation of Alghamdi et al. [10]. In addition, t-distributed stochastic neighbour embedding (t-SNE) projects both the selected input attributes and the 64-dimensional penultimate embedding of ShPCN-Net into two dimensions, as applied by Shah et al. [8]. These projections show whether the network separates the two classes more clearly than the raw selected attributes.

4. Experimental Procedure

This section describes the experimental environment, the dataset, the parameter settings and the evaluation metrics, and then reports the results of ShPCN-Net and the competing methods. The results follow the processing order of the framework, from the characterization of the input attributes and the feature selection behaviour to the classification, comparison, ablation and explanation results.

4.1 Experimental Setup

All experiments were conducted on a laptop with an Intel Core i5-1334U processor and 16.85 GB of memory, without a discrete graphics processor, running Windows 11. Table 2 summarizes the software environment. The framework was implemented in Python 3.10.11, with PyTorch 2.5.1 executed on the CPU for the neural networks, scikit-learn 1.6.1 for preprocessing and the wrapper classifier, XGBoost 2.1.3 for the tree baseline and SHAP 0.49.1 for the explanations. Independent training jobs were distributed over four worker processes with one computation thread each, and every run used the random seed 42. The complete experimental pipeline, including the configuration search, required approximately 24 minutes of computation, of which the 10-fold comparison of the six methods took 404 s with the selected configurations.

Table 2. Experimental hardware and software environment.

Item	Specification
Processor	Intel Core i5-1334U (10 physical cores, 12 logical cores)
Memory	16.85 GB
Graphics processor	None used (CPU-only execution)
Operating system	Windows 11 (build 10.0.26200)

Item	Specification
Language	Python 3.10.11
Libraries	PyTorch 2.5.1, scikit-learn 1.6.1, XGBoost 2.1.3, SHAP 0.49.1, NumPy 1.26.4, pandas 2.3.3, SciPy 1.15.3, Matplotlib 3.8.4
Parallelism and seed	Four worker processes, one thread each; random seed 42

4.2 Dataset Description

The experiments used the Cleveland heart disease dataset from the UCI Machine Learning Repository [11], which contains 303 patient records with 13 clinical attributes and a binary diagnosis. Table 3 lists the attributes and their observed ranges. Of these records, 164 are Normal (54.1%) and 139 are Disease (45.9%), so the class distribution is approximately balanced. This version contains no missing values and no duplicate records, so all 303 records were retained after preprocessing. Every experiment used the complete dataset.

Table 3. Attributes of the Cleveland heart disease dataset (N = 303).

Attribute	Description	Type	Observed range
age	Age (years)	Continuous	29–77
sex	Sex (1 = male, 0 = female)	Categorical	0–1
cp	Chest pain type	Categorical	0–3
trestbps	Resting blood pressure (mm Hg)	Continuous	94–200
chol	Serum cholesterol (mg/dl)	Continuous	126–564
fbs	Fasting blood sugar > 120 mg/dl	Categorical	0–1
restecg	Resting electrocardiographic result	Categorical	0–2
thalach	Maximum heart rate achieved (beats/min)	Continuous	71–202
exang	Exercise-induced angina	Categorical	0–1
oldpeak	ST depression induced by exercise (mm)	Continuous	0–6.2
slope	Slope of the peak exercise ST segment	Categorical	0–2
ca	Number of major vessels coloured by fluoroscopy	Categorical	0–3
thal	Thallium scan result	Categorical	1–3
target	Diagnosis (0 = Normal, 1 = Disease)	Label	164 / 139

Three partitioning schemes were applied. The headline comparison used stratified 10-fold cross-validation, so each test fold contained 30 or 31 records and every record was tested exactly once. The data-fraction analysis trained each method on 50%, 60%, 70%, 80% and 90% of the records and tested it on the remaining records, using stratified splits with the same seed. Finally, a reference split with 90% of the records for training and 10% for testing produced the learning curves, the representative samples and the explanations. In this split, 231 records were used for fitting, 41 for early-stopping validation and 31 for testing.

4.3 Parameter Settings

Table 4 reports the parameters used in the reported experiments. For the six methods, the values are the configurations chosen by the validation-based search described in Section 3.5. The GLO settings follow the original algorithm [18], and the population and iteration budget follow Qiu et al. [19]. Identical early-stopping and learning-rate rules were applied to every neural network.

Table 4. Parameter settings of the GLO, ShPCN-Net and the competing methods in the reported experiments.

Component	Settings
GLO feature selection	Population 30; iterations 100; α = 0.99; bounds [0, 1]; threshold 0.5; ω = 0.05; φ = 5; one-male females 40%; multi-male males 30% and females 40%; wrapper KNN (k = 5) with 5-fold CV
ShPCN-Net	Channels 32; ShuffleNet-V2 units 2; SE reduction 4; PCNN kernels 11/7/5/3 with 32 filters; hidden units 64; dropout 0.4; Adam learning rate 0.001; weight decay 0.001; batch size 32; maximum 150 epochs
Training rules (all networks)	Cross-entropy loss; 15% validation split; learning rate halved after 10 epochs without improvement; early stopping after 25 epochs; best weights restored
ShuffleNet (ablation)	Channels 16; units 2; dropout 0.3; learning rate 0.003; weight decay 0.001; batch size 16
PCNN (ablation)	Channels 16; 48 filters per branch; dropout 0.4; learning rate 0.0005; weight decay 0.0001; batch size 32
HCBiLSTM [9]	Extra Trees top 8 attributes; Conv1d 64 filters (k = 3); BiLSTM 32 units; dense 32; dropout 0.2; learning rate 0.001; batch size 32; maximum 100 epochs
Dense Autoencoder [10]	Autoencoder 13–32–8–32–13 (50 epochs); DenseNet growth 12; dense 512; dropout 0.4; learning rate 0.0001; weight decay 0.0001; batch size 32; maximum 150 epochs
XGBoost [12]	z-score threshold 3 (training records); 200 trees; maximum depth 4; learning rate 0.03; subsample 0.8; column sample 0.6

4.4 Evaluation Metrics

The primary metrics were accuracy, the Matthews correlation coefficient (MCC) and the true positive rate (TPR), and precision, specificity, F1-score and the area under the receiver operating characteristic curve (AUC) were reported as secondary metrics. All metrics are dimensionless; accuracy, TPR, precision, specificity and F1-score are expressed as percentages. Let TP, TN, FP and FN denote the numbers of true positives, true negatives, false positives and false negatives, where Disease is the positive class. The metrics are calculated as follows:

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN}, \quad Precision = \frac{TP}{TP + FP} \tag{25}$$

$$TPR = \frac{TP}{TP + FN}, \quad Specificity = \frac{TN}{TN + FP}, \quad F1 = \frac{2 \cdot Precision \cdot TPR}{Precision + TPR} \tag{26}$$

$$MCC = \frac{TP \cdot TN - FP \cdot FN}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}} \tag{27}$$

where MCC ranges from −1 to 1 and remains informative when the classes are unequal in size. The AUC is the area under the curve of TPR against the false positive rate FP/(FP + TN) as the decision threshold varies. For the cross-validation results, each metric was computed on every test fold and then averaged, and the standard deviation across the ten folds was reported. The paired comparison between ShPCN-Net and each baseline used the per-fold accuracies, and the paired t statistic is defined as follows:

$$t = \frac{\bar{d}}{s_d / \sqrt{n}} \tag{28}$$

where $\bar{d}$ and s_d denote the mean and standard deviation of the per-fold accuracy differences and $n = 10$ is the number of folds, giving nine degrees of freedom. Because the normality of ten differences cannot be verified reliably, the non-parametric Wilcoxon signed-rank test was also applied, following the significance analysis of Calzoni et al. [1]. A two-sided significance level of 0.05 was used for both tests.

4.5 Experimental Results

4.5.1 Characterization of the Clinical Attributes

Fig. 4 presents the kernel density estimates of four continuous attributes for the two classes. The maximum heart rate (thalach) of the Disease class is shifted towards lower values, with a peak near 145

beats/min compared with approximately 160 beats/min for the Normal class. ST depression (oldpeak) concentrates near zero for Normal records, whereas the Disease class shows a broad distribution that extends beyond 3 mm. In contrast, the cholesterol distributions of the two classes overlap almost completely, and age shows only a moderate shift towards older patients in the Disease class. These distributions suggest that individual continuous attributes provide partial separation at best, which motivates the search for informative attribute combinations.

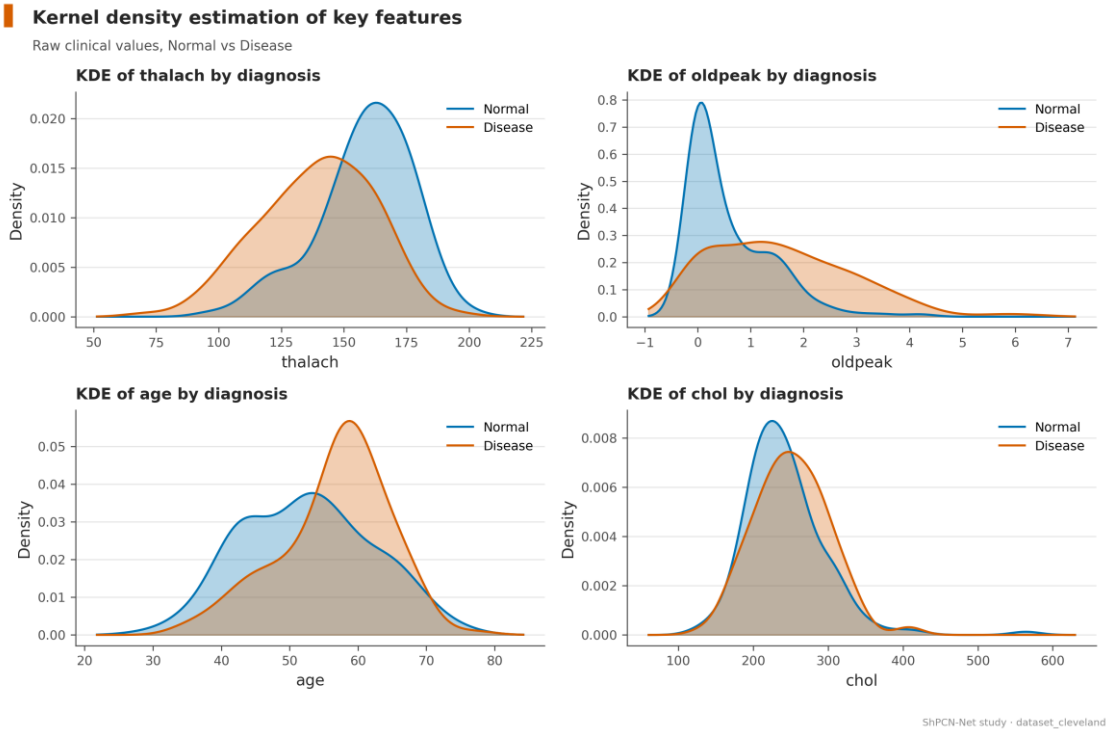


Fig. 4. Kernel density estimates of maximum heart rate, ST depression, age and serum cholesterol for Normal and Disease records.

Fig. 5 and Fig. 6 compare t-SNE projections of the selected input attributes and of the ShPCN-Net penultimate embedding for all 303 records. The projection of the selected inputs forms several clusters that contain records of both classes, because the categorical attributes produce many records with identical or similar codes. The projection of the learned embedding arranges the Disease records on one side and the Normal records on the other, with a mixed region in the centre. This arrangement indicates that the network transforms the selected attributes into a representation in which the two classes are more clearly separated, although the overlapping region corresponds to records that remain difficult to classify.

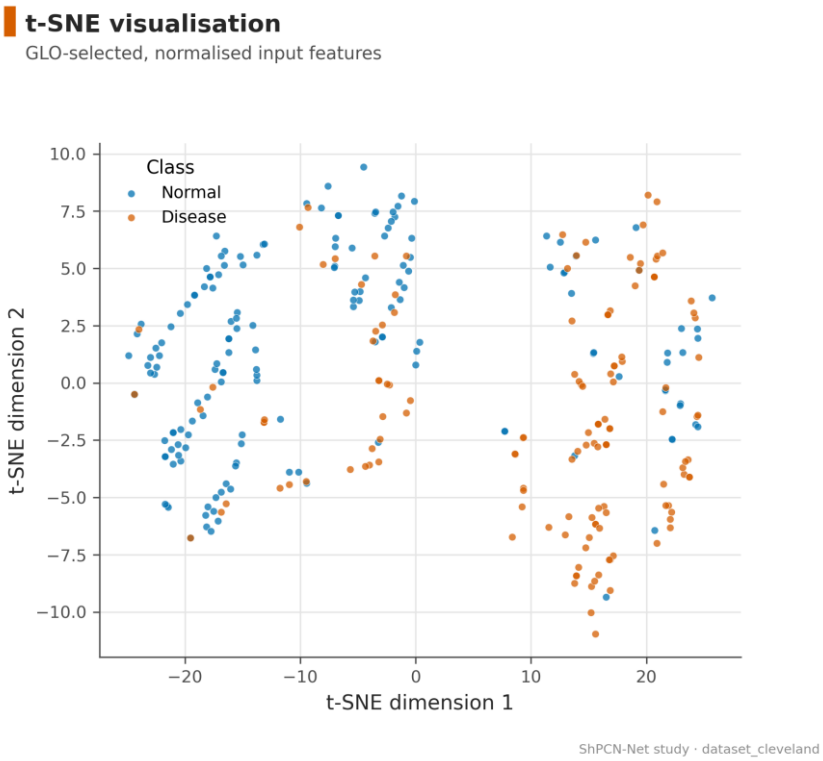


Fig. 5. t-SNE projection of the GLO-selected and normalized input attributes.

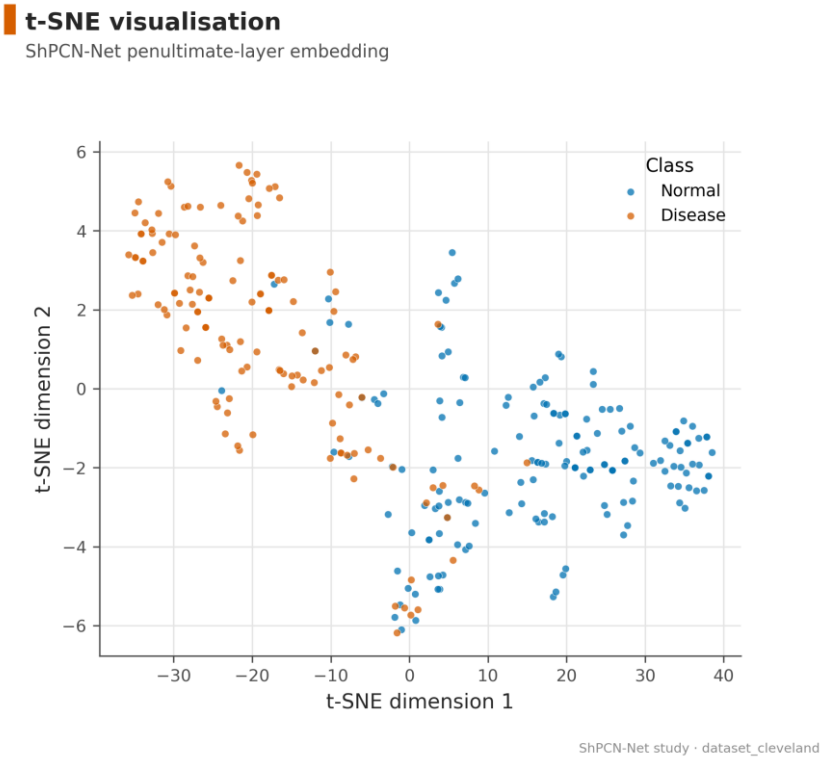


Fig. 6. t-SNE projection of the ShPCN-Net penultimate-layer embedding.

4.5.2 Behaviour of the Gray Langurs Optimizer

Fig. 7 illustrates the convergence of the GLO on the training records of the reference split. The best fitness decreased from 0.1772 at the first iteration to 0.1604 after 100 iterations, while the mean fitness of the population approached the best value. During the search, the size of the best subset fell from eight

to five attributes, and the wrapper accuracy increased from 82.7% to 84.2%. The final subset contained age, cp, slope, ca and thal. These trends indicate that the fitness function rewarded smaller subsets without sacrificing wrapper accuracy.

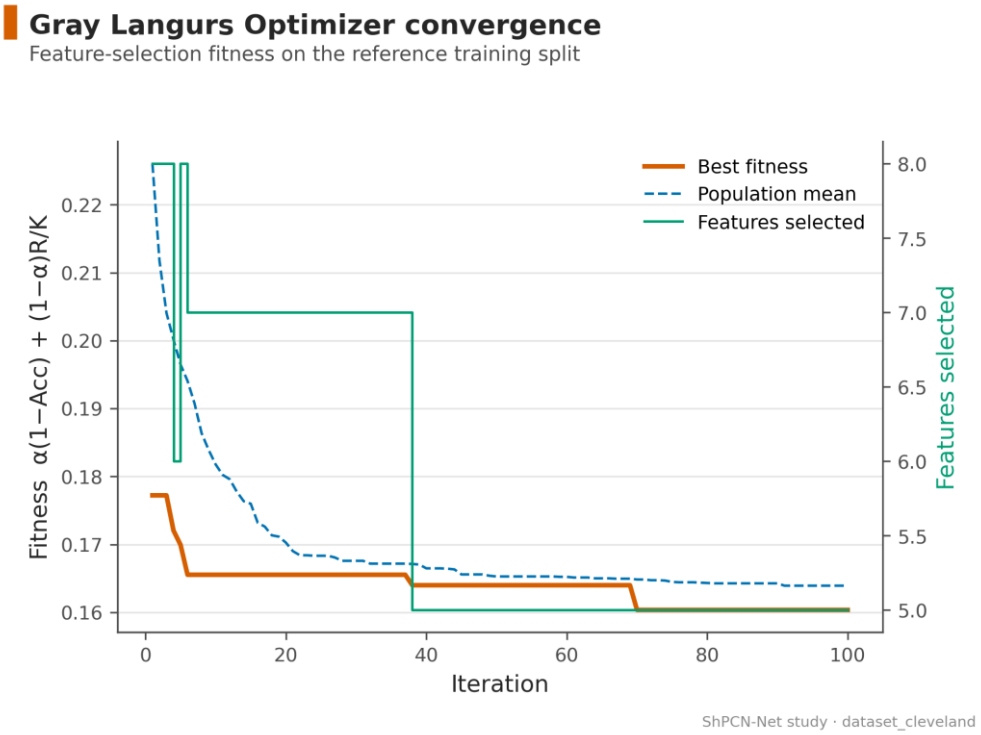


Fig. 7. Convergence of the GLO feature selection, showing the best and mean fitness and the number of selected attributes.

Fig. 8 summarizes how often each attribute was selected across the ten cross-validation folds. The number of major vessels (ca) was selected in every fold, chest pain type (cp) in nine folds and the thallium scan result (thal) in eight folds, while slope and exang were selected in six folds. The GLO retained 5.7 attributes per fold on average, which corresponds to a reduction of 56% relative to the full set of 13 attributes. Notably, the continuous attributes were selected rarely; thalach was retained in only one fold and oldpeak in two folds, despite the class differences visible in Fig. 4. This preference can be attributed to the strong discriminative power of the categorical attributes and to the symmetric Gaussian transform, which maps low and high values of a continuous attribute to similar normalized values.

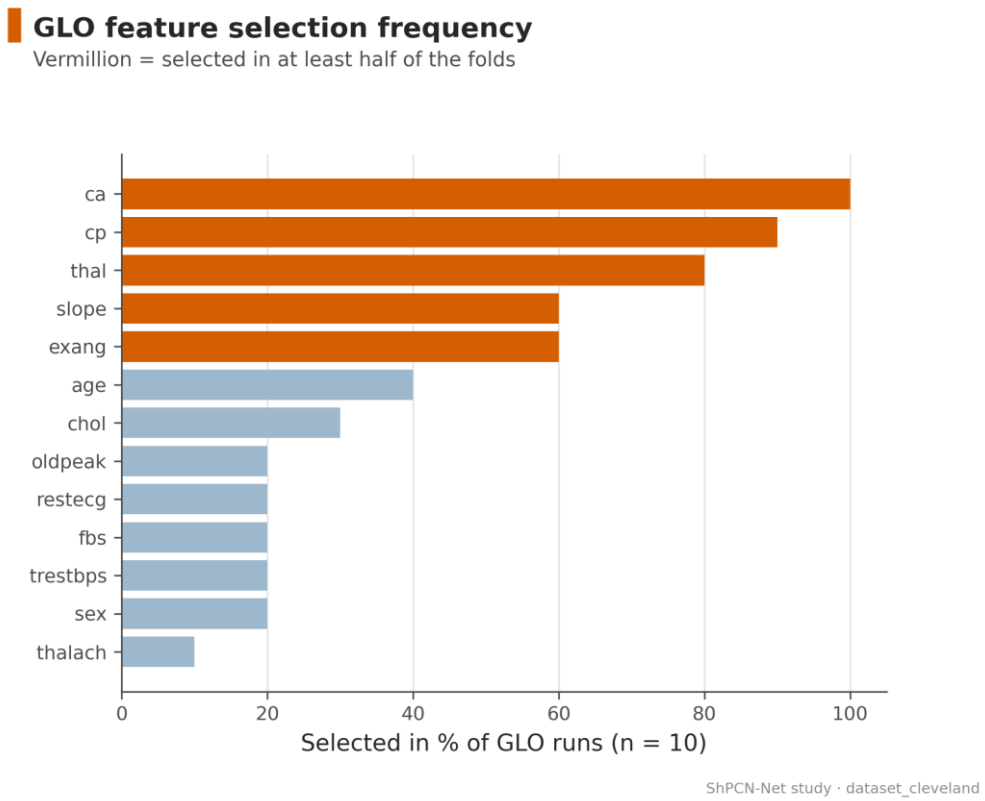


Fig. 8. Selection frequency of each attribute by the GLO across the ten cross-validation folds.

4.5.3 Representative Record Through the Processing Pipeline

Fig. 9 to Fig. 13 follow one held-out test record (record 213, a Normal patient) through the complete framework. In Fig. 9, the preprocessing stage normalizes the 13 raw attribute values and highlights the five attributes selected by the GLO. Fig. 10 depicts the output of the ShuffleNet-V2 and SE branch: the heat map shows the recalibrated representation O_1 over the 32 channels and five attribute positions, and the bar chart shows the channel weights produced by the SE module. Several channels receive weights close to 0.6, whereas others are suppressed to approximately 0.4, which indicates that the SE module redistributes emphasis among the channels.

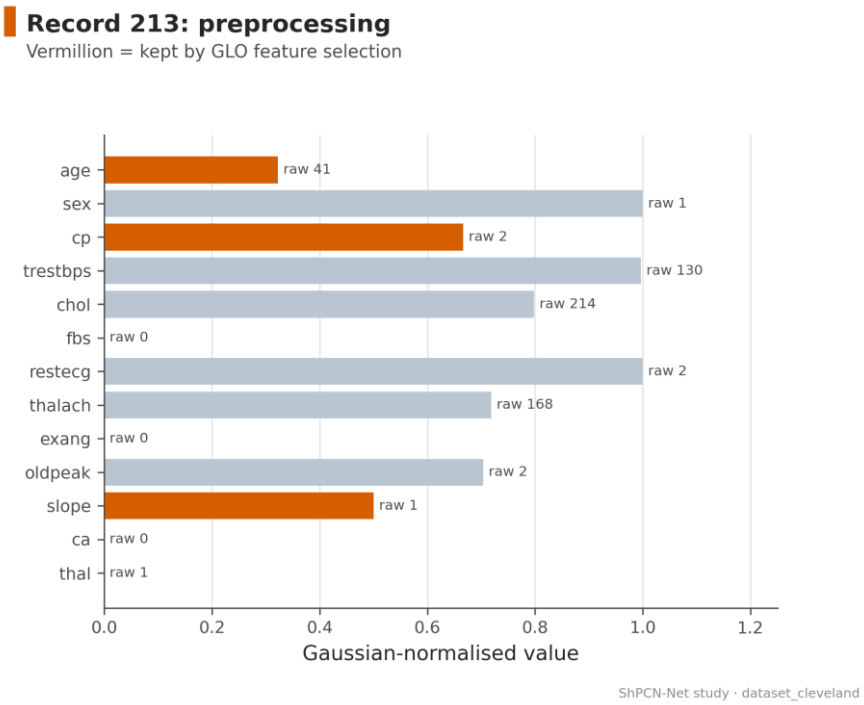


Fig. 9. Preprocessing stage for record 213: normalized attribute values with the GLO-selected attributes highlighted.

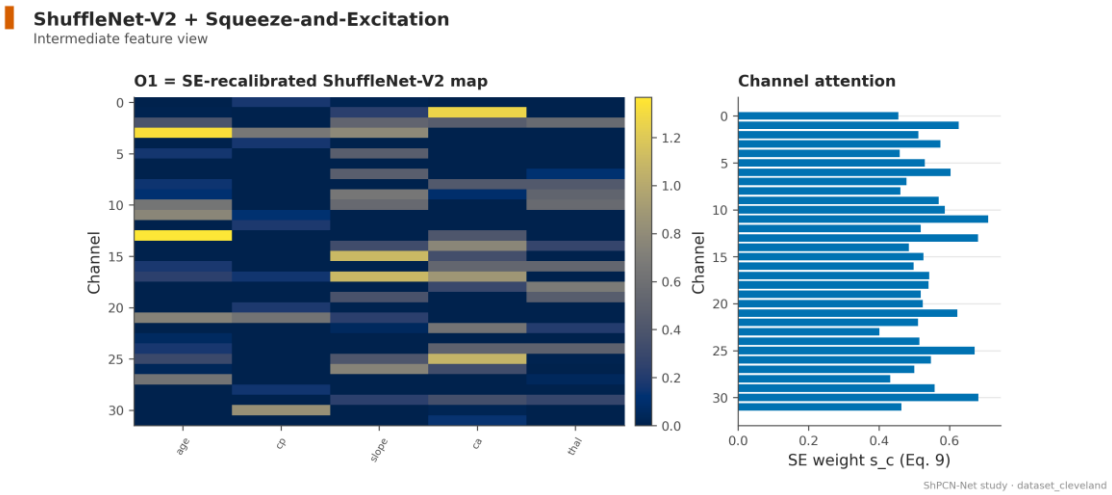


Fig. 10. ShuffleNet-V2 and SE output for record 213: recalibrated representation O1 and SE channel weights.

Fig. 11 presents the three maps involved in the Taylor-series fusion, namely the projected regression output Y_1, the convolutional output O_1 and the fused representation $O_2 = 2Y_1 - O_1$. The fused map inherits the structure of the regression output while subtracting the activations of the convolutional branch, so channels with strong convolutional responses are attenuated. Fig. 12 shows the mean responses of the four PCNN branches across the attribute positions. The branches with larger kernels produce smoother responses, whereas the branch with kernel size three responds more strongly to individual attribute positions, which confirms that the branches capture different scales of interaction.

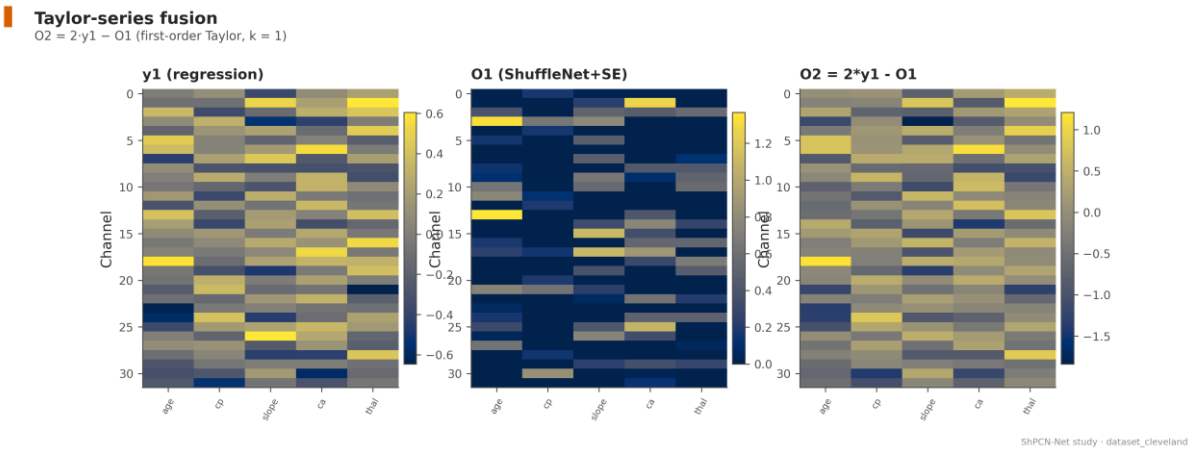


Fig. 11. Taylor-series fusion for record 213: regression output Y1, convolutional output O1 and fused representation O2.

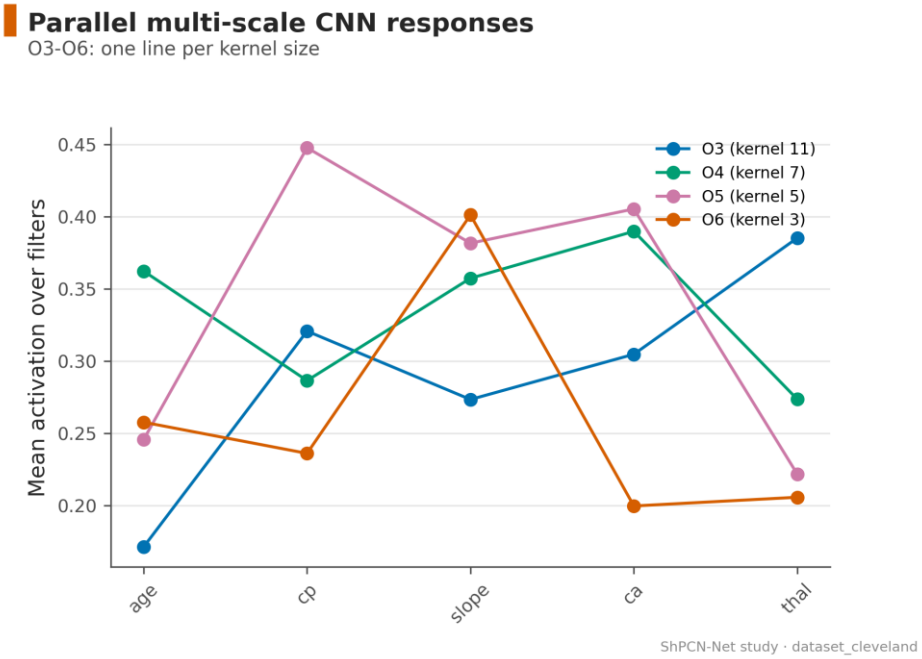


Fig. 12. Mean responses of the four PCNN branches (kernel sizes 11, 7, 5 and 3) for record 213.

Fig. 13 compares the final prediction of ShPCN-Net with the predictions of the five competing methods for the same record. ShPCN-Net assigned a disease probability of 0.066 and correctly classified the record as Normal, and all five competing methods also predicted Normal, with probabilities between 0.077 (XGBoost) and 0.367 (Dense Autoencoder). Among the six representative test records analysed, ShPCN-Net classified five correctly; the misclassified case (record 197) was a Normal patient assigned a disease probability of 0.840.

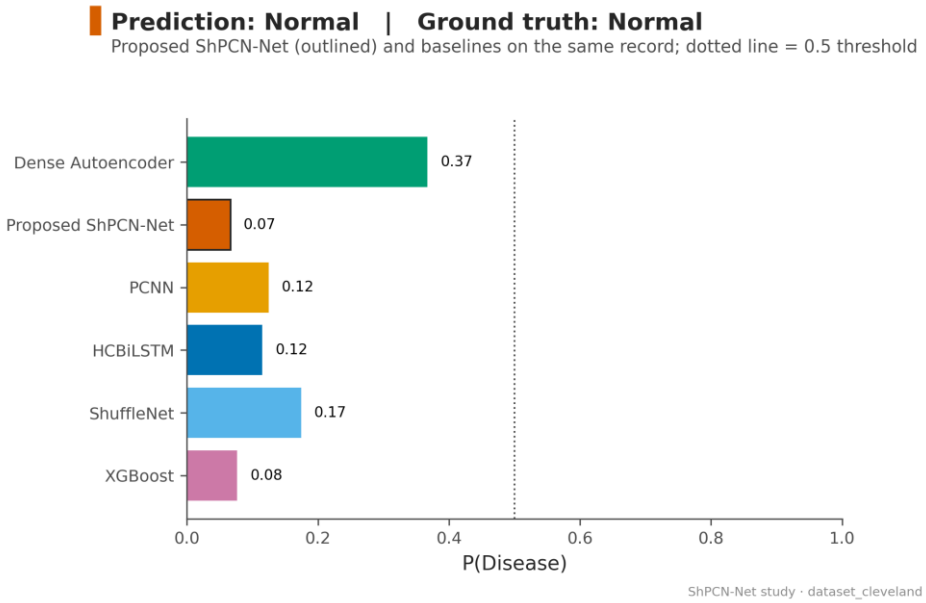


Fig. 13. Predicted disease probability of ShPCN-Net and the competing methods for record 213 compared with the ground truth.

4.5.4 Training Behaviour and Classification Outcome

Fig. 14 depicts the training and validation curves of ShPCN-Net on the reference split. The training loss decreased from 0.646 to 0.221, and the validation loss reached its minimum at epoch 83, after which early stopping ended training at epoch 108 and restored the best weights. The validation accuracy stabilized between 0.83 and 0.88 after approximately 40 epochs, while the training accuracy continued to rise towards 0.90. This gap indicates mild overfitting in the later epochs, which the early-stopping rule limits. On the 31 held-out test records of this split, ShPCN-Net classified 26 records correctly, which corresponds to an accuracy of 83.87%, an MCC of 0.693 and a TPR of 92.86%.

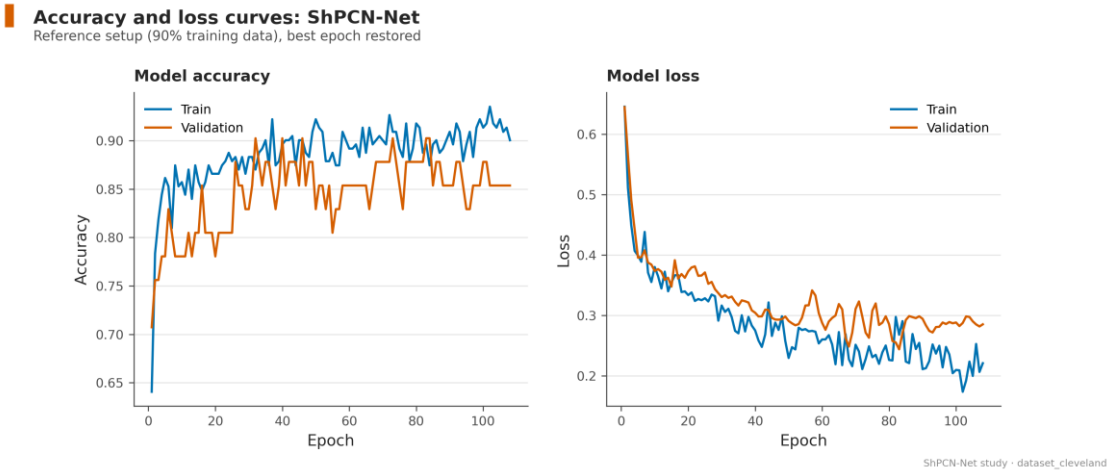


Fig. 14. Training and validation accuracy and loss of ShPCN-Net on the reference split.

Fig. 15 reports the confusion matrix of the out-of-fold predictions pooled over the ten folds. ShPCN-Net correctly identified 107 of the 139 Disease records and 139 of the 164 Normal records, with 32 false negatives and 25 false positives. The resulting pooled TPR of 76.98% is lower than the specificity of 84.76%, so missed disease cases remain the main source of error. Fig. 16 compares the receiver operating characteristic curves of the six methods. XGBoost achieved the largest pooled AUC of 0.902,

whereas ShPCN-Net reached 0.876, and the curves of all methods overlap closely across most of the threshold range.

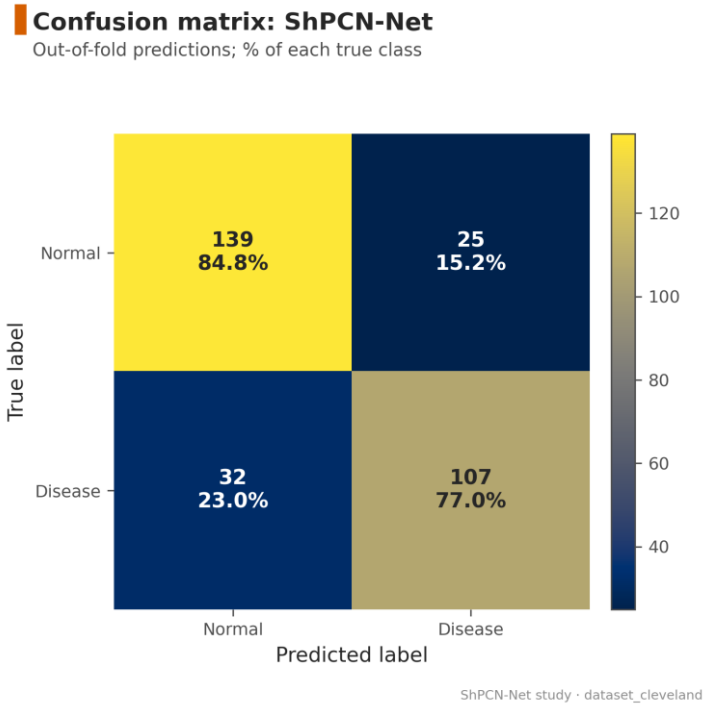


Fig. 15. Confusion matrix of the ShPCN-Net out-of-fold predictions over the ten folds.

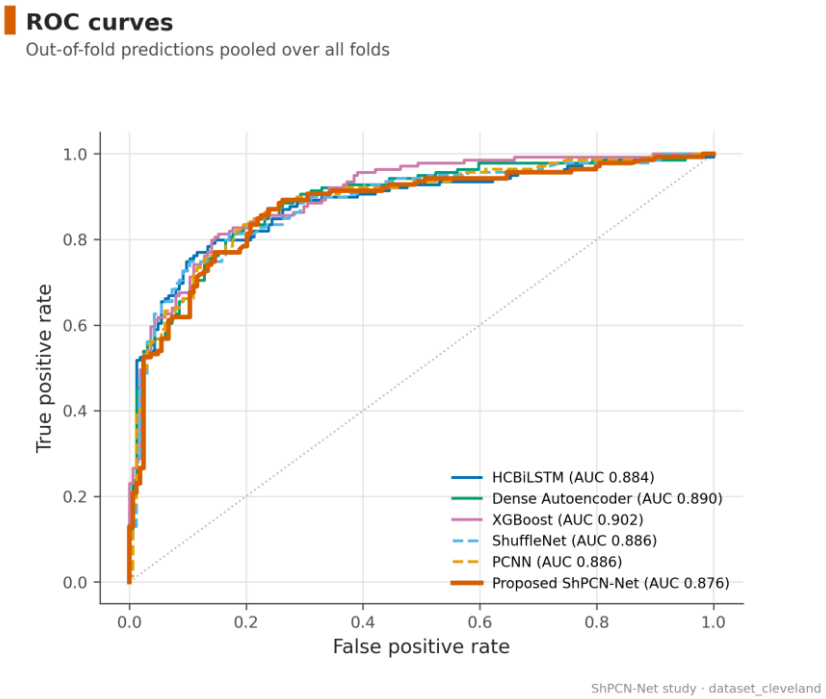


Fig. 16. Receiver operating characteristic curves of the six methods from pooled out-of-fold predictions.

Fig. 17 presents a deviation analysis that quantifies how far the predictions depart from the ground truth. In panel (a), the deviation of each test record of the reference split is defined as the predicted disease probability minus the true label, so values near zero indicate confident correct predictions and values beyond ±0.5 indicate errors. The mean absolute deviation was 0.197, and five records exceeded the threshold distance: four Normal records received high disease probabilities, and one Disease record

received a low probability. Panel (b) shows the deviation of each fold accuracy from the mean accuracy of each method. ShPCN-Net deviated between −11.15 and +8.85 percentage points, which is comparable to the spread of the baselines and confirms that fold composition strongly affects accuracy on this small dataset.

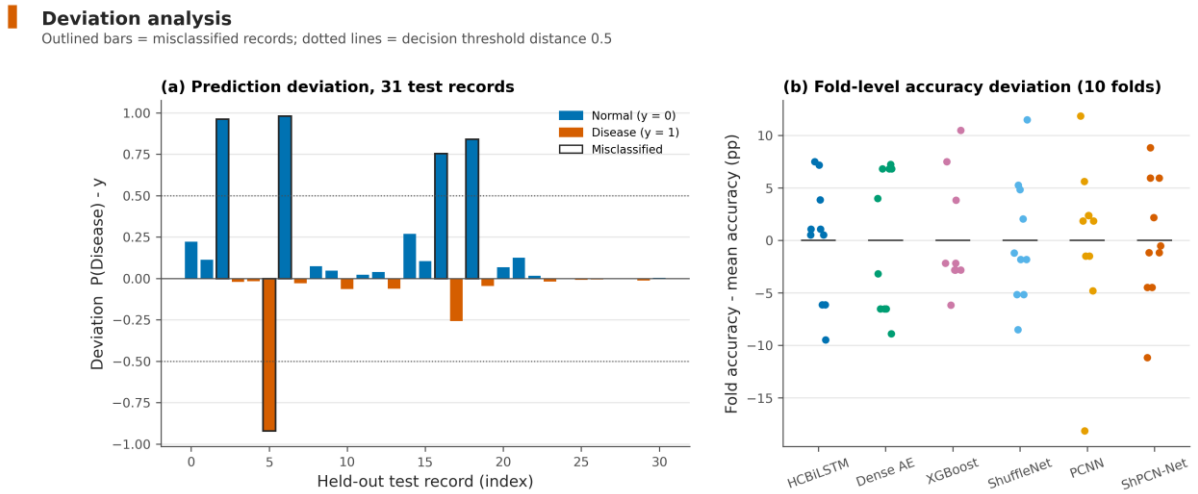


Fig. 17. Deviation analysis: (a) predicted probability minus true label for the 31 test records of the reference split; (b) deviation of each fold accuracy from the method mean.

4.5.5 Comparative Analysis

Table 5 summarizes the 10-fold cross-validation results of all methods, and Fig. 18 visualizes the means and standard deviations. XGBoost and HCBiLSTM achieved the highest accuracies of 82.83% and 82.81%, followed by the ShuffleNet variant with 81.83%, the PCNN variant with 81.48%, ShPCN-Net with 81.15% and the Dense Autoencoder with 79.86%. ShPCN-Net recorded an MCC of 0.630, compared with 0.661 for XGBoost and 0.659 for HCBiLSTM. In terms of TPR, ShPCN-Net reached 77.03%, which exceeded every method except XGBoost (79.12%). However, the standard deviations of approximately 6 percentage points in accuracy are several times larger than the differences between the means.

Table 5. Stratified 10-fold cross-validation results on the Cleveland dataset (mean ± standard deviation).

Method	Accuracy (%)	MCC	TPR (%)	Precision (%)	Specificity (%)	F1-score (%)	AUC
HCBiLSTM [9]	82.81 ± 5.69	0.659 ± 0.118	76.26 ± 9.53	85.47 ± 8.91	88.42 ± 7.40	80.18 ± 6.65	0.893 ± 0.048
Dense Autoencoder [10]	79.86 ± 6.87	0.604 ± 0.142	69.73 ± 13.12	84.68 ± 10.53	88.42 ± 8.82	75.66 ± 9.00	0.895 ± 0.056
XGBoost [12]	82.83 ± 5.38	0.661 ± 0.110	79.12 ± 7.90	83.80 ± 10.05	86.07 ± 9.43	80.90 ± 5.95	0.906 ± 0.039
ShuffleNet (ablation)	81.83 ± 5.97	0.639 ± 0.121	74.67 ± 12.82	84.45 ± 7.53	87.79 ± 7.09	78.59 ± 8.33	0.880 ± 0.066
PCNN (ablation)	81.48 ± 7.80	0.631 ± 0.163	76.10 ± 11.37	82.60 ± 11.13	86.03 ± 8.98	78.85 ± 9.60	0.890 ± 0.053
ShPCN-Net (proposed)	81.15 ± 5.97	0.630 ± 0.120	77.03 ± 11.48	82.22 ± 9.87	84.78 ± 10.04	78.77 ± 7.15	0.878 ± 0.057

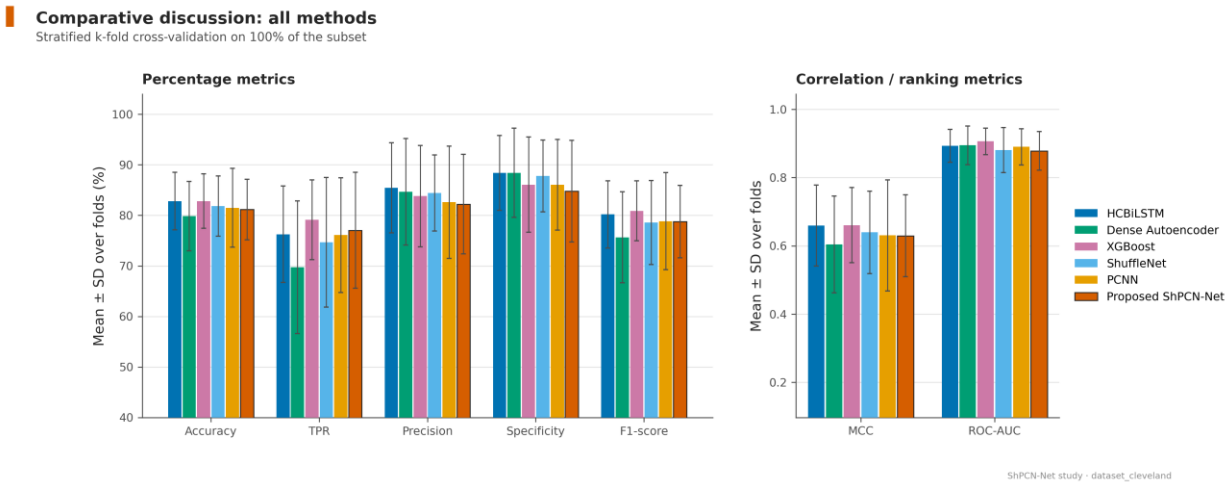


Fig. 18. Comparison of all methods on seven metrics under stratified 10-fold cross-validation (mean ± standard deviation).

Fig. 19, Fig. 20 and Fig. 21 show how accuracy, MCC and TPR change as the training portion increases from 50% to 90%, and Table 6 lists the accuracy and MCC values. ShPCN-Net improved from 68.42% accuracy with 50% training data to 85.25% with 80%, which is the steepest increase among the methods. In contrast, XGBoost and the Dense Autoencoder remained above 81% at every fraction and reached their highest values of 90.16% and 87.10% at 80% and 90% training data. ShPCN-Net attained the highest or joint-highest TPR at 50% (90.00%) and 90% (92.86%) training data. These trends suggest that the hybrid network requires more training records than the tree ensemble to reach its best accuracy, which is consistent with its larger number of trainable parameters.

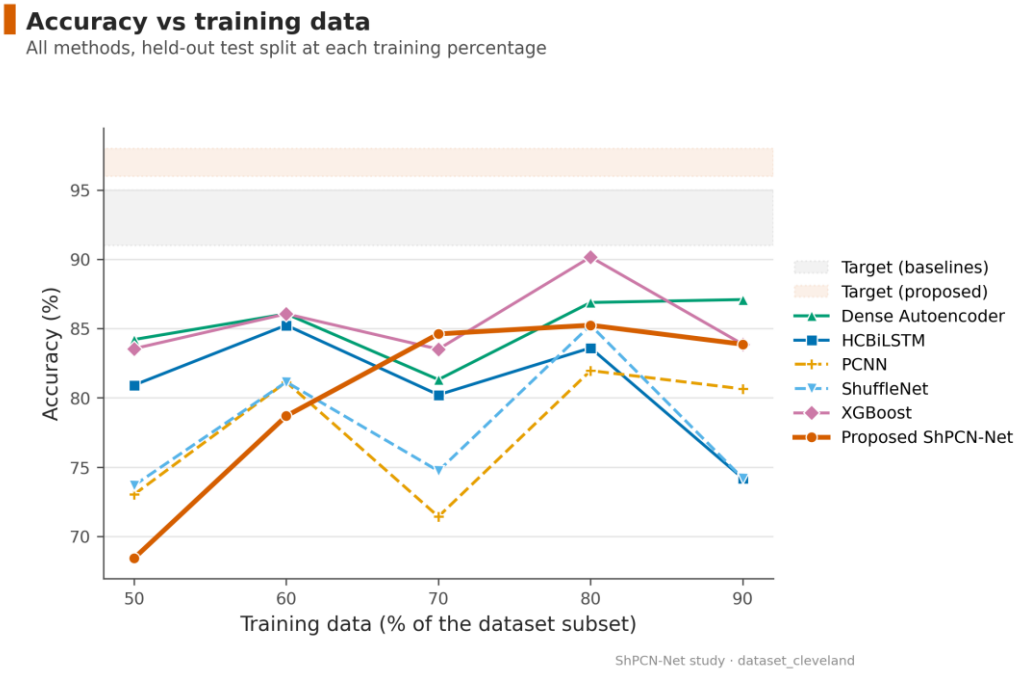


Fig. 19. Accuracy of all methods as a function of the training data percentage.

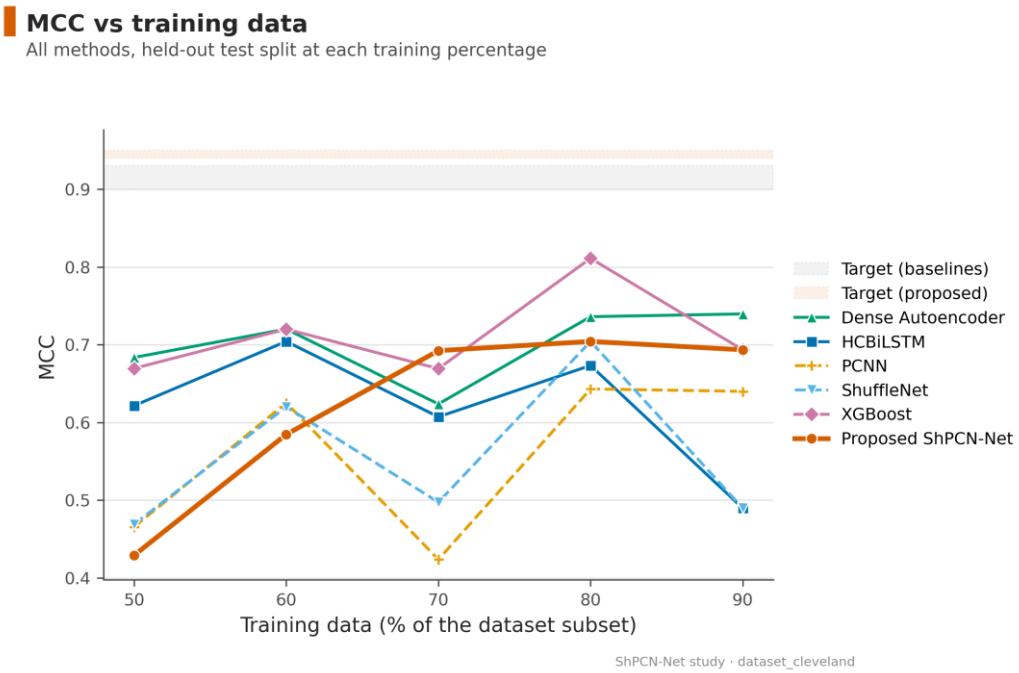


Fig. 20. MCC of all methods as a function of the training data percentage.

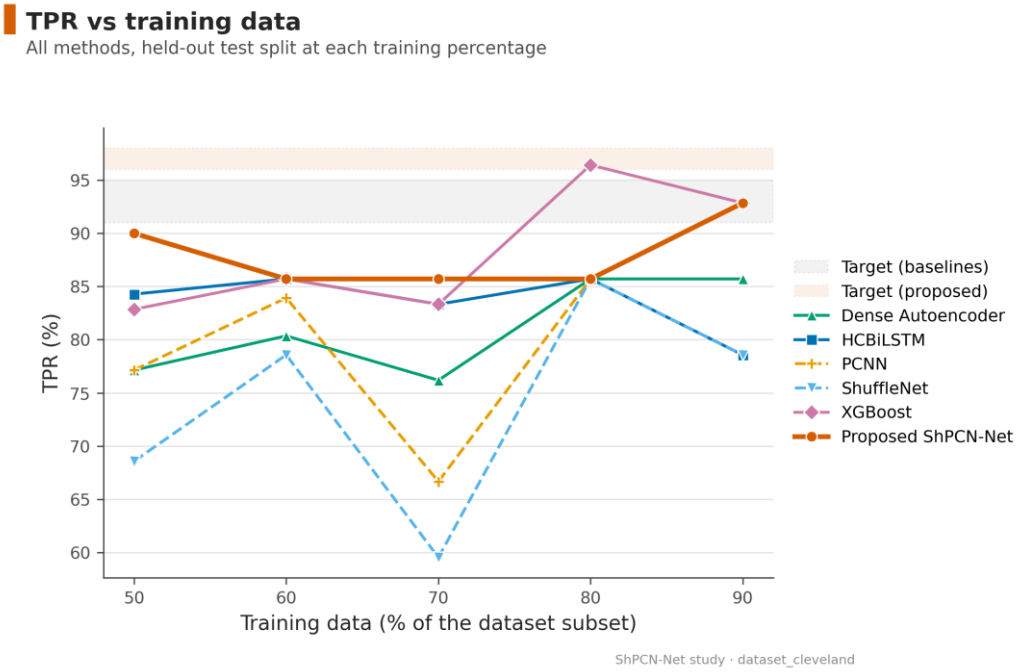


Fig. 21. TPR of all methods as a function of the training data percentage.

Table 6. Accuracy (%) and MCC of all methods for training data percentages from 50% to 90%.

Method	**50%**	**60%**	**70%**	**80%**	**90%**
HCBiLSTM	80.92 / 0.621	85.25 / 0.704	80.22 / 0.607	83.61 / 0.673	74.19 / 0.490
Dense Autoencoder	84.21 / 0.683	86.07 / 0.720	81.32 / 0.624	86.89 / 0.736	87.10 / 0.739
XGBoost	83.55 / 0.669	86.07 / 0.720	83.52 / 0.669	90.16 / 0.811	83.87 / 0.693
ShuffleNet	73.68 / 0.469	81.15 / 0.620	74.73 / 0.497	85.25 / 0.704	74.19 / 0.490
PCNN	73.03 / 0.465	81.15 / 0.625	71.43 / 0.424	81.97 / 0.643	80.65 / 0.640
ShPCN-Net	68.42 / 0.429	78.69 / 0.584	84.62 / 0.692	85.25 / 0.704	83.87 / 0.693

Fig. 22 depicts the AUC of each method in every fold together with 95% bootstrap intervals, following the fold-wise presentation of Srivastava et al. [24]. The mean fold AUC of ShPCN-Net was 0.878 ± 0.057, with a range from 0.795 to 0.977, whereas XGBoost reached 0.906 ± 0.039. The bootstrap intervals of all methods overlap in every fold, and the fold-to-fold variation of each method exceeds the differences between methods within a fold.

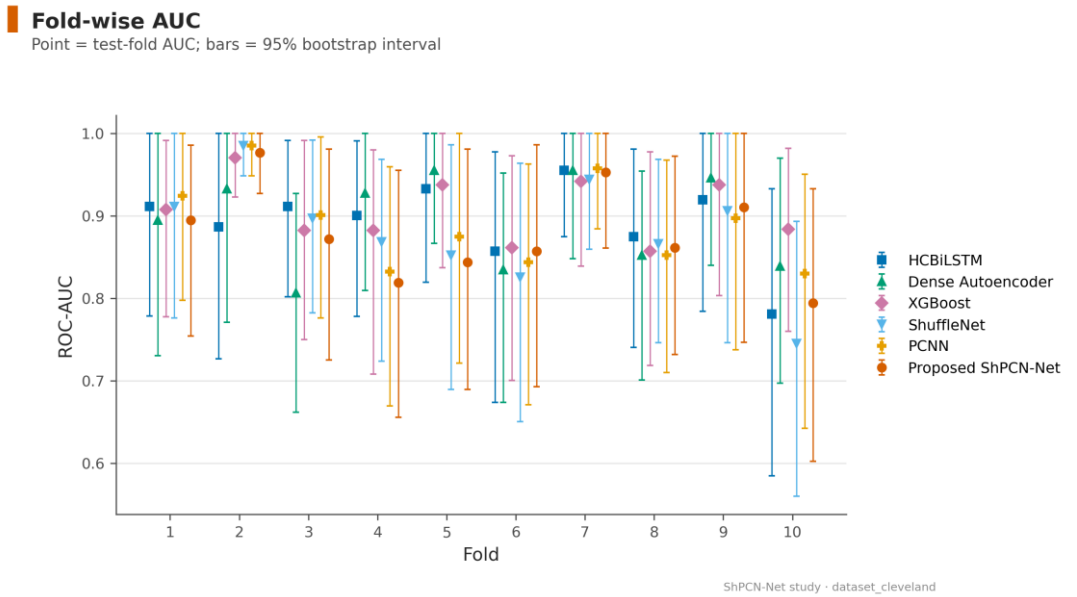


Fig. 22. Fold-wise AUC of the six methods with 95% bootstrap intervals.

4.5.6 Ablation, Scalability and Sensitivity Analyses

Fig. 23 compares ShPCN-Net with its two ablation variants, using the component-removal analysis adopted by Xiaomeng et al. [25]. The ShuffleNet variant achieved 81.83% accuracy and an MCC of 0.639, the PCNN variant reached 81.48% and 0.631, and the full ShPCN-Net recorded 81.15% and 0.630. The full model obtained the highest TPR of 77.03%, compared with 74.67% and 76.10% for the variants. Therefore, the Taylor-series fusion and the combination of both branches did not improve accuracy or MCC on this dataset, although the full model detected a slightly larger share of disease cases.

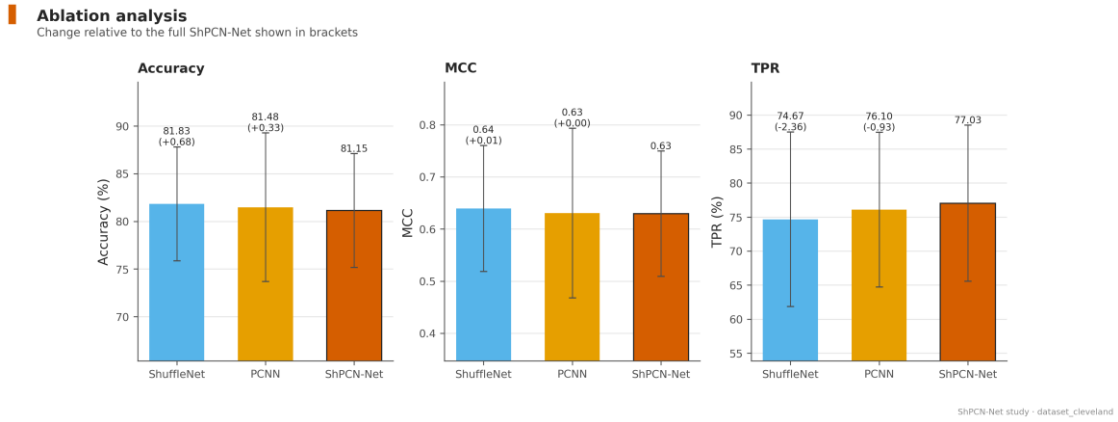


Fig. 23. Ablation analysis of ShPCN-Net: accuracy, MCC and TPR of the ShuffleNet variant, the PCNN variant and the full model.

Fig. 24 presents the scalability analysis, in which each method received the four to thirteen attributes most frequently selected by the GLO and was evaluated with five-fold cross-validation. ShPCN-Net achieved its highest mean accuracy of 83.83% with only four attributes, close to the 84.15% of

HCBiLSTM with the same attributes. XGBoost reached its best value of 84.83% with six attributes, and the Dense Autoencoder peaked at 83.50% with eight attributes. With all 13 attributes, the accuracies of the four methods ranged from 81.51% to 83.16%. This pattern indicates that the most frequently selected attributes carry most of the diagnostic information, so small subsets match or exceed the performance obtained with the complete attribute set.

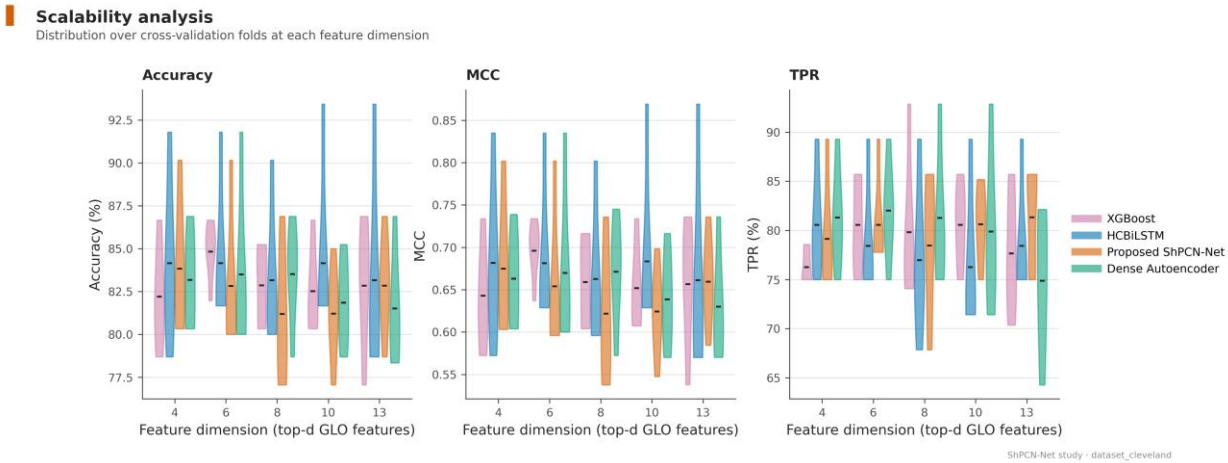


Fig. 24. Scalability analysis: distribution of accuracy, MCC and TPR over five folds for feature dimensions from 4 to 13.

Fig. 25 depicts the sensitivity of ShPCN-Net to the number of ShuffleNet-V2 units for training fractions from 50% to 90%, averaged over two repeated splits. Two units achieved the highest accuracy at most fractions, reaching 84.43% with 80% training data, whereas eight units reduced accuracy to 81.97% at the same fraction and to 64.52% at 90%. This decline with deeper configurations suggests that additional units increase the capacity beyond what the small training sets support. At 90% training data, the low values for all depths reflect the small test set of 31 records, where each misclassified record changes accuracy by 3.2 percentage points.

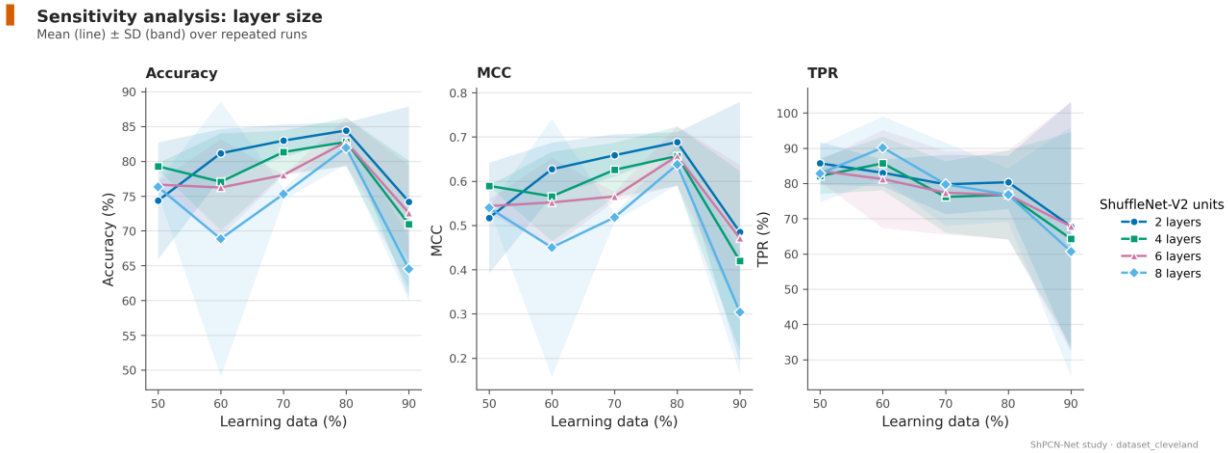


Fig. 25. Sensitivity of ShPCN-Net to the number of ShuffleNet-V2 units across training data percentages (mean ± standard deviation).

4.5.7 Explainability Results

Fig. 26 presents the SHAP summary of ShPCN-Net for the 60 explained records. The number of major vessels (ca) had the largest mean absolute SHAP value of 0.173, followed by thal (0.120), slope (0.118), cp (0.114) and age (0.049). High values of ca and thal increased the predicted disease probability, whereas low values decreased it, which agrees with the clinical meaning of these attributes. Age had a

small influence, with SHAP values mostly between −0.05 and 0.05. These results show that the network relies mainly on angiographic and thallium scan findings rather than on demographic attributes.

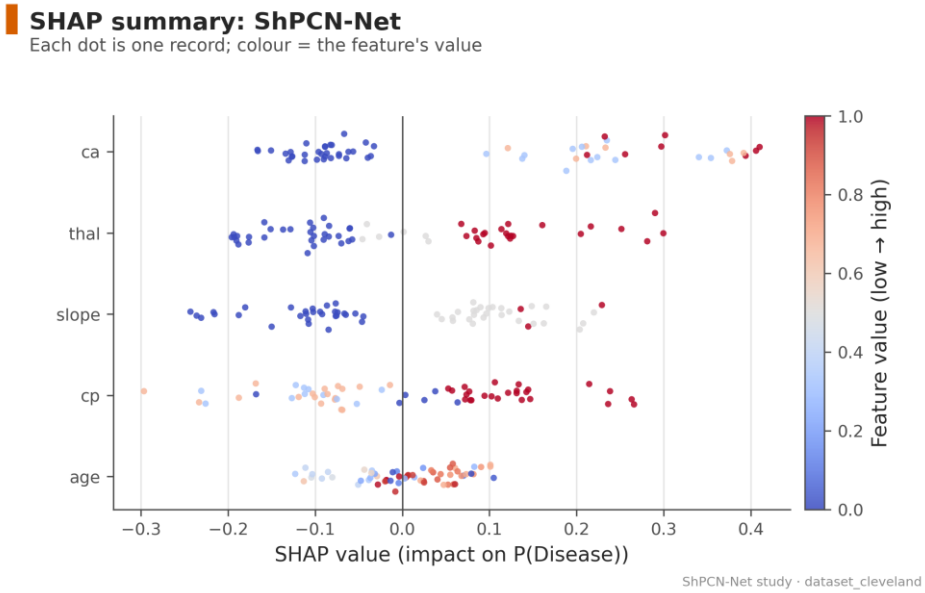


Fig. 26. SHAP summary of ShPCN-Net; each point is one record, and the colour indicates the normalized attribute value.

Fig. 27 shows the dependence of the SHAP value of ca on its value, coloured by thal. Records with no coloured vessels received negative SHAP values between −0.17 and −0.03, whereas records with one or more coloured vessels received positive values between 0.10 and 0.41. The colouring by thal shows that the positive contributions are not determined by thal alone, which suggests that the network combines both attributes rather than relying on a single threshold.

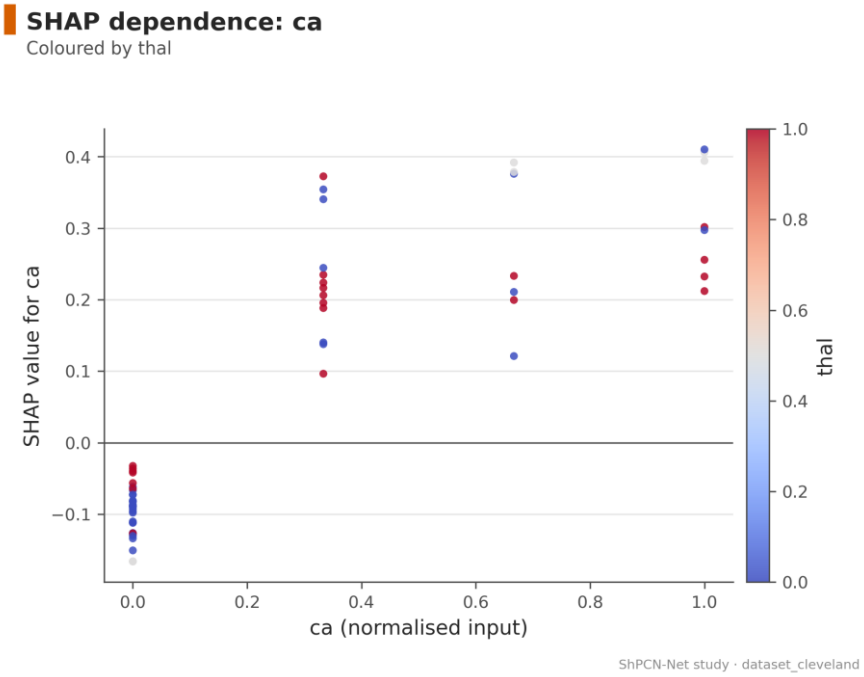


Fig. 27. SHAP dependence plot for the number of major vessels (ca), coloured by the thallium scan result (thal).

4.6 Statistical Findings and Experimental Discussion

Table 7 reports the paired t-test and Wilcoxon signed-rank test between ShPCN-Net and each competing method over the ten folds, and Fig. 28 displays the per-fold accuracy distributions. The mean accuracy differences ranged from −1.68 percentage points against XGBoost to +1.29 percentage points against the Dense Autoencoder. None of the differences was significant at the 0.05 level; the smallest p-values were 0.213 for the t-test and 0.266 for the Wilcoxon test, both against HCBiLSTM. Therefore, the experiments do not support a claim that ShPCN-Net improves or degrades accuracy relative to the competing methods on this dataset.

Table 7. Paired statistical comparison of ShPCN-Net with each competing method over the ten folds.

Comparison	Mean difference (pp)	t (df = 9)	p (t-test)	W	p (Wilcoxon)
ShPCN-Net vs HCBiLSTM	−1.66	−1.342	0.213	16.0	0.266
ShPCN-Net vs Dense Autoencoder	+1.29	0.696	0.504	21.5	0.563
ShPCN-Net vs XGBoost	−1.68	−0.972	0.357	20.5	0.508
ShPCN-Net vs ShuffleNet	−0.68	−0.820	0.433	18.0	0.375
ShPCN-Net vs PCNN	−0.33	−0.191	0.853	17.0	0.328

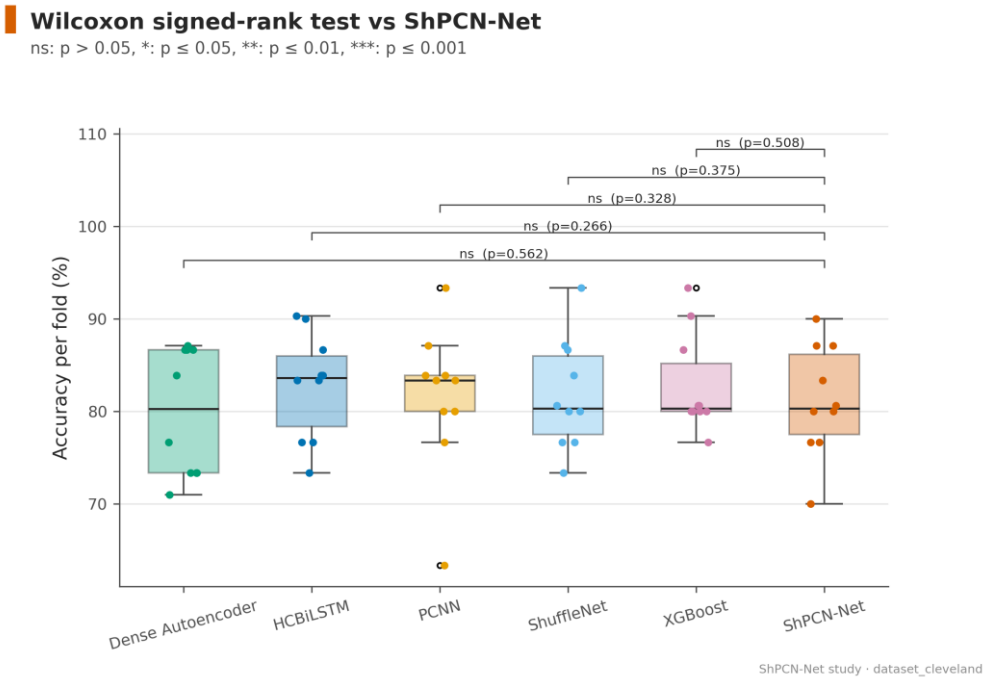


Fig. 28. Per-fold accuracy distributions of the six methods with Wilcoxon signed-rank test results against ShPCN-Net.

Overall, the experiments show that the GLO reduced the attribute set by more than half while the hybrid network maintained an accuracy comparable to the reproduced baselines. However, the full network did not outperform its own ablation variants, which indicates that the Taylor-series fusion added capacity without a measurable accuracy benefit on 303 records. Its strongest evidence lies in compactness and transparency: the four most frequent attributes reproduced the accuracy of the full attribute set, and the SHAP analysis attributed the predictions to clinically meaningful findings. The main limitation observed during experimentation was the high variance between folds, which reached standard deviations of 5–8 percentage points and prevented any method from establishing a significant advantage.

5. Discussion

The main finding of this study is that an optimized subset of approximately six clinical attributes allowed ShPCN-Net to reach 81.15% accuracy, which is statistically indistinguishable from published architectures evaluated under the same folds. Table 8 organizes the results by evaluation method. Under 10-fold cross-validation, XGBoost achieved the highest accuracy and MCC, and the gap to ShPCN-Net was 1.68 percentage points and 0.031. Averaged over the five training fractions, the gap widened to 5.27 percentage points, because ShPCN-Net performed poorly with only 50% training data. On the reference split, ShPCN-Net matched XGBoost exactly, while the Dense Autoencoder obtained the best result. These patterns indicate that the ranking of methods depends on the partition, so single-split comparisons can favour any of the methods.

Table 8. Summary of ShPCN-Net and the best competing method under each evaluation method.

Evaluation method	ShPCN-Net accuracy / MCC	Best competing method	Accuracy / MCC
Stratified 10-fold CV (mean)	81.15% / 0.630	XGBoost	82.83% / 0.661
Mean over 50–90% training data	80.17% / 0.621	XGBoost	85.43% / 0.713
Reference 90/10 split	83.87% / 0.693	Dense Autoencoder	87.10% / 0.739
Best feature dimension (5-fold CV)	83.83% (4 attributes)	XGBoost	84.83% (6 attributes)

Table 9 compares these results with values reported in published studies. Shrivastava et al. [9] reported 96.66% accuracy for HCBiLSTM on a test set of 60 records, and Alghamdi et al. [10] reported a mean accuracy of 99.67% with a 95/5 partition. In contrast, the re-implementations of the same pipelines reached 82.81% and 79.86% under stratified 10-fold cross-validation. Javeed et al. [7] obtained 93.33% with seven attributes on a single 70/30 split of the Cleveland records. The published values are not directly comparable with the present results, because they differ in partitions, attribute sets and, for Alharbi [12], Khan et al. [15] and Shorewala [17], in the dataset itself. Nevertheless, the difference between the published and reproduced values suggests that small test partitions may yield optimistic estimates on this benchmark.

Table 9. Accuracy reported in related studies and in this study; published values were obtained under different protocols and are not directly comparable.

Study	Dataset	Protocol	Accuracy
Shrivastava et al. [9]	Cleveland	Single split, 60 test records	96.66%
Alghamdi et al. [10]	Cleveland	95/5 split	99.67% (mean)
Javeed et al. [7]	Cleveland	70/30 hold-out, 7 attributes	93.33%
Nashif et al. [3]	Two open-access databases	10-fold CV	97.53%
Alharbi [12]	Combined, 1,190 records	80/20 split	89%
Khan et al. [15]	Health indicators, 253,680 records	10-fold CV	91%
Shorewala [17]	Cardiovascular, 70,000 records	K-fold CV	75.1%
This study (reproduced baselines)	Cleveland	Stratified 10-fold CV, identical folds	79.86–82.83%
This study (ShPCN-Net)	Cleveland	Stratified 10-fold CV, identical folds	81.15%

The comparable performance of ShPCN-Net can be attributed mainly to the attribute subset rather than to the hybrid architecture. Removing either branch did not reduce accuracy in the ablation analysis, and the sensitivity analysis showed that deeper configurations lowered accuracy, which suggests that the 303 records do not provide enough evidence to exploit the additional capacity. In contrast, the GLO consistently selected ca, cp and thal, and the SHAP analysis confirmed that these attributes drove the

predictions. This agreement between selection and explanation is consistent with the clinical role of angiographic and perfusion findings in diagnosing coronary disease. Moreover, the size of the selected subsets is close to the seven attributes with which Javeed et al. [7] obtained their best result on the same benchmark. From a practical perspective, a detector that requires five or six attributes reduces the number of inputs, although ca and thal still depend on fluoroscopy and thallium imaging. Consequently, the compact subsets simplify the input record but do not remove the dependence on specialised diagnostic tests.

Despite these results, several limitations should be acknowledged. First, the evaluation used a single cohort of 303 records, so the fold-to-fold variance prevents the detection of differences smaller than several percentage points. Second, the configuration search shared records with the later cross-validation folds, which may make the post-search estimates slightly optimistic for all methods alike. Third, the symmetric Gaussian transform discards the direction of deviation in continuous attributes, which may explain why thalach and oldpeak were rarely selected. Finally, the deep baselines were re-implemented with assumed layer sizes where the original studies omitted them.

6. Conclusion

This paper presented ShPCN-Net, an optimization-based feature selection and hybrid learning framework for heart disease detection from clinical records. The framework combined mean–mode imputation and Gaussian-function normalization with Gray Langurs Optimizer feature selection, and it processed the selected attributes with a ShuffleNet-V2 branch with squeeze-and-excitation recalibration, a regression branch, a Taylor-series fusion and a multi-scale parallel convolutional network. Under stratified 10-fold cross-validation on the Cleveland dataset, the framework retained 5.7 of 13 attributes on average and achieved 81.15% accuracy, an MCC of 0.630 and a TPR of 77.03%. These values were statistically indistinguishable from three re-implemented baselines and two ablation variants, which reached accuracies between 79.86% and 82.83%. The four most frequently selected attributes reproduced the accuracy of the complete attribute set, and Shapley additive explanations attributed the predictions mainly to the number of major vessels, thallium scan result, ST slope and chest pain type. However, the hybrid network did not improve accuracy over its ablation variants or the strongest baselines. In future work, the framework will be evaluated on larger multi-centre cohorts, and a direction-preserving normalization will be investigated to retain the information carried by continuous attributes.

References

[1] A. Calzoni, M. Savardi, M. Silvestri, S. Benini, A. Signoroni, "Bimodal ECG and PCG Cardiovascular Disease Detection: Exploring the Potential and Modality Contribution", Journal of Medical Systems, Vol. 49, Art. no. 113, 2025, DOI: 10.1007/s10916-025-02245-5

[2] A. Abdellatif, H. Abdellatef, J. Kanesan, C.-O. Chow, J. H. Chuah, H. M. Gheni, "An Effective Heart Disease Detection and Severity Level Classification Model Using Machine Learning and Hyperparameter Optimization Methods", IEEE Access, Vol. 10, pp. 79974–79985, 2022, DOI: 10.1109/ACCESS.2022.3191669

[3] S. Nashif, M. R. Raihan, M. R. Islam, M. H. Imam, "Heart Disease Detection by Using Machine Learning Algorithms and a Real-Time Cardiovascular Health Monitoring System", World Journal of Engineering and Technology, Vol. 6, No. 4, pp. 854–873, November, 2018, DOI: 10.4236/wjet.2018.64057

[4] B. A. Tama, S. Im, S. Lee, "Improving an Intelligent Detection System for Coronary Heart Disease Using a Two-Tier Classifier Ensemble", BioMed Research International, Vol. 2020, Art. ID 9816142, 2020, DOI: 10.1155/2020/9816142

[5] V. Chang, V. R. Bhavani, A. Q. Xu, M. A. Hossain, "An artificial intelligence model for heart disease detection using machine learning algorithms", Healthcare Analytics, Vol. 2, Art. no. 100016, 2022, DOI: 10.1016/j.health.2022.100016

[6] U. Nagavelli, D. Samanta, P. Chakraborty, "Machine Learning Technology-Based Heart Disease Detection Models", Journal of Healthcare Engineering, Vol. 2022, Art. ID 7351061, 2022, DOI: 10.1155/2022/7351061

[7] A. Javeed, S. Zhou, L. Yongjian, I. Qasim, A. Noor, R. Nour, "An Intelligent Learning System Based on Random Search Algorithm and Optimized Random Forest Model for Improved Heart Disease Detection", IEEE Access, Vol. 7, pp. 180235–180243, 2019, DOI: 10.1109/ACCESS.2019.2952107

[8] P. Shah, M. Shukla, N. H. Dholakia, H. Gupta, "Predicting cardiovascular risk with hybrid ensemble learning and explainable AI", Scientific Reports, Vol. 15, Art. no. 17927, 2025, DOI: 10.1038/s41598-025-01650-7

[9] P. K. Shrivastava, M. Sharma, P. Sharma, A. Kumar, "HCBiLSTM: A hybrid model for predicting heart disease using CNN and BiLSTM algorithms", Measurement: Sensors, Vol. 25, Art. no. 100657, 2023, DOI: 10.1016/j.measen.2022.100657

[10] N. S. Alghamdi, M. Zakariah, A. Shankar, W. Viriyasitavat, "Heart disease prediction using autoencoder and DenseNet architecture", Egyptian Informatics Journal, Vol. 28, Art. no. 100559, 2024, DOI: 10.1016/j.eij.2024.100559

[11] UCI Machine Learning Repository, "Heart Disease" dataset, https://archive.ics.uci.edu/dataset/45/heart+disease

[12] L. A. Alharbi, "Heart Disease Prediction of Cleveland Clinic Patients using Advanced Machine Learning Algorithms", Journal of Advanced Research Design, Vol. 126, No. 1, pp. 1–14, 2025, DOI: 10.37934/ard.126.1.114

[13] B. Ramesh, K. Lakshmanna, "A Novel Early Detection and Prevention of Coronary Heart Disease Framework Using Hybrid Deep Learning Model and Neural Fuzzy Inference System", IEEE Access, Vol. 12, pp. 26683–26695, 2024, DOI: 10.1109/ACCESS.2024.3366537

[14] S. Ketu, P. K. Mishra, "Empirical Analysis of Machine Learning Algorithms on Imbalance Electrocardiogram Based Arrhythmia Dataset for Heart Disease Detection", Arabian Journal for Science and Engineering, 2021, DOI: 10.1007/s13369-021-05972-2

[15] H. Khan, N. Javaid, T. Bashir, M. Akbar, N. Alrajeh, S. Aslam, "Heart Disease Prediction Using Novel Ensemble and Blending Based Cardiovascular Disease Detection Networks: EnsCVDD-Net and BlCVDD-Net", IEEE Access, Vol. 12, pp. 109230–109254, 2024, DOI: 10.1109/ACCESS.2024.3421241

[16] A. A. Hipparkar, R. R. Chakre, "Hybrid Framework Combining Swarm Optimization and Deep Learning for Cardiovascular Disease Classification", Journal of Intelligent Decision Making and Information Science, Vol. 3, No. 9s, pp. 2420–2445, 2026

[17] V. Shorewala, "Early detection of coronary heart disease using ensemble techniques", Informatics in Medicine Unlocked, Vol. 26, Art. no. 100655, 2021, DOI: 10.1016/j.imu.2021.100655

[18] S. Barshandeh, N. Khodadadi, B. Abdollahzadeh, A. Mohammadzadeh, E.-S. M. El-Kenawy, M. M. Eid, K. M. Mosalam, "Gray langurs optimizer: a multi-group bio-inspired optimization algorithm", Artificial Intelligence Review, Vol. 59, Art. no. 166, 2026, DOI: 10.1007/s10462-026-11529-2

[19] Y. Qiu, R. Li, X. Zhang, "Simultaneous SVM Parameters and Feature Selection Optimization Based on Improved Slime Mould Algorithm", IEEE Access, Vol. 12, pp. 18215–18236, 2024, DOI: 10.1109/ACCESS.2024.3351943

[20] J. Liu, Y. Chen, K. Huang, X. Guan, "Enhancing Missense Variant Pathogenicity Prediction with MissenseNet: Integrating Structural Insights and ShuffleNet-Based Deep Learning Techniques", Biomolecules, Vol. 14, No. 9, Art. no. 1105, 2024, DOI: 10.3390/biom14091105

[21] C. Li, M. Liang, W. Song, K. Xiao, "A Multi-Scale Parallel Convolutional Neural Network Based Intelligent Human Identification Using Face Information", Journal of Information Processing Systems, Vol. 14, No. 6, pp. 1494–1507, December, 2018, DOI: 10.3745/JIPS.02.0103

[22] R. W. Anwer, M. Abrar, M. Ullah, A. Salam, F. Ullah, "Advanced intrusion detection in the industrial Internet of Things using federated learning and LSTM models", Ad Hoc Networks, Vol. 178, Art. no. 103991, 2025, DOI: 10.1016/j.adhoc.2025.103991

[23] Y.-S. Kim, M. K. Kim, N. Fu, J. Liu, J. Wang, J. Srebric, "Investigating the impact of data normalization methods on predicting electricity consumption in a building using different artificial neural network models", Sustainable Cities and Society, Vol. 118, Art. no. 105570, 2025, DOI: 10.1016/j.scs.2024.105570

[24] E. Srivastava, S. Mohan, T. K. Gandhi, A. K. Choudhury, S. Kumar, "Attention-driven graph-based machine learning for non-invasive diagnosis of NAFLD", Intelligence-Based Medicine, Vol. 12, Art. no. 100288, 2025, DOI: 10.1016/j.ibmed.2025.100288

[25] L. Xiaomeng, L. Bingjie, L. Huimin, "Intelligent Detection and Early Warning of Power System Cybersecurity Threats Based on Multi-modal Large Language Models", Journal of Cyber Security and Mobility, Vol. 14, No. 6, pp. 1347–1372, 2025, DOI: 10.13052/jcsm2245-1439.1463