A Langur-Optimized Hybrid Convolutional Method for Heart Disease Detection

[First A. Author], *[Member, IEEE]*, [Second B. Author], and [Third C. Author], *[Member, IEEE]*

***Abstract*— Heart disease remains a leading cause of death worldwide, and clinical decision support depends on classifiers that stay reliable when patient records are few, redundant, and heterogeneous. Many reported prediction models for the Cleveland heart disease records combine feature selection and deep learning, but their gains are rarely tested under protocols that keep all data-dependent steps inside the training folds. This study proposes and evaluates ShPCN-Net, a hybrid framework that cleans the records, normalizes numerical attributes with a Gaussian function, selects attributes with a Gray Langurs Optimizer wrapper, and classifies the selected attributes with a learned feature map processed by a ShuffleNet-V2 attention branch, a Taylor regression fusion stage, and a multiscale parallel convolutional head. The framework is assessed on all 303 Cleveland records with stratified ten-fold cross-validation, against two ablation networks and three published model families, using accuracy, the Matthews correlation coefficient, the true positive rate, the area under the receiver operating characteristic curve, Wilcoxon signed-rank tests, and Shapley additive explanations. The optimizer reduced the 13 attributes to 6.9 attributes per fold on average, consistently retaining the vessel count and thalassemia test. ShPCN-Net reached an accuracy of 0.805, a Matthews correlation coefficient of 0.610, a true positive rate of 0.755, and an area under the curve of 0.897, whereas the six evaluated models spanned accuracies from 0.785 to 0.815 without any significant pairwise difference. The results show that compact, interpretable attribute subsets are attainable, but the hybrid fusion offers no measurable advantage over simpler networks on this dataset.**

***Index Terms*— Convolutional neural network, explainable artificial intelligence, feature selection, heart disease detection, metaheuristic optimization**

I. Introduction

CARDIOVASCULAR diseases are the leading cause of death worldwide; the World Health Organization estimates that they caused about 17.9 million deaths in 2019, nearly one third of all global deaths [1]. Coronary heart disease accounts for a large share of this burden, and its diagnosis still relies on a combination of symptoms, exercise testing, imaging, and invasive angiography. Routine clinical attributes, such as age, chest pain type, resting blood pressure, serum cholesterol, exercise response, and fluoroscopic findings, are collected for nearly every patient and therefore offer an inexpensive basis for computer-aided risk stratification [2]. The Cleveland heart disease records, published through the University of California Irvine repository, have become the most widely used benchmark for this task because they couple 13 such attributes with an angiographic diagnosis [2], [3].

Automated prediction from these records matters because a model that flags high-risk patients early can prioritize scarce diagnostic resources, whereas a model that misses diseased patients or raises frequent false alarms can erode clinical trust. The records are, however, small, mixed in type, and partly redundant: the numerical vital signs and the coded test results carry overlapping information, and several attributes contribute little to discrimination [4]. Such properties make predictive performance sensitive to the choice of attributes and to the way preprocessing interacts with validation. When attribute selection or normalization uses the complete dataset before cross-validation, the estimated accuracy can be markedly optimistic [5], [6]. Reliable, interpretable, and honestly validated predictors are therefore as important as raw accuracy for clinical decision support.

Existing approaches span classical classifiers, ensembles, and deep networks. Gradient boosting, notably extreme gradient boosting (XGBoost) [7], and stacked ensembles of boosting and neural models have been compared on the Cleveland data and related risk-factor datasets [8]–[10]. Deep hybrids combine convolutional and recurrent layers, such as the convolutional bidirectional long short-term memory network HCBiLSTM [11], or pair autoencoders with densely connected classifiers [12]; several of these studies report accuracies above 0.96 on the 303 Cleveland records. Multimodal signal models fuse electrocardiogram and phonocardiogram channels to detect cardiovascular disease [13]. Metaheuristic wrappers, including the improved slime mould algorithm, select attribute subsets by optimizing classifier accuracy and subset size jointly [14], [15], and a recent multi-group metaheuristic, the Gray Langurs Optimizer (GLO), has been proposed for general optimization [16]. Convolutional designs such as ShuffleNet [17], [18], squeeze-and-excitation attention [19], ShuffleNet-based tabular classification [20], and multiscale parallel convolutional networks [21] are efficient feature extractors, and tabular records can be mapped onto image-like arrays for convolutional processing [22]. The research problem addressed here is whether combining such a metaheuristic attribute search with a hybrid convolutional classifier improves heart disease detection when every data-dependent step is confined to the training folds. Three gaps remain open: the GLO has

Manuscript received [Month Day, Year]; revised [Month Day, Year]; accepted [Month Day, Year]. This work was supported in part by [funding agency] under Grant [number]. (Corresponding author: [Name].)

[First A. Author] is with [Department, Institution, City, Country] (e-mail: [email]). [Second B. Author] is with [Department, Institution, City, Country] (e-mail: [email]). [Third C. Author] is with [Department, Institution, City, Country] (e-mail: [email]).

not been examined as an attribute selector for clinical records, efficient image-domain architectures have rarely been fused for tabular cardiac data, and the reported gains of hybrid predictors are seldom accompanied by paired statistical tests, class-balanced metrics, and explanations.

This paper develops ShPCN-Net, a hybrid framework for binary heart disease detection from routine clinical attributes. The framework first cleans and normalizes the records, then searches for a compact attribute subset with a GLO wrapper, and finally classifies the selected attributes with a network that joins a lightweight attention branch and a multiscale convolutional branch through a fusion rule derived from a first-order Taylor expansion. The framework is evaluated under a leakage-controlled protocol, compared with ablated and published model families, and interpreted with model-agnostic explanations.

The contributions of this paper are as follows. First, it formulates a GLO-based wrapper for clinical attribute selection, with a solution encoding, a multi-group update scheme, and a fitness function that balances wrapper accuracy against subset size. Second, it introduces Taylor regression fusion, which combines a linear regression of the selected attributes with the output of a ShuffleNet-V2 attention branch before a multiscale parallel convolutional head. Third, it provides a leakage-controlled evaluation in which imputation, normalization, and attribute selection are refitted inside each of ten stratified folds, and in which ShPCN-Net is compared with two ablation networks and three published model families using the Matthews correlation coefficient (MCC) [23], [24], Wilcoxon signed-rank tests [25], [26], and Shapley additive explanations (SHAP) [27]. Fourth, it reports the outcome transparently, including the finding that the proposed fusion does not significantly outperform simpler networks on this dataset.

The remainder of this paper is organized as follows. Section II presents the motivation, the literature survey, and the research gaps. Section III describes the proposed methodology, from record preprocessing through GLO attribute selection to the ShPCN-Net classifier and its training. Section IV reports the experimental setup, dataset, parameter settings, metrics, intermediate outputs, comparative results, and statistical findings. Section V discusses the findings in relation to previous studies, and Section VI concludes the paper with limitations and future directions.

II. Motivation

The motivation for this study arises from a tension in the heart disease prediction literature. Increasingly complex hybrids are reported with very high accuracy on a dataset of only 303 records, yet a small dataset leaves little room to separate genuine architectural gains from variance and from optimistic validation. A model that selects attributes and learns representations should therefore be judged by its behavior under strict cross-validation, by class-balanced metrics, by paired statistical tests against simpler alternatives, and by explanations that clinicians can inspect. This study pursues such an evaluation for a hybrid built from components that the concept of this work identified as promising: a new multi-group optimizer for attribute selection and two efficient convolutional designs.

A. Literature Survey

Shrivastava *et al.* [11] proposed HCBiLSTM, which selects attributes with an extra-trees classifier, handles missing and imbalanced data, and classifies the Cleveland records with a convolutional bidirectional long short-term memory network, reporting an accuracy of 96.66% together with precision, recall, and F1-score; the reported evaluation does not include a class-balanced correlation metric, a paired statistical test against simpler models, or an explanation of the predictions, which the present study adds. Alghamdi *et al.* [12] combined a dense autoencoder with a DenseNet-style classifier on the same 13 Cleveland attributes and reported a mean accuracy of 99.67%; near-perfect figures on 303 records are highly sensitive to how preprocessing and validation are partitioned [5], [6], which motivates the fold-internal protocol adopted here. Alharbi [8] compared XGBoost with multilayer perceptron, k-nearest neighbor, support vector, and decision tree classifiers on Cleveland Clinic patients and identified XGBoost as the preferred model; that comparison concerns the choice of classifier rather than a search for compact attribute subsets, a question the present wrapper addresses.

Shorewala [9] studied coronary heart disease prediction with k-nearest neighbor, logistic, and naive Bayes classifiers and their bagged, boosted, and stacked ensembles under k-fold validation, reporting that bagging raised accuracy by 1.96% on average; the study used a 70 000-record risk-factor dataset rather than angiography-confirmed records and did not examine learned representations or explanations, both of which are evaluated here on the Cleveland data. Shah *et al.* [10] proposed a stacked ensemble of gradient boosting, CatBoost, and neural networks for cardiovascular risk prediction and interpreted it with SHAP values and t-distributed stochastic neighbor embedding (t-SNE) and principal component projections, reporting an area under the receiver operating characteristic curve of 0.82; its explanations are computed on the full attribute set, without an explicit search for a smaller subset, whereas the present framework explains a model trained on an optimized subset. Calzoni *et al.* [13] fused electrocardiogram and phonocardiogram branches by late fusion, reached an area under the curve of 96.4% on an augmented bimodal dataset, and quantified modality contributions with three explanation techniques; the approach requires synchronized signal acquisition, which is unavailable when only routine clinical attributes are recorded, the setting targeted here.

Liu *et al.* [20] built MissenseNet on the ShuffleNet architecture with an encoder-decoder framework and a squeeze-and-excitation module to classify the pathogenicity of missense variants, achieving the highest areas under the receiver operating characteristic and precision-recall curves on an independent test set; its design was validated on genomic and structural features, so its transfer to clinical cardiac attributes remained untested before the present study. Li *et al.* [21] proposed a multiscale parallel convolutional neural

network that extracts and fuses features at several scales for face recognition and validated it on the CMU-PIE, extended FERET, and AR databases; the parallel multiscale idea was developed for images, and the present work examines it on learned maps of tabular records. Qiu *et al.* [14] introduced an improved slime mould algorithm that optimizes support vector machine parameters and attribute subsets simultaneously, with a fitness that weighs classification error against subset size, and reported accuracies of at least 90% on seven of ten repository datasets; its wrapper is coupled to a support vector machine, whereas the present framework transfers the selected subset to a deep hybrid classifier. Barshandeh *et al.* [16] introduced the GLO, which models the one-male, multi-male, and all-male groups of gray langurs, their social hierarchy, and their migration after alpha death, mate finding, and puberty; the algorithm was presented as a general-purpose optimizer, and its use for clinical attribute selection is examined here for the first time to our knowledge, although its relative merit against other optimizers is not tested in this study.

B. Research Gap

The survey identifies the following practical problems, unresolved issues, and expected value, each mapped to a component of the proposed method and to the evidence that this study provides.

- *Optimistic validation of small-sample predictors.* Very high accuracies are reported on the 303 Cleveland records, but attribute selection and normalization are not always shown to be confined to training data [5], [6]. *Component:* fold-internal preprocessing and GLO search (Section III-G). *Status:* addressed methodologically; the resulting estimates (Section IV) are lower than the reported literature values, as expected under strict separation.
- *Redundant and weakly informative attributes.* Mixed-type records contain overlapping attributes [4]. *Component:* GLO wrapper attribute selection (Section III-D). *Status:* partially addressed; the subset size fell from 13 to 6.9 attributes per fold on average, and the scalability analysis shows similar accuracy for compact and full subsets, but the GLO was not compared with other selectors under the same protocol.
- *Untested transfer of efficient image-domain networks to tabular cardiac data.* ShuffleNet-style attention and multiscale parallel convolutions were validated on genomic and image data [20], [21]. *Component:* the learned input map and the two branches of ShPCN-Net (Section III-E). *Status:* addressed as an evaluation; both branches perform comparably to established baselines.
- *Fusion of complementary representations.* Hybrid predictors are expected to benefit from combining a linear view of the attributes with convolutional features. *Component:* Taylor regression fusion (Section III-E.3). *Status:* not supported by the evidence; ShPCN-Net did not significantly outperform its ablated branches.
- *Limited statistical testing, class-balanced metrics, and explanations. Component:* MCC, Wilcoxon signed-rank tests, and SHAP analysis (Sections III-G and IV). *Status:* addressed; all comparisons are reported with paired tests and explanations.

III. Proposed Langur-Optimized ShPCN-Net Methodology

A. Overview of the ShPCN-Net Detection Pipeline

B. Clinical Record Representation

Let $D = \{(\mathbf{x}_j, c_j)\}_{j=1}^{n}$ denote the heart disease dataset, where $\mathbf{x}_j \in \mathbb{R}^K$ holds the $K = 13$ clinical attributes of record j, $c_j \in \{0, 1\}$ is its diagnosis (0 for Normal and 1 for Disease), and $n = 303$. The attributes split into a numerical set $\mathcal{N}$ (age, resting blood pressure, serum cholesterol, maximum heart rate, and ST depression) and a categorical set $\mathcal{C}$ (sex, chest pain type, fasting blood sugar, resting electrocardiogram, exercise-induced angina, ST slope, number of major vessels, and thalassemia test), which Section IV-B describes in detail. The value of attribute k in record j is written x_{jk}. During cross-validation the index set of the training records of a fold is denoted $\mathcal{T}$, and every statistic below, such as a mean or a mode, is computed over $\mathcal{T}$ only.

C. Leakage-Controlled Preprocessing of Clinical Records

1) Duplicate Removal: Repeated records inflate the apparent agreement between training and test data when copies fall on both sides of a split. The first step therefore removes exact duplicates, so that the cleaned dataset D' keeps one instance of every distinct attribute-label combination. This operation uses no learned statistic and is applied once, before the folds are formed. The Cleveland records used in this study contain no duplicate rows, so $D' = D$, but the step is retained so that the pipeline remains valid for other record collections.

2) Missing-Value Imputation: Missing entries are replaced by the training mean for numerical attributes and by the training mode for categorical attributes, following the imputation rule of the concept design,

$$\tilde{x}_{jk} = \begin{cases} \mu_k = \frac{1}{|\mathcal{T}|} \sum_{i \in \mathcal{T}} x_{ik}, & k \in \mathcal{N} \\ \mathrm{mode}_{i \in \mathcal{T}}\,(x_{ik}), & k \in \mathcal{C} \end{cases} \tag{1}$$

where $\tilde{x}_{jk}$ is the imputed value, μ_k is the training mean of attribute k, and the mode is the most frequent training value. The mean preserves the first moment of a continuous attribute, and the mode keeps a categorical attribute inside its valid code set, which a mean would violate. The imputed matrix is denoted M_v. The version of the Cleveland records analyzed here contains no missing entries, so the rule leaves the data unchanged; it is nevertheless fitted per fold for consistency.

3) Categorical Encoding: Categorical attributes are kept as integer codes rather than one-hot vectors so that the attribute count remains $K = 13$ and the search space of the optimizer stays at 2^{13} subsets. To give the codes a range comparable with the normalized numerical attributes, each code is divided by the largest code observed in the training fold,

$$z_{jk} = \frac{\tilde{x}_{jk}}{\max_{i \in \mathcal{T}} \tilde{x}_{ik}}, \quad k \in \mathcal{C} \tag{2}$$

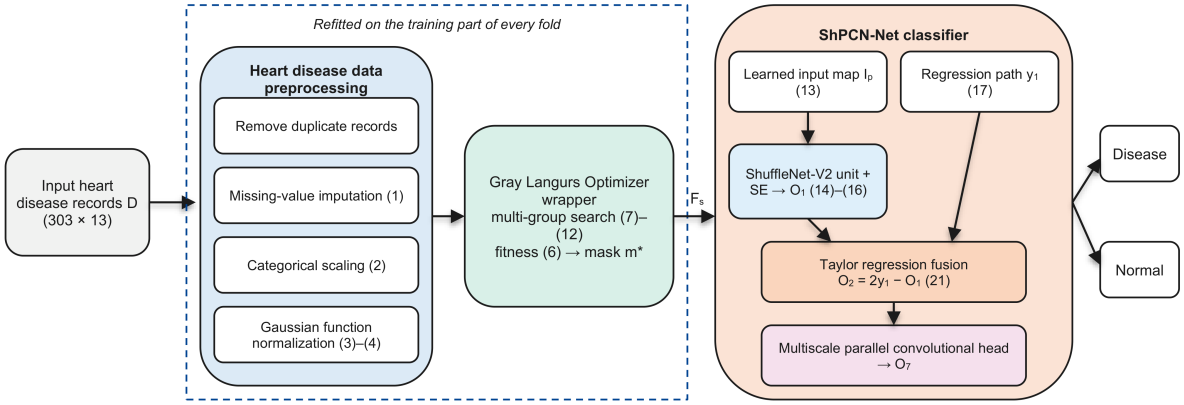


Fig. 1. Block diagram of the proposed ShPCN-Net framework. Preprocessing and Gray Langurs attribute selection are refitted inside every training fold; the selected attributes feed the ShuffleNet-V2 attention branch and the regression path, which Taylor regression fusion combines before the multiscale parallel convolutional head.

where $z_{jk} \in [0,1]$ is the scaled code. Binary attributes such as sex and exercise-induced angina remain 0 or 1.

4) Gaussian Function Normalization: Numerical attributes are transformed with the Gaussian probability density function evaluated at the attribute value, a transformation that Kim *et al.* [28] examined among several normalization methods for neural predictors,

$$g_k(\tilde{x}_{jk}) = \frac{1}{\sigma_k\sqrt{2\pi}} \exp\left(-\frac{(\tilde{x}_{jk}-\mu_k)^2}{2\sigma_k^2}\right), \quad k \in \mathcal{N} \tag{3}$$

where μ_k and σ_k are the training mean and standard deviation of attribute k. The raw density in (3) has a maximum of $1/(\sigma_k\sqrt{2\pi})$, which differs between attributes by more than an order of magnitude for the Cleveland data (for example, σ is 9.02 years for age and 51.69 mg/dL for cholesterol). A wrapper based on a radial basis function kernel would then effectively ignore attributes with a large standard deviation. The framework therefore multiplies (3) by the per-attribute constant $\sigma_k\sqrt{2\pi}$,

$$\hat{x}_{jk} = \sigma_k\sqrt{2\pi}\, g_k(\tilde{x}_{jk}) = \exp\left(-\frac{(\tilde{x}_{jk}-\mu_k)^2}{2\sigma_k^2}\right) \tag{4}$$

which preserves the Gaussian shape but maps every numerical attribute to the interval $(0,1]$. The Gaussian mapping compresses extreme values smoothly and bounds the influence of outliers, which motivated its choice in the concept design. It is, however, symmetric about the mean: a value one standard deviation above the mean and a value one standard deviation below it receive the same $\hat{x}_{jk}$. This property removes the direction of the deviation, a limitation that the experiments examine indirectly through the selection frequency of the numerical attributes. The normalized matrix N_d concatenates $\hat{x}_{jk}$ for $k \in \mathcal{N}$ and z_{jk} for $k \in \mathcal{C}$ in the original attribute order.

D. Gray Langurs Wrapper Attribute Selection

Attribute selection is posed as a binary optimization problem over the K normalized attributes. A wrapper evaluates candidate subsets with a classifier and therefore captures interactions that filter criteria miss [4], [15]. The GLO was chosen as the search engine because the concept of this work targets this recent multi-group optimizer, whose social structure keeps several partly independent subpopulations and thus supports exploration in a rugged and discrete fitness landscape [16]. The no-free-lunch theorem [29] implies that no optimizer dominates on all problems, so the suitability of the GLO for clinical attribute selection is an empirical question; this study addresses it by reporting the convergence and selection behavior of the GLO, not by claiming superiority over other optimizers.

1) Solution Encoding: Each langur i holds a continuous position $\mathbf{p}_i \in [0,1]^K$. The position decodes into a binary attribute mask by thresholding,

$$m_{ik} = \begin{cases} 1, & p_{ik} > \tau \\ 0, & \text{otherwise} \end{cases} \tag{5}$$

where $m_{ik} = 1$ means that attribute k is selected and $\tau = 0.5$. If no attribute exceeds the threshold, the attribute with the largest position value is selected, so every langur represents a valid subset. The decoded mask is equivalent to the list of selected attribute indices of the concept design, and the continuous representation allows the arithmetic update rules below.

2) Subset Fitness: The quality of a mask $\mathbf{m}$ is measured by the fitness of Qiu *et al.* [14], which is minimized:

$$f(\mathbf{m}) = \alpha\,(1 - A_w(\mathbf{m})) + (1-\alpha)\frac{R}{K} \tag{6}$$

where $A_w(\mathbf{m})$ is the wrapper accuracy of the subset, $R = \sum_k m_k$ is the number of selected attributes, and $\alpha \in [0,1]$ weighs accuracy against compactness. With $\alpha = 0.99$, one percentage point of accuracy outweighs the removal of one attribute, so the search prefers small subsets only when they cost little accuracy. The wrapper accuracy is the mean accuracy of a support vector machine with a radial basis function kernel [30] under stratified five-fold cross-validation on the training

records of the current outer fold. A support vector machine was chosen as the wrapper because it trains in milliseconds on a few hundred records, whereas training ShPCN-Net inside each of roughly one thousand fitness evaluations would be prohibitive. Identical masks are cached, so repeated visits to the same subset cost nothing.

3) Multi-Group Position Update: The population of P langurs is divided into G groups whose types cycle through the three social structures described by Barshandeh *et al.* [16]: one-male, multi-male, and all-male groups. A control parameter decreases linearly over the T iterations,

$$a(t) = a_0 - (a_0 - a_1)\frac{t}{T-1} \tag{7}$$

where t is the iteration index, $a_0 = 2$, and $a_1 = 0$; a large a favors exploration and a small a favors exploitation, in the same spirit as the encircling coefficient of the grey wolf optimizer [31]. Let $\mathbf{A}_g$ be the best position (the alpha male) of group g, $\mathbf{M}_g$ the mean of its three best positions, $\mathbf{G}^*$ the best position of the whole population, and $\mathbf{r}_1$, $\mathbf{r}_2$, and $\mathbf{r}_3$ vectors of independent uniform random numbers in $[0, 1]$. A member of a one-male group follows its single alpha,

$$\mathbf{p}_i' = \mathbf{A}_g - a\,(2\mathbf{r}_1 - 1) \odot |\mathbf{r}_2 \odot \mathbf{A}_g - \mathbf{p}_i| \tag{8}$$

where $\odot$ denotes element-wise multiplication. A member of a multi-male group follows the mean of its leading males and is additionally drawn toward the population best as the search progresses,

$$\begin{aligned}\mathbf{p}_i' = \mathbf{M}_g &- a\,(2\mathbf{r}_1 - 1) \odot |\mathbf{r}_2 \odot \mathbf{M}_g - \mathbf{p}_i| \\ &+ \left(1 - \frac{a}{a_0}\right) \mathbf{r}_3 \odot (\mathbf{G}^* - \mathbf{p}_i)\end{aligned} \tag{9}$$

A member of an all-male group behaves as a roaming bachelor that moves relative to a randomly chosen langur j and, for mate finding, toward the population best,

$$\begin{aligned}\mathbf{p}_i' = \mathbf{p}_i &+ a\,(2\mathbf{r}_1 - 1) \odot (\mathbf{p}_j - \mathbf{p}_i) \\ &+ \left(1 - \frac{a}{a_0}\right) \mathbf{r}_2 \odot (\mathbf{G}^* - \mathbf{p}_i)\end{aligned} \tag{10}$$

With probability $\rho = 0.1$, a langur instead performs an arbitrary movement within its territory,

$$\mathbf{p}_i' = \mathbf{p}_i + a\,\boldsymbol{\xi}, \quad \boldsymbol{\xi} \sim \mathcal{N}\left(\mathbf{0},\, (0.1\,(u-l))^2\mathbf{I}\right) \tag{11}$$

where $l = 0$ and $u = 1$ are the lower and upper bounds. Every candidate is clipped to $[l, u]$, and the langur moves only if the candidate improves its fitness,

$$\mathbf{p}_i \leftarrow \begin{cases} \mathbf{p}_i', & f(\mathbf{p}_i') < f(\mathbf{p}_i) \\ \mathbf{p}_i, & \text{otherwise} \end{cases} \tag{12}$$

so the best fitness of the population never increases. The published description of the GLO specifies the group types, the hierarchy, and the migration events, but its exact update coefficients were not available when this study was implemented. Equations (7)–(12) therefore constitute this study's formulation of the GLO, and the label GLO in this paper refers to this formulation.

Algorithm 1 Gray Langurs wrapper attribute selection for one training fold.

Require: normalized training matrix $N_d^{\mathcal{T}}$, labels $c^{\mathcal{T}}$, settings $P, G, T, \alpha, \tau, L, M$
Ensure: attribute mask $\mathbf{m}^*$ and convergence curve
Initialize $\mathbf{p}_i \sim U(0,1)^K$ for $i = 1, \ldots, P$; evaluate f by (6) with the cached RBF-SVM wrapper
Assign langur i to group $i \bmod G$; group types cycle one-male, multi-male, all-male
for $t = 0$ to $T - 1$ **do**
  compute $a(t)$ by (7); find $\mathbf{G}^*$
  for each group g **do**
    find $\mathbf{A}_g$ and $\mathbf{M}_g$
    for each member i of g **do**
      with probability ρ draw $\mathbf{p}_i'$ by (11); **otherwise** apply (8), (9), or (10) by group type
      clip $\mathbf{p}_i'$ to $[l, u]$; decode the mask by (5); accept by (12)
    end for
    if the group best has not improved for L iterations **then** rebirth of the worst member near $\mathbf{G}^*$
  end for
  if $(t+1) \bmod M = 0$ **then** swap the worst member of each group with a member of the next group
  record $\min_i f(\mathbf{p}_i)$
end for
return the mask of the best langur by (5)

4) Alpha Death and Inter-Group Migration: Two population events model the group dynamics of gray langurs. If the best fitness of a group does not improve for $L = 6$ consecutive iterations, its alpha is considered dead: the worst member of the group is reborn near the population best, $\mathbf{p}_w \leftarrow \mathrm{clip}(\mathbf{G}^* + \boldsymbol{\eta})$ with $\boldsymbol{\eta} \sim \mathcal{N}(\mathbf{0}, 0.15^2\mathbf{I})$, which represents a new male joining after puberty. Every $M = 5$ iterations, the worst member of each group swaps places with a random member of the next group, which represents migration between groups and keeps the group sizes constant. These events inject diversity when a group stagnates while preserving the best solutions found so far.

5) Search Cost: Each iteration evaluates at most P candidate positions, plus one rebirth per stagnating group, so the search requires at most about $P(T + 1) + GT$ fitness evaluations. With $P = 20$, $G = 4$, and $T = 50$, this bound is about 1 220 evaluations, but caching reduces the number of distinct subsets that are actually trained to a few hundred per fold. Algorithm 1 summarizes the complete selection procedure.

E. ShPCN-Net Hybrid Classifier

1) Learned Input Map: Convolutional layers exploit local spatial structure, but a vector of clinical attributes has no natural spatial order. Rather than fixing an arbitrary arrangement, ShPCN-Net learns one. A fully connected layer projects the selected attributes onto $H \times W = 8 \times 8$ positions,

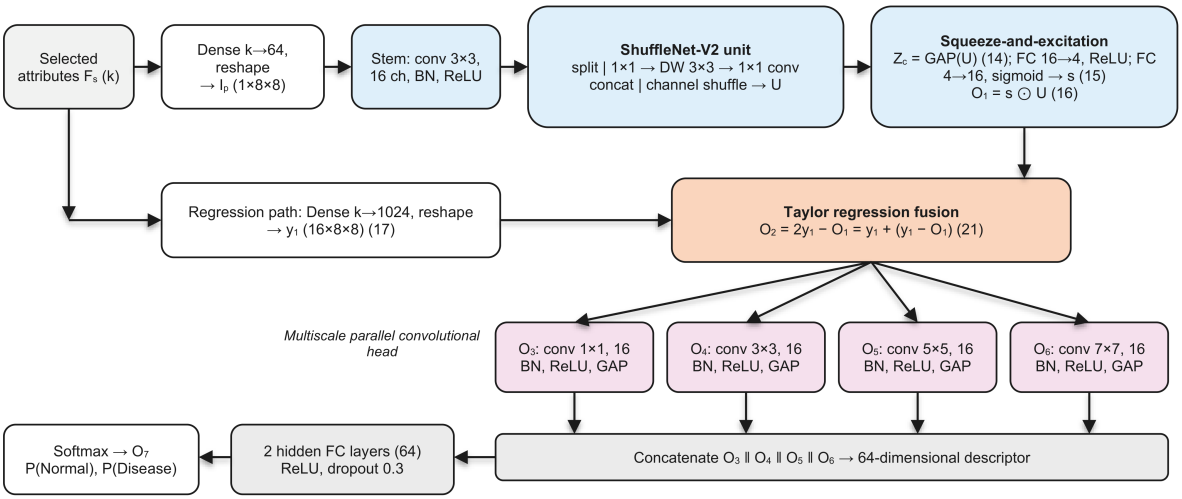


Fig. 2. Network architecture of ShPCN-Net for k selected attributes. The learned input map $\boldsymbol{I_p}$ passes through a stem convolution, a ShuffleNet-V2 unit, and squeeze-and-excitation attention to give $\boldsymbol{O_1}$; a regression of the same attributes gives $\boldsymbol{y_1}$; Taylor regression fusion forms $\boldsymbol{O_2 = 2y_1 - O_1}$; four parallel convolution branches with $\mathbf{1 \times 1}$ to $\mathbf{7 \times 7}$ kernels give $\boldsymbol{O_3}$ to $\boldsymbol{O_6}$, whose pooled outputs are concatenated and classified.

$$I_p = \mathrm{reshape}_{1\times H\times W}\left(\mathbf{W}_e F_s + \mathbf{b}_e\right) \tag{13}$$

where $\mathbf{W}_e \in \mathbb{R}^{HW\times k}$ and $\mathbf{b}_e \in \mathbb{R}^{HW}$ are learned and I_p is the single-channel input map. Methods such as DeepInsight place tabular features on an image grid by a fixed similarity layout [22]; the learned projection used here lets the training objective decide which combinations of attributes appear at neighboring positions, at a cost of only $65k$ parameters. An 8×8 map is the smallest square that holds all 13 attributes with room for learned combinations while keeping the convolutions inexpensive.

2) ShuffleNet-V2 Attention Branch: A 3×3 stem convolution with batch normalization [32] and a rectified linear unit expands I_p to $C = 16$ channels. A ShuffleNet-V2 basic unit [18] then splits the channels into two halves, passes one half through a 1×1 convolution, a 3×3 depthwise convolution, and a second 1×1 convolution, concatenates the result with the untouched half, and applies a channel shuffle so that information flows between the halves in later layers [17]. The output map $U \in \mathbb{R}^{C\times H\times W}$ is recalibrated with squeeze-and-excitation attention [19], as in MissenseNet [20]. The squeeze step averages each channel,

$$Z_c = \frac{1}{H \cdot W}\sum_{i=1}^{H}\sum_{j=1}^{W} U_c(i,j) \tag{14}$$

where Z_c is the descriptor of channel c. The excitation step computes channel weights,

$$\mathbf{s} = \sigma\left(\mathbf{W}_2\,\delta\left(\mathbf{W}_1\mathbf{Z}\right)\right) \tag{15}$$

where $\mathbf{Z} = [Z_1, \ldots, Z_C]^\top$, $\mathbf{W}_1 \in \mathbb{R}^{(C/r)\times C}$ and $\mathbf{W}_2 \in \mathbb{R}^{C\times(C/r)}$ are learned with reduction ratio $r = 4$, δ is the rectified linear unit, and σ is the logistic sigmoid. The branch output rescales every channel by its weight,

$$O_1 = \mathbf{s} \odot U \tag{16}$$

where the weight s_c multiplies all positions of channel c. ShuffleNet-V2 was selected because its channel split and shuffle give a good accuracy-to-cost ratio on small inputs [18], and squeeze-and-excitation was added because it lets the branch emphasize the channels that respond to discriminative attribute combinations at negligible cost [19].

3) Taylor Regression Fusion: The central idea of ShPCN-Net is a fusion rule that combines a direct linear view of the attributes with the attention features. The regression path computes a weighted sum of the selected attributes for every position and channel of the feature map,

$$y_1 = \mathrm{reshape}_{C\times H\times W}\left(\mathbf{W}_r F_s + \mathbf{b}_r\right) \tag{17}$$

where $\mathbf{W}_r \in \mathbb{R}^{CHW\times k}$ and $\mathbf{b}_r$ are learned. Each element of y_1 is a regression $\sum_{q=1}^{k} F_{s,q}\, w_q$ of the attributes, as in the concept design, but with its own weights, so that y_1 has the same shape as O_1. To fuse the two maps, the representation is viewed as a state $x(t)$ evolving over abstract steps, and a first-order Taylor expansion predicts the next state,

$$x(t+1) = x(t) + \frac{x'(t)}{1!} \tag{18}$$

where the derivative is approximated by the backward difference

$$x'(t) \approx \frac{x(t) - x(t-\kappa)}{\kappa} \tag{19}$$

with step κ. Setting $\kappa = 1$ and substituting (19) into (18) gives the linear extrapolation

$$x(t+1) = 2x(t) - x(t-1) \tag{20}$$

Taylor regression fusion assigns the attention output to the earlier state, $x(t-1) = O_1$, the regression output to the current state, $x(t) = y_1$, and the fused map to the predicted state, $x(t+1) = O_2$, which yields

$$O_2 = 2y_1 - O_1 = y_1 + (y_1 - O_1) \tag{21}$$

The second form of (21) shows the logic of the rule: the fused map starts from the regression view and adds the discrepancy between the regression and attention views, extrapolating along the direction from the convolutional representation to the linear one. Because the coefficients 2 and -1 are fixed, the fusion adds no parameters, and both paths receive gradients during training, with $\partial O_2/\partial y_1 = 2$ and $\partial O_2/\partial O_1 = -1$. The rule was preferred over plain addition or concatenation in the concept design because it encodes an explicit relation between the two views and keeps the channel count unchanged. These properties are design rationale; whether the rule improves detection is tested by the ablation experiments in Section IV.

4) Multiscale Parallel Convolutional Head: The fused map is processed by four parallel convolution branches that differ only in kernel size, following the multiscale parallel network of Li *et al.* [21]:

$$O_{2+b} = \delta\left(\mathrm{BN}\left(\mathrm{conv}_{\kappa_b}(O_2)\right)\right), \quad b = 1, \ldots, 4 \tag{22}$$

where $\kappa_b \in \{1, 3, 5, 7\}$ is the kernel size of branch b, each branch has 16 filters with same padding, and BN denotes batch normalization. On an 8×8 map the four kernels cover receptive fields from a single position to almost the whole map, so the head can combine local and global interactions among the learned attribute combinations. Each branch output is reduced by global average pooling, and the four 16-dimensional vectors are concatenated into a 64-dimensional descriptor $\mathbf{h}_0 = [\bar{O}_3 \,\|\, \bar{O}_4 \,\|\, \bar{O}_5 \,\|\, \bar{O}_6]$, where $\|$ denotes concatenation. A stack of L hidden layers with 64 neurons, rectified linear units, and dropout with rate 0.3 [33] follows, and the output layer produces the class probabilities,

$$O_7 = \mathrm{softmax}\left(\mathbf{W}_o\,\mathbf{h}_L + \mathbf{b}_o\right) \tag{23}$$

where $\mathbf{h}_L$ is the output of the last hidden layer and $O_7 = [\hat{P}(\text{Normal}), \hat{P}(\text{Disease})]$. A record is labeled Disease when $\hat{P}(\text{Disease}) \geq 0.5$. For the consensus subset of $k = 7$ attributes, ShPCN-Net has 39 358 trainable parameters.

5) Training Objective and Optimization: ShPCN-Net is trained end to end by minimizing the cross-entropy loss over a mini-batch B,

$$\mathcal{L} = -\frac{1}{|B|}\sum_{j\in B}\sum_{c\in\{0,1\}} \mathbb{1}[c_j = c]\,\log \hat{P}_j(c) \tag{24}$$

where $\mathbb{1}[\cdot]$ is the indicator function and $\hat{P}_j(c)$ is the predicted probability of class c for record j. The parameters are updated by the Adam optimizer [34] with a learning rate of 10^{-3}, a weight decay of 10^{-4}, and mini-batches of 16 records. Within each training fold, a stratified 15% of the records is held out as a validation set; training stops when the validation loss has not improved for 20 epochs or after 150 epochs, and the weights of the epoch with the lowest validation loss are restored. Early stopping and dropout limit overfitting, which is a central risk when a network with tens of thousands of parameters is trained on fewer than 300 records.

6) Computational Footprint: The parameter budget of ShPCN-Net is dominated by its head. For $k = 7$ attributes, the learned input map uses 512 parameters, the stem convolution with batch normalization 176, the ShuffleNet-V2 unit 248, and the squeeze-and-excitation block 148. The regression path of Taylor regression fusion uses 8 192 parameters, because it assigns a separate weight vector to each of the $C \cdot H \cdot W = 1\,024$ positions of y_1. The four parallel branches use 21 632 parameters (about 55% of the total), the two hidden layers 8 320, and the output layer 130, which adds up to the 39 358 parameters stated above. The parallel branches also dominate the computation: on an 8×8 map with 16 input and 16 output channels, the four kernels require $64{\cdot}16{\cdot}16{\cdot}(1{+}9{+}25{+}49) \approx 1.38\times10^6$ multiply-accumulate operations per record, whereas the attention branch requires fewer than 3×10^4. ShPCN-Net therefore remains small enough to train on a laptop processor within seconds per fold, but it carries about six times the parameters of the ShuffleNet-only ablation, a difference that matters when only 270 records are available for training in each fold.

F. Ablation and Comparison Models

Five models were implemented to test the contribution of the ShPCN-Net components and to position it against published model families. Table I summarizes their architectures. Two ablation networks remove parts of ShPCN-Net: the ShuffleNet-only network keeps the learned input map and the attention branch but replaces the fusion and the parallel head by global average pooling and the hidden stack, and the PCNN-only network keeps the learned input map and a stem convolution but removes the attention branch and the fusion. Three comparison models represent the families studied in the literature survey: HCBiLSTM [11], built from a one-dimensional convolution and a bidirectional long short-term memory layer [35], [36]; a dense autoencoder [12], [37] whose encoder is pretrained by reconstruction and then fine-tuned with a densely connected classifier [38]; and XGBoost [7], [8]. The source papers of the comparison families were not available in full, so the models are re-implementations with documented settings rather than reproductions of the published pipelines. All models use the same folds, the same preprocessing, and the same GLO subset in each fold, so that differences in their scores reflect the classifiers alone.

G. Evaluation Protocol and Explainability

1) Leakage-Controlled Cross-Validation: The records are divided into ten stratified folds with a fixed random seed of 42 [39]. For each fold, the imputation and normalization statistics of (1)–(4) are fitted on the nine training folds, the GLO searches for a mask on the same nine folds only, every model is trained on the selected training attributes, and the held-out fold is predicted. The out-of-fold predictions of all ten folds cover every record exactly once and are pooled for the receiver operating characteristic curves and the confusion matrix, whereas the metrics are averaged over folds. Confining the search to the training folds prevents the selection bias that arises when attributes are chosen with knowledge of the test records [5], [6]. Algorithm 2 summarizes the protocol.

TABLE I
ARCHITECTURES OF THE PROPOSED, ABLATION, AND COMPARISON MODELS (PARAMETERS COUNTED FOR $k = 7$ ATTRIBUTES)

Model	Role	Architecture	Parameters
ShPCN-Net	Proposed	Input map 8×8; stem conv; ShuffleNet-V2 unit + SE ($C = 16$, $r = 4$); Taylor regression fusion; four parallel conv branches (1, 3, 5, 7; 16 filters); 2 hidden layers of 64; softmax	39 358
ShuffleNet	Ablation	Input map; stem conv; ShuffleNet-V2 unit + SE; global average pooling; 2 hidden layers of 64; softmax	6 462
PCNN	Ablation	Input map; stem conv; four parallel conv branches; concatenation; 2 hidden layers of 64; softmax	30 770
HCBiLSTM	Comparison	Conv1D (32, kernel 3) + max pooling; BiLSTM (32 per direction); dense 32; dropout 0.3; softmax	19 170
Dense autoencoder	Comparison	Encoder k–32–16, decoder 16–32–k (MSE pretraining, 60 epochs); dense block (3 layers, growth 16); softmax	3 273
XGBoost	Comparison	200 trees, depth 4, learning rate 0.05, subsample 0.8, column subsample 0.8	tree model

Algorithm 2 Leakage-controlled training and evaluation of ShPCN-Net and the comparison models.

Require: dataset D, models $\mathcal{M}$, number of folds $F = 10$, seed 42
Ensure: per-fold metrics, out-of-fold probabilities, masks, and training curves
remove duplicate records; split D into F stratified folds
for $f = 1$ to F **do**
  $\mathcal{T} \leftarrow$ training indices, $\mathcal{V} \leftarrow$ held-out indices of fold f
  fit (1)–(4) on $\mathcal{T}$; transform $\mathcal{T}$ and $\mathcal{V}$ into N_d
  $\mathbf{m}_f^* \leftarrow$ Algorithm 1 on $N_d^{\mathcal{T}}$
  for each model in $\mathcal{M}$ **do**
    train on $N_d^{\mathcal{T}}[:, \mathbf{m}_f^*]$ with early stopping on a stratified 15% validation split
    predict $\hat{P}$(Disease) for $\mathcal{V}$; compute the metrics of Section IV-D
  end for
end for
consensus mask $\leftarrow$ attributes selected in at least half of the folds
train the consensus ShPCN-Net on all records for explanation and sensitivity analysis

2) Consensus Subset and Explanations: Because each fold may select a different subset, the stability of the search is summarized by the selection frequency of each attribute,

$$\nu_q = \frac{1}{F} \sum_{f=1}^{F} m_{f,q}^* \tag{25}$$

where $m_{f,q}^*$ indicates whether attribute q was selected in fold f and $F = 10$. A consensus subset is formed from the attributes with $\nu_q \geq 0.5$, and a consensus ShPCN-Net is trained on all records with this subset. The frequency ν_q also ranks the attributes for the scalability analysis described below. The consensus model is used only for explanation and for the sensitivity analysis, never for the reported cross-validated performance. Explanations use SHAP [27], which attributes a prediction to the attributes through an additive decomposition,

$$\hat{P}_j(\text{Disease}) = \phi_0 + \sum_{q=1}^{k} \phi_{jq} \tag{26}$$

where ϕ_0 is the expected prediction over a background set and ϕ_{jq} is the Shapley value of attribute q for record j. The model-agnostic kernel estimator is applied with a background of 30 k-means centroids, 80 randomly selected records, and 200 coalition samples per record. The SHAP values describe the behavior of the consensus model on records it was trained on and therefore explain the model rather than estimate its generalization. Two exploratory views complete the analysis: a t-SNE embedding [40] of the normalized records and kernel density estimates [41] of key raw attributes by diagnosis.

3) Robustness Analyses: Three analyses probe the stability of the conclusions. The scalability analysis ranks the attributes by their GLO selection frequency and trains ShPCN-Net and the three comparison models on the top 4, 6, 8, 10, and 13 attributes with stratified five-fold cross-validation, which shows how performance changes with the feature dimension. The sensitivity analysis trains ShPCN-Net with 2, 4, 6, and 8 hidden layers on 60%, 70%, 80%, and 90% of the records as learning data, testing on the remainder with three repeated stratified splits. Finally, the paired fold scores of ShPCN-Net and each other model are compared with the two-sided Wilcoxon signed-rank test [25], which is recommended for comparing classifiers because it does not assume normally distributed differences [26].

4) Reproducibility Controls: All random elements, namely the fold assignment, the GLO initialization and moves, the network initialization, the mini-batch order, and the validation split, are seeded from a single run seed of 42, with the fold index added for fold-specific components. Each GLO result is cached together with a key that encodes the seed, the fold, and every optimizer and preprocessing setting, so a rerun either reproduces the stored subset or recomputes it when a setting changes. The fitted preprocessing statistics, the selected masks, the trained fold models, the out-of-fold predictions, and the training curves are stored for every fold, and all figures in Section IV are drawn from these stored results. Redrawing the complete figure set from the stored results reproduced byte-identical underlying data files, which confirms that the reported values are deterministic for the stated seed. The framework

was implemented in Python with PyTorch [42] for the networks, scikit-learn [43] for the support vector wrapper, cross-validation, and metrics, and the reference XGBoost library [7]. Section IV-A lists the exact versions.

IV. Experimental Procedure

A. Experimental Setup

All experiments ran on a laptop with a 13th-generation Intel Core i5-1334U processor (12 hardware threads), 15.7 GB of memory, and Windows 11 Home (build 26200). The integrated Intel Iris Xe graphics adapter was not used; PyTorch executed on the processor with four computation threads. The software environment comprised Python 3.10.11, PyTorch 2.5.1 [42], scikit-learn 1.6.1 [43], XGBoost 2.1.3 [7], SHAP 0.49.1 [27], NumPy 1.26.4, pandas 2.3.3, SciPy 1.15.3, and matplotlib 3.8.4, installed in a dedicated virtual environment. The complete pipeline, comprising preprocessing, the ten-fold comparison of six models with fold-internal GLO search, the consensus model, the scalability and sensitivity analyses, SHAP, the statistical tests, and all figures, finished in 16.7 min. The ten-fold comparison took 327 s, the scalability analysis 290 s, and the sensitivity analysis 340 s.

B. Dataset Description and Partitioning

The experiments use the Cleveland heart disease records of the University of California Irvine repository [3], collected from patients referred for coronary angiography at the Cleveland Clinic Foundation [2]. This standard benchmark is the dataset used by most of the Cleveland studies in the literature survey [8], [11], [12]. The file analyzed here contains 303 records, 13 attributes, and a binary label that marks the presence (Disease, 139 records, 45.9%) or absence (Normal, 164 records, 54.1%) of angiographic disease. It is a cleaned version of the repository data: it contains no missing values and no duplicate records, so (1) and the duplicate removal step leave it unchanged. Table II lists the attributes, their types, and their units; clinical units are given with SI equivalents where applicable. Stratified ten-fold cross-validation assigns 272 or 273 records to training and 30 or 31 records to testing in each fold, and the networks hold out a stratified 15% of each training part (about 41 records) for early stopping. Every record is therefore tested exactly once, by models that never saw it during preprocessing, attribute selection, or training.

TABLE II
Attributes of the Cleveland Heart Disease Records (Training Means over All 303 Records)

Attribute	Type	Description and unit
age	numerical	age in years (mean 54.4)
sex	binary	0 = female, 1 = male
cp	categorical	chest pain type, four codes
trestbps	numerical	resting blood pressure in mm Hg (mean 131.7 mm Hg = 17.6 kPa)
chol	numerical	serum cholesterol in mg/dL (mean 246.7 mg/dL = 6.38 mmol/L)
fbs	binary	fasting blood sugar above 120 mg/dL (6.7 mmol/L)
restecg	categorical	resting electrocardiogram result, three codes
thalach	numerical	maximum heart rate in min^{-1} (mean 149.6)
exang	binary	exercise-induced angina
oldpeak	numerical	exercise ST depression in mm (mean 1.04 mm; 1 mm = 0.1 mV at standard gain)
slope	categorical	slope of the peak exercise ST segment, three codes
ca	categorical	major vessels colored by fluoroscopy, 0 to 3
thal	categorical	thalassemia stress test result, coded

TABLE III
Parameter Settings of the Reported Run

Component	Parameter	Value
GLO	population size P; groups G	20; 4 (5 langurs each)
GLO	iterations T; control $a_0 \to a_1$	50; $2 \to 0$
GLO	bounds $[l, u]$; threshold τ	$[0, 1]$; 0.5
GLO	fitness weight α; territory probability ρ	0.99; 0.10
GLO	migration interval M; alpha death limit L	5; 6 iterations
GLO	wrapper	RBF-SVM, $C = 1$, inner stratified 5-fold
Networks	input map; channels C; SE ratio r	8×8; 16; 4
Networks	hidden layers; neurons; activation	2; 64; ReLU
Networks	loss; optimizer; learning rate	cross-entropy; Adam; 10^{-3}
Networks	batch size; maximum epochs; patience	16; 150; 20
Networks	dropout; weight decay; validation split	0.3; 10^{-4}; 15%
Protocol	folds; seed; data used	10 stratified; 42; 100% (303 records)

C. Parameter Settings

Table III reports the settings confirmed for the reported run. The parameter table of the concept design listed the parameter names without values, so the values were chosen before the experiments to suit a processor-only laptop and were not tuned on the test folds. Early stopping ended training well before the 150-epoch limit: the mean number of executed epochs was 38.7 for ShPCN-Net, 46.1 for the ShuffleNet ablation, 39.4 for the PCNN ablation, 49.9 for HCBiLSTM, and 67.1 for the dense autoencoder.

D. Evaluation Metrics

With TP, TN, FP, and FN denoting the true positives, true negatives, false positives, and false negatives, Disease being the positive class, the metrics are

$$\mathrm{Acc} = \frac{TP + TN}{TP + TN + FP + FN}, \qquad \mathrm{TPR} = \frac{TP}{TP + FN} \tag{27}$$

$$\text{Spec} = \frac{TN}{TN+FP}, \qquad \text{Prec} = \frac{TP}{TP+FP}, \qquad F_1 = \frac{2\,\text{Prec}\cdot\text{TPR}}{\text{Prec}+\text{TPR}} \tag{28}$$

$$\text{MCC} = \frac{TP\cdot TN - FP\cdot FN}{\sqrt{(TP+FP)(TP+FN)(TN+FP)(TN+FN)}} \tag{29}$$

where Acc is the accuracy, TPR the true positive rate (sensitivity or recall), Spec the specificity, Prec the precision, and F_1 the F1-score. The MCC ranges from -1 to 1 and remains informative when the classes are unbalanced or when a model favors one class [23], [24]. The area under the receiver operating characteristic curve (AUC) is the probability that a randomly chosen Disease record receives a higher $\hat{P}(\text{Disease})$ than a randomly chosen Normal record. All metrics are dimensionless ratios. Each metric is computed per fold and reported as the mean and the sample standard deviation over the ten folds; the standard deviation is the deviation measure shown in the fold-wise plots. Accuracy, MCC, and TPR are the primary metrics of the concept design, and precision, specificity, F1-score, and AUC are reported as secondary metrics.

E. Representative Record Through the Pipeline

F. Data Structure and Attribute Selection Results

G. Classification Performance and Comparative Analysis

Table IV and Fig. 7 report the cross-validated performance of all models. All six models reached accuracies between 0.785 and 0.815. The ShuffleNet ablation had the highest mean accuracy (0.815) and MCC (0.634), the dense autoencoder the highest TPR (0.806), and the PCNN ablation the highest AUC (0.902). ShPCN-Net reached an accuracy of 0.805, an MCC of 0.610, a TPR of 0.755, and an AUC of 0.897, ranking fourth in accuracy and third in AUC; it had the highest specificity (0.847), the second-highest precision (0.814), and one of the lowest TPRs. XGBoost scored lowest on accuracy, MCC, and AUC. The fold-to-fold standard deviations, between 0.048 and 0.077 for accuracy, are two to three times larger than the largest difference between the mean accuracies (0.029).

H. Fold-Wise Deviation, Ablation, Scalability, and Sensitivity

I. Explainability Results

J. Statistical Findings and Experimental Discussion

Table V and Fig. 16 report the two-sided Wilcoxon signed-rank tests between ShPCN-Net and every other model on the ten paired fold scores. None of the 20 comparisons was significant at the 0.05 level. The smallest p-values were 0.055 for the MCC against the ShuffleNet ablation, where ShPCN-Net was lower (median difference -0.010), and 0.094 for the TPR against the dense autoencoder (median difference -0.071). Against HCBiLSTM, XGBoost, and the PCNN ablation, all p-values exceeded 0.33.

Taken together, the experiments show a consistent pattern. Under the leakage-controlled protocol, all six models reach accuracies near 0.80 and AUCs near 0.89, and the differences between them are smaller than the variation between folds. The proposed hybrid trades sensitivity for specificity relative to its ablations: it produces the fewest false alarms but misses more Disease records. The Taylor regression fusion and the larger parallel head did not yield a measurable gain, which is plausible given that ShPCN-Net has about six times the parameters of the ShuffleNet ablation and is trained on about 231 records per fold after the validation split. The GLO produced stable and clinically plausible subsets; the four most frequent attributes alone matched the accuracy of the full attribute set. The main experimental limitations are the single dataset, the statistical power of ten paired folds, the fixed hyperparameters, and the symmetric normalization, which may discard useful directional information.

V. Discussion

The central finding of this study is that the proposed ShPCN-Net performs on a par with, but not better than, simpler networks and published model families when every data-dependent step is confined to the training folds. Table VI organizes the results by evaluation method. Under ten-fold cross-validation, ShPCN-Net reached an accuracy of 0.805, compared with 0.815 for its best ablation. In the scalability folds with four attributes, it reached 0.835, compared with 0.838 for the best comparison model. Its best sensitivity configuration reached 0.847 on repeated holdout splits, and none of the 20 paired Wilcoxon tests was significant. The consistent observation across evaluation methods is parity, not superiority.

Table VII places these values beside the results reported by the surveyed studies. The comparison is not like for like. The cited values come from the respective publications and were not reproduced here; several studies used other datasets or modalities, and their validation protocols were not available in sufficient detail. The Cleveland accuracies of 0.967 for HC-BiLSTM [11] and 0.997 for the dense autoencoder [12] exceed the 0.798 and 0.812 obtained here by re-implementations of the same model families. One plausible explanation, which this study cannot confirm without the original pipelines, is the protocol: selecting attributes or fitting preprocessing on all records before validation is known to inflate estimates on small datasets [5], [6]. Differences in architecture details, tuning, and dataset versions may also contribute. The present estimates should therefore be read as conservative reference values for leakage-controlled evaluation.

Several interpretations, which remain hypotheses, may explain why the hybrid did not improve detection. First, model capacity is high relative to the data: ShPCN-Net has 39 358 parameters for about 231 training records per fold, six times the ShuffleNet ablation, and its training curves show overfitting after about ten epochs. Second, the fixed coefficients of Taylor regression fusion impose a particular combination of the regression and attention views; a learned gate might suit small data better. Third, the symmetric Gaussian normalization removes the direction of deviation, which may explain why maximum heart rate and ST depression were rarely selected

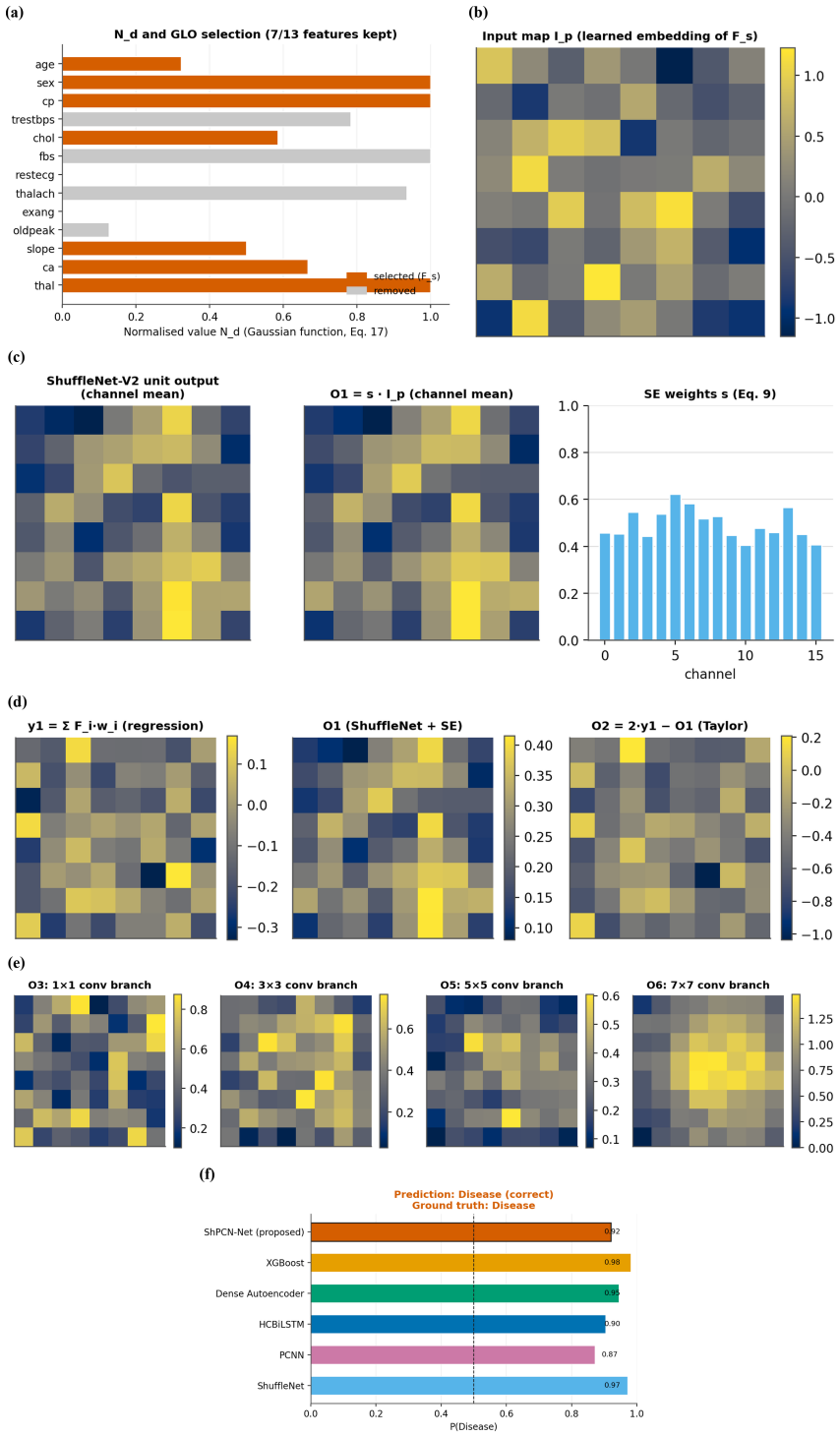


Fig. 3. Representative Disease record (record 299, fold 9) traced through ShPCN-Net. (a) Normalized attributes and the GLO mask (orange: selected). (b) Learned input map $\boldsymbol{I_p}$. (c) ShuffleNet-V2 unit output, attention output $\boldsymbol{O_1}$, and squeeze-and-excitation weights. (d) Regression map $\boldsymbol{y_1}$, $\boldsymbol{O_1}$, and fused map $\boldsymbol{O_2 = 2y_1 - O_1}$. (e) Outputs $\boldsymbol{O_3}$ to $\boldsymbol{O_6}$ of the four parallel branches. (f) Out-of-fold Disease probabilities of all models against the ground truth. Equation labels inside the panels follow the implementation's specification numbering; they correspond to (4) and (15) in this paper.

despite their clear class differences. Benchmarks of tabular learning likewise show that deep architectures rarely dominate simpler models on small tabular datasets [44].

The study also has clear strengths. The GLO consistently selected the vessel count, the thalassemia test, the chest pain type, and exercise-induced angina, and four attributes sufficed to match the accuracy of all 13. The SHAP analysis assigned the largest contributions to the same attributes in clinically expected directions [27], so the selection and the explanations corroborate each other. ShPCN-Net also achieved the highest specificity, 0.847, which may suit screening settings where false alarms are costly. Its lower TPR is a limitation for such

TABLE IV
TEN-FOLD CROSS-VALIDATED PERFORMANCE (MEAN ± STANDARD DEVIATION OVER FOLDS; BEST MEAN IN BOLD)

Model	Accuracy	MCC	TPR	Precision	Specificity	F1-score	AUC
ShuffleNet	**0.815** ± 0.064	**0.634** ± 0.132	0.784 ± 0.101	**0.817** ± 0.097	0.841 ± 0.099	0.795 ± 0.067	0.899 ± 0.056
PCNN	0.808 ± 0.048	0.623 ± 0.094	0.792 ± 0.091	0.803 ± 0.089	0.824 ± 0.102	0.791 ± 0.048	**0.902** ± 0.050
HCBiLSTM	0.798 ± 0.051	0.600 ± 0.107	0.776 ± 0.065	0.794 ± 0.100	0.817 ± 0.101	0.780 ± 0.050	0.890 ± 0.049
Dense autoencoder	0.812 ± 0.074	0.626 ± 0.150	**0.806** ± 0.101	0.794 ± 0.093	0.817 ± 0.091	**0.797** ± 0.079	0.894 ± 0.054
XGBoost	0.785 ± 0.077	0.575 ± 0.157	0.749 ± 0.122	0.785 ± 0.112	0.818 ± 0.104	0.760 ± 0.093	0.885 ± 0.054
ShPCN-Net (proposed)	0.805 ± 0.070	0.610 ± 0.143	0.755 ± 0.085	0.814 ± 0.098	**0.847** ± 0.090	0.781 ± 0.074	0.897 ± 0.061

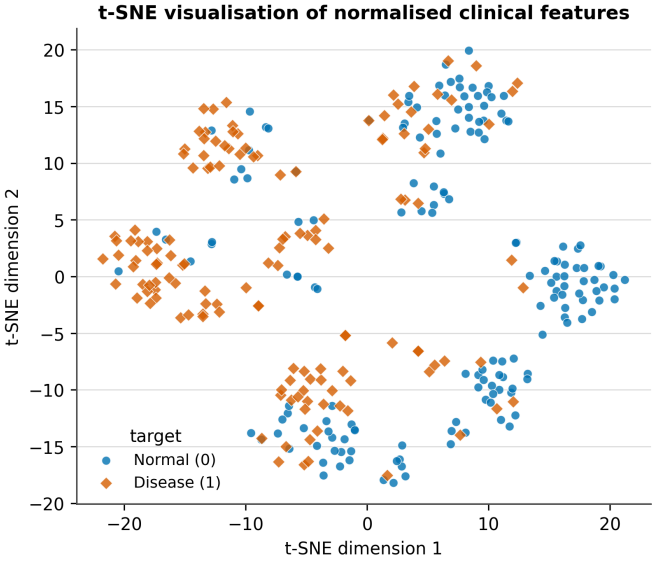


Fig. 4. t-SNE embedding of the 303 normalized Cleveland records colored by diagnosis.

TABLE V
TWO-SIDED WILCOXON SIGNED-RANK P-VALUES, SHPCN-NET VERSUS EACH MODEL OVER TEN PAIRED FOLDS

Compared model	Accuracy	MCC	TPR	AUC
ShuffleNet	0.250	0.055	0.188	0.539
PCNN	0.703	0.625	0.375	0.461
HCBiLSTM	0.750	0.883	0.531	0.625
Dense autoencoder	0.500	0.275	0.094	0.730
XGBoost	0.594	0.846	0.850	0.338

TABLE VI
SUMMARY OF SHPCN-NET RESULTS BY EVALUATION METHOD ON THE CLEVELAND DATASET

Evaluation method	ShPCN-Net	Best other model
10-fold CV, accuracy (Table IV)	0.805	0.815 (ShuffleNet)
10-fold CV, AUC	0.897	0.902 (PCNN)
5-fold CV, top-4 attributes, accuracy	0.835	0.838 (dense autoencoder)
Holdout, 2 layers, 80% learning data	0.847	not evaluated
Wilcoxon tests with $p < 0.05$	0 of 20	

TABLE VII
REPORTED RESULTS OF SURVEYED STUDIES AND OF THIS STUDY (REPORTED VALUES; DIFFERENT DATASETS AND PROTOCOLS, NOT DIRECTLY COMPARABLE)

Study	Dataset	Model	Reported result
[11]	Cleveland	CNN-BiLSTM, extra-trees selection	accuracy 0.967
[12]	Cleveland	dense autoencoder + DenseNet	mean accuracy 0.997
[8]	Cleveland	XGBoost vs. five classifiers	XGBoost preferred
[9]	risk factors, 70 000 records	bagging, boosting, stacking	+0.0196 accuracy from bagging
[10]	public CVD risk data	stacked boosting and neural ensemble	AUC 0.82
[13]	MITHSDB, ECG and PCG	bimodal late-fusion CNN	AUC 0.964
This study	Cleveland, 303 records	ShPCN-Net, fold-internal GLO	accuracy 0.805, AUC 0.897

settings, however.

Statistical power is the main constraint on these conclusions. With ten paired folds, the Wilcoxon test can detect only large and consistent differences [26], and the observed mean differences of at most 0.03 are smaller than the fold-to-fold standard deviations of 0.05 to 0.08. The absence of significance therefore shows that no advantage was detected, not that the models are equivalent. Repeated cross-validation, external cohorts, tuned hyperparameters, and a direction-preserving normalization are needed before a firmer conclusion can be drawn.

VI. CONCLUSION

This study set out to detect heart disease from routine clinical attributes with ShPCN-Net, which combines Gaussian normalization, Gray Langurs Optimizer attribute selection, and a hybrid of a ShuffleNet-V2 attention branch, Taylor regression fusion, and a multiscale parallel convolutional head. The study also evaluated this hybrid under a leakage-controlled protocol. On all 303 Cleveland records with stratified ten-fold cross-validation, the optimizer reduced 13 attributes to 6.9 per fold and consistently retained the vessel count and the thalassemia test. Four top-ranked attributes matched the accuracy of the full set, and SHAP explanations confirmed the clinical plausibility of these attributes. ShPCN-Net achieved an accuracy of 0.805,

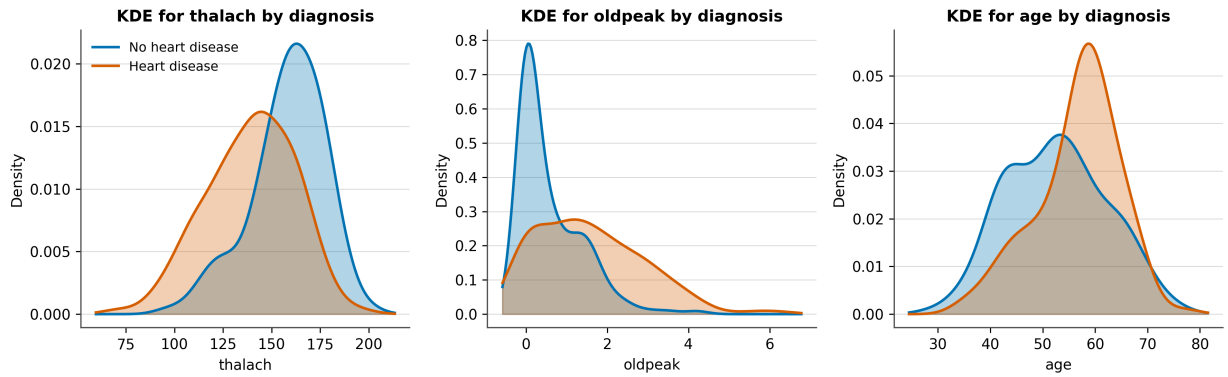


Fig. 5. Kernel density estimates of maximum heart rate (thalach, $\mathbf{min}^{-1}$), ST depression (oldpeak, mm), and age (years) by diagnosis.

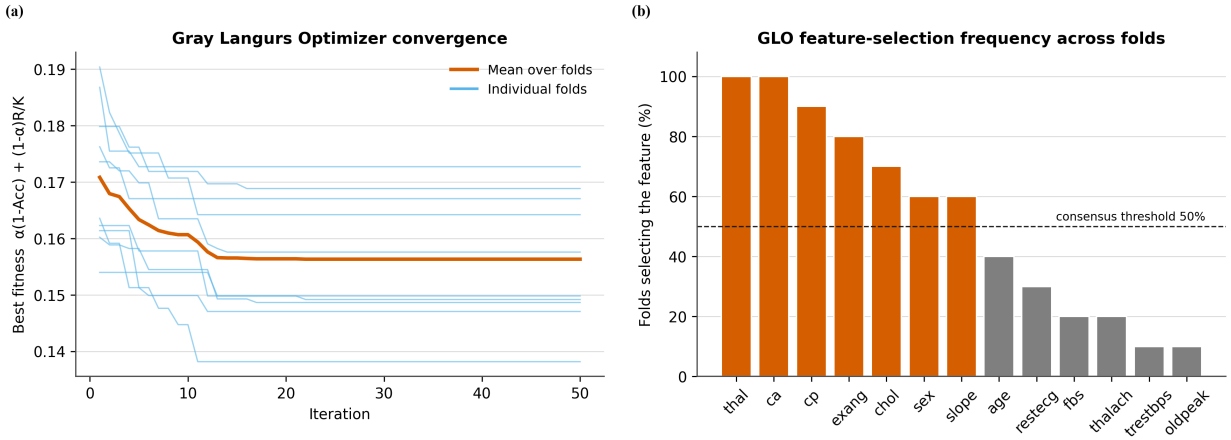


Fig. 6. Gray Langurs Optimizer search. (a) Best fitness per iteration for each fold and the mean over folds. (b) Selection frequency of each attribute over the ten folds; the dashed line marks the 50% consensus threshold.

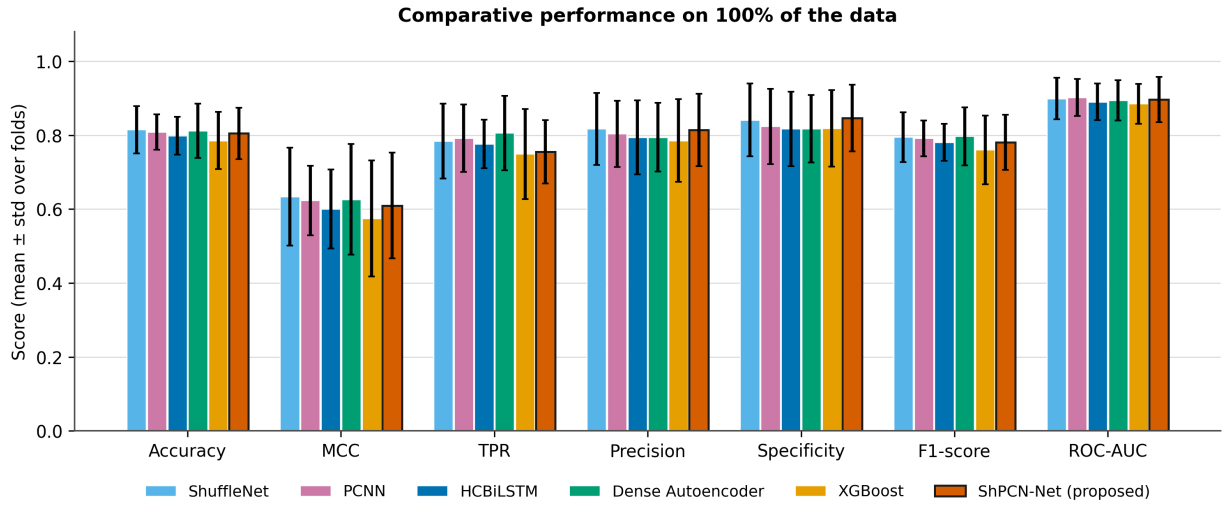


Fig. 7. Comparative performance of all models on 100% of the records (mean ± standard deviation over ten folds).

an MCC of 0.610, a TPR of 0.755, a specificity of 0.847, and an AUC of 0.897. All six evaluated models reached accuracies between 0.785 and 0.815, and no Wilcoxon comparison was significant, so the fusion provided no measurable advantage on this dataset. The main limitations are the single small dataset, untuned hyperparameters, and the symmetric normalization. Future work will evaluate the framework on larger multicenter cohorts with repeated cross-validation, replace the fixed fusion with a learned gate, adopt direction-preserving normalization, and compare the optimizer with established selectors.

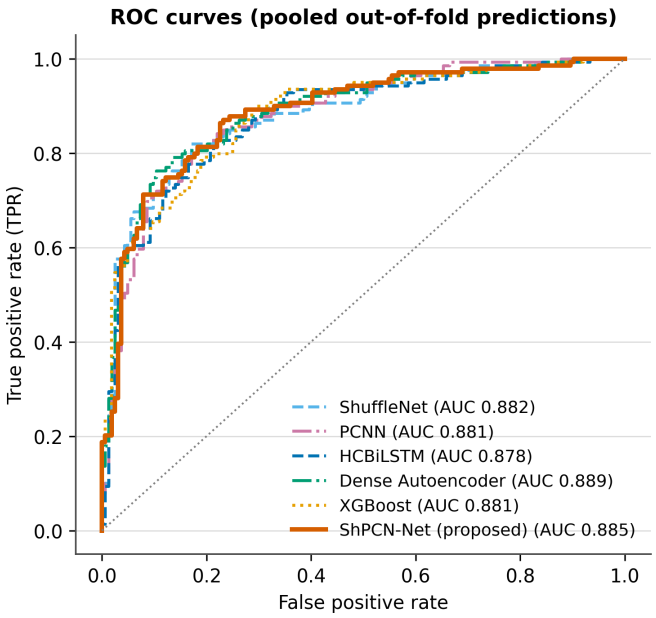


Fig. 8. Receiver operating characteristic curves of the pooled out-of-fold predictions of all models.

Declaration of Generative AI Use

During the preparation of this work, the authors used an AI assistant (Claude, Anthropic) to draft and organize the manuscript text, to write and run the experimental code under the authors' specification, and to check the reported numbers against the stored experiment outputs. All numerical results come from the executed code and its stored outputs, not from the AI assistant. The authors reviewed and edited the content and take full responsibility for the publication. [Authors to confirm or revise this statement before submission.]

Acknowledgment

[Funding and acknowledgment details to be provided by the authors.]

References

[1] World Health Organization, "Cardiovascular diseases (CVDs)," WHO Fact Sheets, Jun. 11, 2021. Accessed: Nov. 7, 2026. [Online]. Available: https://www.who.int/news-room/fact-sheets/detail/cardiovascular-diseases-(cvds)

[2] R. Detrano *et al.*, "International application of a new probability algorithm for the diagnosis of coronary artery disease," *Am. J. Cardiol.*, vol. 64, no. 5, pp. 304–310, Aug. 1989, doi: 10.1016/0002-9149(89)90524-9.

[3] A. Janosi, W. Steinbrunn, M. Pfisterer, and R. Detrano, "Heart disease," UCI Machine Learning Repository, 1988, doi: 10.24432/C52P4X. Accessed: Nov. 7, 2026. [Online]. Available: https://archive.ics.uci.edu/dataset/45/heart+disease

[4] I. Guyon and A. Elisseeff, "An introduction to variable and feature selection," *J. Mach. Learn. Res.*, vol. 3, pp. 1157–1182, Mar. 2003.

[5] C. Ambroise and G. J. McLachlan, "Selection bias in gene extraction on the basis of microarray gene-expression data," *Proc. Natl. Acad. Sci. USA*, vol. 99, no. 10, pp. 6562–6566, May 2002, doi: 10.1073/pnas.102102699.

[6] S. Varma and R. Simon, "Bias in error estimation when using cross-validation for model selection," *BMC Bioinformatics*, vol. 7, 2006, Art. no. 91, doi: 10.1186/1471-2105-7-91.

[7] T. Chen and C. Guestrin, "XGBoost: A scalable tree boosting system," in *Proc. 22nd ACM SIGKDD Int. Conf. Knowl. Discov. Data Mining*, San Francisco, CA, USA, 2016, pp. 785–794, doi: 10.1145/2939672.2939785.

[8] L. A. Alharbi, "Heart disease prediction of Cleveland Clinic patients using advanced machine learning algorithms," *J. Adv. Res. Des.*, vol. 126, no. 1, pp. 1–14, Mar. 2025, doi: 10.37934/ard.126.1.114.

[9] V. Shorewala, "Early detection of coronary heart disease using ensemble techniques," *Inform. Med. Unlocked*, vol. 26, 2021, Art. no. 100655, doi: 10.1016/j.imu.2021.100655.

[10] P. Shah, M. Shukla, N. H. Dholakia, and H. Gupta, "Predicting cardiovascular risk with hybrid ensemble learning and explainable AI," *Sci. Rep.*, vol. 15, no. 1, May 2025, Art. no. 17927, doi: 10.1038/s41598-025-01650-7.

[11] P. K. Shrivastava, M. Sharma, P. Sharma, and A. Kumar, "HCBiLSTM: A hybrid model for predicting heart disease using CNN and BiLSTM algorithms," *Meas.: Sensors*, vol. 25, Feb. 2023, Art. no. 100657, doi: 10.1016/j.measen.2022.100657.

[12] N. S. Alghamdi, M. Zakariah, A. Shankar, and W. Viriyasitavat, "Heart disease prediction using autoencoder and DenseNet architecture," *Egyptian Inform. J.*, vol. 28, Dec. 2024, Art. no. 100559, doi: 10.1016/j.eij.2024.100559.

[13] A. Calzoni, M. Savardi, M. Silvestri, S. Benini, and A. Signoroni, "Bimodal ECG and PCG cardiovascular disease detection: Exploring the potential and modality contribution," *J. Med. Syst.*, vol. 49, no. 1, Sep. 2025, Art. no. 113, doi: 10.1007/s10916-025-02245-5.

[14] Y. Qiu, R. Li, and 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.

[15] R. Kohavi and G. H. John, "Wrappers for feature subset selection," *Artif. Intell.*, vol. 97, no. 1–2, pp. 273–324, Dec. 1997, doi: 10.1016/S0004-3702(97)00043-X.

[16] S. Barshandeh *et al.*, "Gray langurs optimizer: A multi-group bio-inspired optimization algorithm," *Artif. Intell. Rev.*, vol. 59, no. 7, Apr. 2026, Art. no. 166, doi: 10.1007/s10462-026-11529-2.

[17] X. Zhang, X. Zhou, M. Lin, and J. Sun, "ShuffleNet: An extremely efficient convolutional neural network for mobile devices," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Salt Lake City, UT, USA, 2018, pp. 6848–6856, doi: 10.1109/CVPR.2018.00716.

[18] N. Ma, X. Zhang, H.-T. Zheng, and J. Sun, "ShuffleNet V2: Practical guidelines for efficient CNN architecture design," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, Munich, Germany, 2018, pp. 116–131, doi: 10.1007/978-3-030-01264-9_8.

[19] J. Hu, L. Shen, and G. Sun, "Squeeze-and-excitation networks," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Salt Lake City, UT, USA, 2018, pp. 7132–7141, doi: 10.1109/CVPR.2018.00745.

[20] J. Liu, Y. Chen, K. Huang, and X. Guan, "Enhancing missense variant pathogenicity prediction with MissenseNet: Integrating structural insights and ShuffleNet-based deep learning techniques," *Biomolecules*, vol. 14, no. 9, Sep. 2024, Art. no. 1105, doi: 10.3390/biom14091105.

[21] C. Li, M. Liang, W. Song, and K. Xiao, "A multi-scale parallel convolutional neural network based intelligent human identification using face information," *J. Inf. Process. Syst.*, vol. 14, no. 6, pp. 1494–1507, Dec. 2018, doi: 10.3745/JIPS.02.0103.

[22] A. Sharma, E. Vans, D. Shigemizu, K. A. Boroevich, and T. Tsunoda, "DeepInsight: A methodology to transform a non-image data to an image for convolution neural network architecture," *Sci. Rep.*, vol. 9, 2019, Art. no. 11399, doi: 10.1038/s41598-019-47765-6.

[23] B. W. Matthews, "Comparison of the predicted and observed secondary structure of T4 phage lysozyme," *Biochim. Biophys. Acta*, vol. 405, no. 2, pp. 442–451, Oct. 1975, doi: 10.1016/0005-2795(75)90109-9.

[24] D. Chicco and G. Jurman, "The advantages of the Matthews correlation coefficient (MCC) over F1 score and accuracy in binary classification evaluation," *BMC Genomics*, vol. 21, 2020, Art. no. 6, doi: 10.1186/s12864-019-6413-7.

[25] F. Wilcoxon, "Individual comparisons by ranking methods," *Biometrics Bull.*, vol. 1, no. 6, pp. 80–83, Dec. 1945, doi: 10.2307/3001968.

[26] J. Demšar, "Statistical comparisons of classifiers over multiple data sets," *J. Mach. Learn. Res.*, vol. 7, pp. 1–30, 2006.

[27] S. M. Lundberg and S.-I. Lee, "A unified approach to interpreting model predictions," in *Proc. Adv. Neural Inf. Process. Syst. (NeurIPS)*, Long Beach, CA, USA, 2017, pp. 4765–4774.

[28] Y.-S. Kim, M. K. Kim, N. Fu, J. Liu, J. Wang, and J. Srebric, "Investigating the impact of data normalization methods on predicting electricity consumption in a building using different artificial neural network models," *Sustain. Cities Soc.*, vol. 118, Jan. 2025, Art. no.

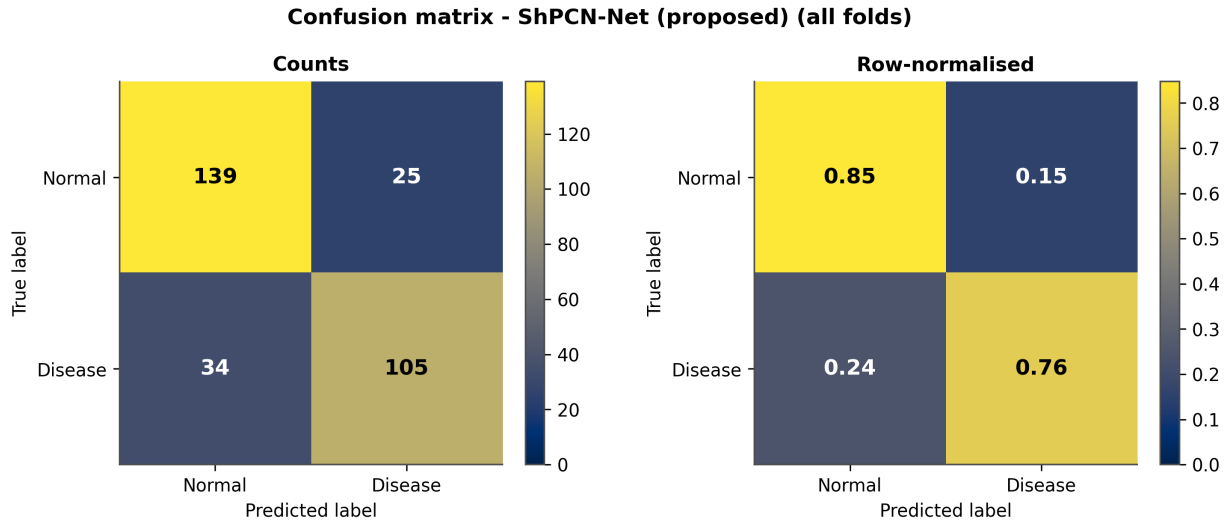


Fig. 9. Pooled out-of-fold confusion matrix of ShPCN-Net: counts (left) and row-normalized rates (right).

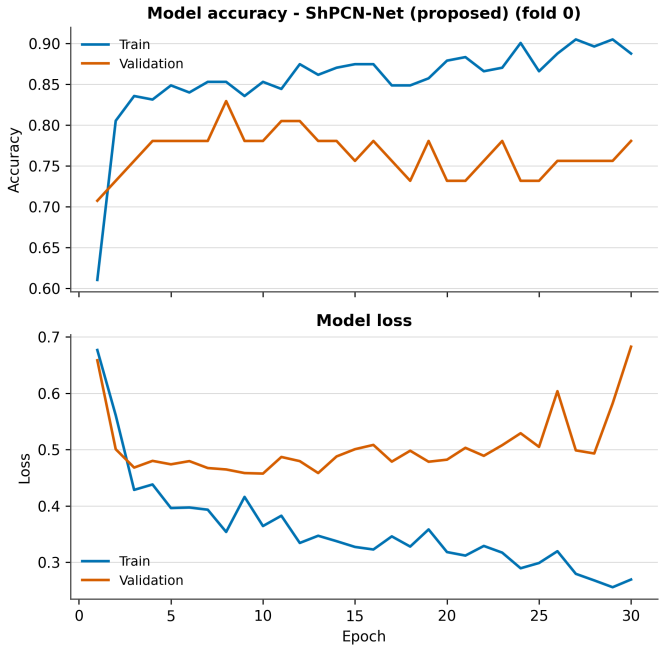


Fig. 10. Training and validation accuracy (top) and loss (bottom) of ShPCN-Net in fold 0.

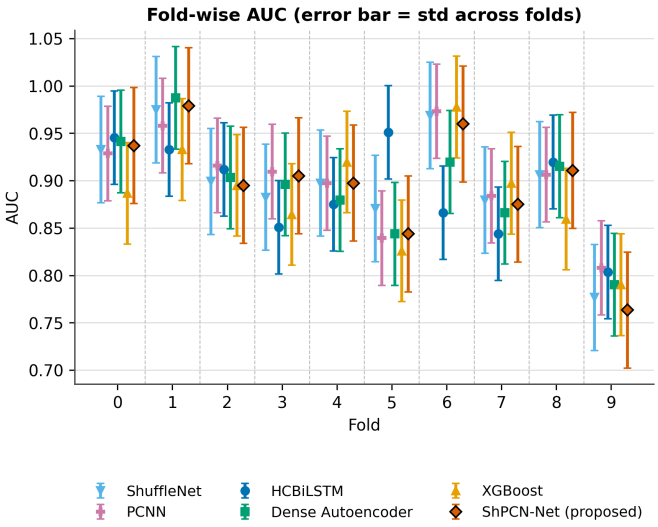


Fig. 11. Fold-wise AUC of all models; error bars show the standard deviation over the ten folds.

105570, doi: 10.1016/j.scs.2024.105570.
[29] D. H. Wolpert and W. G. Macready, "No free lunch theorems for optimization," *IEEE Trans. Evol. Comput.*, vol. 1, no. 1, pp. 67–82, Apr. 1997, doi: 10.1109/4235.585893.
[30] C. Cortes and V. Vapnik, "Support-vector networks," *Mach. Learn.*, vol. 20, no. 3, pp. 273–297, Sep. 1995, doi: 10.1007/BF00994018.
[31] S. Mirjalili, S. M. Mirjalili, and A. Lewis, "Grey wolf optimizer," *Adv. Eng. Softw.*, vol. 69, pp. 46–61, Mar. 2014, doi: 10.1016/j.advengsoft.2013.12.007.
[32] S. Ioffe and C. Szegedy, "Batch normalization: Accelerating deep network training by reducing internal covariate shift," in *Proc. 32nd Int. Conf. Mach. Learn. (ICML)*, Lille, France, 2015, pp. 448–456.
[33] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov, "Dropout: A simple way to prevent neural networks from overfitting," *J. Mach. Learn. Res.*, vol. 15, no. 56, pp. 1929–1958, 2014.
[34] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," in *Proc. 3rd Int. Conf. Learn. Represent. (ICLR)*, San Diego, CA, USA, 2015.
[35] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural Comput.*, vol. 9, no. 8, pp. 1735–1780, Nov. 1997, doi: 10.1162/neco.1997.9.8.1735.
[36] M. Schuster and K. K. Paliwal, "Bidirectional recurrent neural networks," *IEEE Trans. Signal Process.*, vol. 45, no. 11, pp. 2673–2681, Nov. 1997, doi: 10.1109/78.650093.
[37] G. E. Hinton and R. R. Salakhutdinov, "Reducing the dimensionality of data with neural networks," *Science*, vol. 313, no. 5786, pp. 504–507, Jul. 2006, doi: 10.1126/science.1127647.
[38] G. Huang, Z. Liu, L. van der Maaten, and K. Q. Weinberger, "Densely connected convolutional networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Honolulu, HI, USA, 2017, pp. 4700–4708, doi: 10.1109/CVPR.2017.243.
[39] R. Kohavi, "A study of cross-validation and bootstrap for accuracy estimation and model selection," in *Proc. 14th Int. Joint Conf. Artif. Intell. (IJCAI)*, Montreal, QC, Canada, 1995, pp. 1137–1143.
[40] L. van der Maaten and G. Hinton, "Visualizing data using t-SNE," *J. Mach. Learn. Res.*, vol. 9, no. 86, pp. 2579–2605, Nov. 2008.
[41] E. Parzen, "On estimation of a probability density function and mode," *Ann. Math. Statist.*, vol. 33, no. 3, pp. 1065–1076, Sep. 1962, doi: 10.1214/aoms/1177704472.

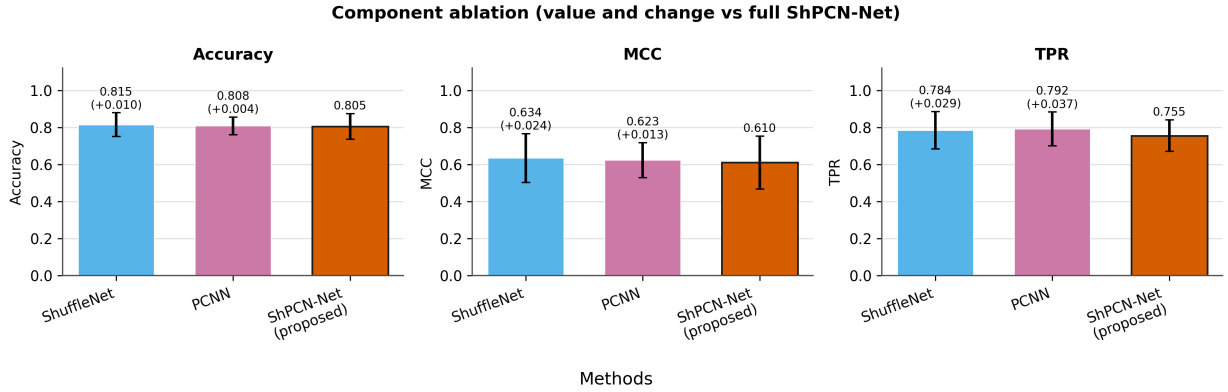


Fig. 12. Ablation analysis: accuracy, MCC, and TPR of the ShuffleNet and PCNN ablations and the full ShPCN-Net; values in parentheses give the difference from the full model.

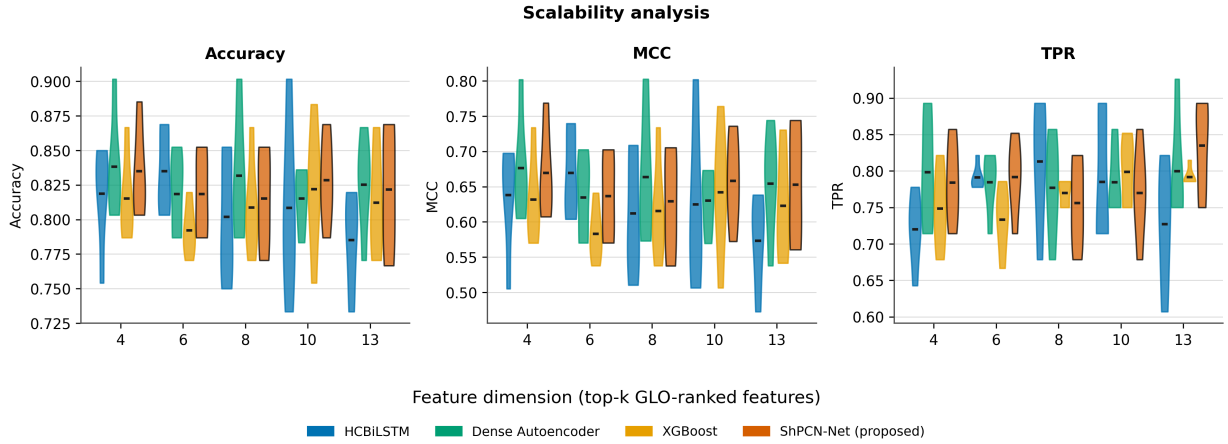


Fig. 13. Scalability analysis: distributions of accuracy, MCC, and TPR over five folds for the top 4, 6, 8, 10, and 13 GLO-ranked attributes.

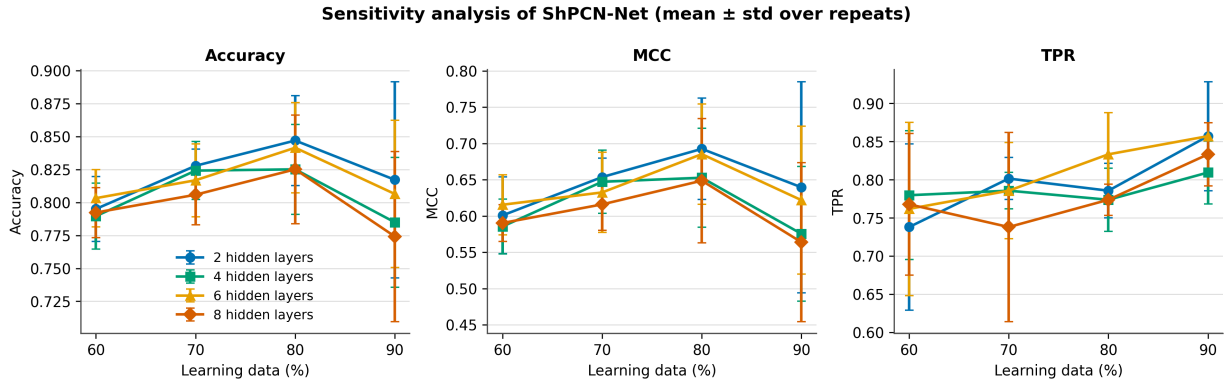


Fig. 14. Sensitivity analysis of ShPCN-Net: mean ± standard deviation over three repeats for 2, 4, 6, and 8 hidden layers against the learning-data percentage.

[42] A. Paszke *et al.*, "PyTorch: An imperative style, high-performance deep learning library," in *Proc. Adv. Neural Inf. Process. Syst. (NeurIPS)*, Vancouver, BC, Canada, 2019, pp. 8024–8035.
[43] F. Pedregosa *et al.*, "Scikit-learn: Machine learning in Python," *J. Mach. Learn. Res.*, vol. 12, pp. 2825–2830, 2011.
[44] L. Grinsztajn, E. Oyallon, and G. Varoquaux, "Why do tree-based models still outperform deep learning on typical tabular data?," in *Proc. Adv. Neural Inf. Process. Syst. (NeurIPS)*, New Orleans, LA, USA, 2022, pp. 507–520.

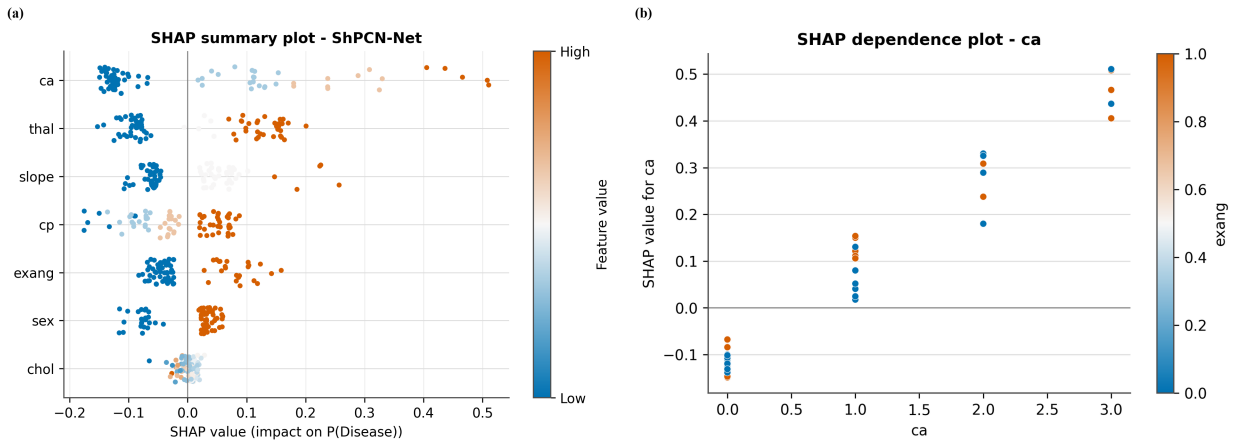


Fig. 15. SHAP explanation of the consensus ShPCN-Net. (a) Summary plot of the Shapley values of the seven consensus attributes (color: attribute value). (b) Dependence plot of the vessel count (ca), colored by exercise-induced angina.

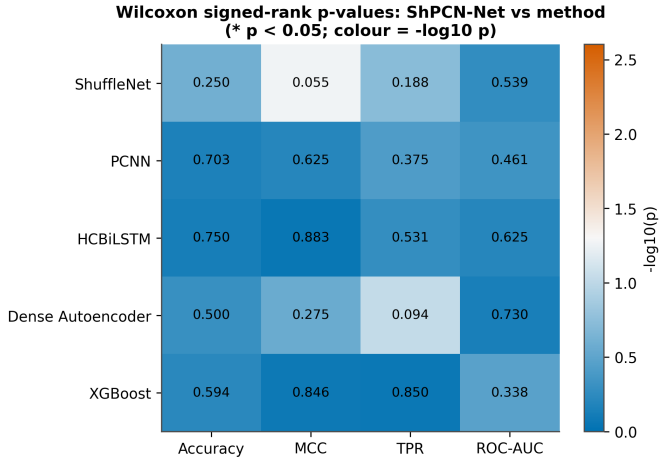


Fig. 16. Wilcoxon signed-rank p-values between ShPCN-Net and each model (color: $-\log_{10} p$; an asterisk would mark $p < 0.05$).