VOL. 12. NO. 1 AUGUST 2026
P-ISSN: 2685-8223 | E-ISSN: 2527-4864
DOI: 10.33480/jitk.v12i1.8299

JITK (JURNAL ILMU PENGETAHUAN DAN TEKNOLOGI KOMPUTER)

# CONTEXTUAL FEATURE NORMALIZATION ON THE PERFORMANCE OF HEART DISEASE CLASSIFICATION MODELS

**Adi Suwondo[1*]; Kusrini[1]; Ema Utami[1]; Kumara Ari Yuana[1]**

Doctoral in Informatics[1]
Amikom University Yogyakarta, Yogyakarta, Indonesia[1]
https:// amikom.ac.id[1]
adisuwondo@students.amikom.ac.id[*]

(*) Corresponding Author
(Responsible for the Quality of Paper Content)



***Abstract***— *Heart disease classification models commonly employ statistical normalization techniques that standardize features according to data distribution but do not explicitly incorporate clinically meaningful cardiovascular information. This study evaluates Clinical Contextual Normalization (CCN) as an alternative feature representation strategy for heart disease classification. Using the Cleveland Heart Disease dataset (303 records), standard numerical representation and CCN were evaluated across five classifiers: Logistic Regression (LR), Support Vector Classifier (SVC), Random Forest (RF), Multilayer Perceptron (MLP), and Naïve Bayes (NB). Model performance was assessed using repeated stratified 10-fold cross-validation with five repetitions (50 evaluation folds), with recall as the primary metric because false negatives may delay clinical screening. The results revealed a classifier-dependent response to CCN. Random Forest showed a small numerical recall increase (ΔRecall = +0.0059), but the difference was not statistically significant (p = 0.6434). MLP produced the largest positive numerical recall change (ΔRecall = +0.0143) and produced 10 more aggregated true-positive predictions while false positives decreased by two, although its recall difference was also not statistically significant (p = 0.2195). In contrast, Logistic Regression showed a statistically significant recall decrease (ΔRecall = −0.0115, p = 0.0186), while Naïve Bayes exhibited the largest significant reduction (ΔRecall = −0.0333, p < 0.001). These findings demonstrate that clinically informed feature representation does not uniformly improve predictive performance but produces classifier-dependent effects. Further validation using larger and more diverse datasets is required before clinical deployment.*

***Keywords***: *Clinical Normalization, Contextual Features, Heart Disease Classification, Recall Optimization, Risk Based Representation.*

**Intisari**— *Model klasifikasi penyakit jantung umumnya menggunakan teknik normalisasi statistik yang menstandarkan fitur berdasarkan distribusi data, tetapi tidak secara eksplisit memasukkan informasi kardiovaskular yang bermakna secara klinis. Penelitian ini mengevaluasi Clinical Contextual Normalization (CCN) sebagai strategi representasi fitur alternatif untuk klasifikasi penyakit jantung. Menggunakan dataset Cleveland Heart Disease yang terdiri atas 303 data pasien, representasi numerik standar dan CCN dievaluasi pada lima algoritma klasifikasi: Logistic Regression (LR), Support Vector Classifier (SVC), Random Forest (RF), Multilayer Perceptron (MLP), dan Naïve Bayes (NB). Kinerja model dievaluasi menggunakan repeated stratified 10-fold cross-validation sebanyak lima pengulangan (50 lipatan evaluasi), dengan recall sebagai metrik utama karena kesalahan negatif palsu (false negative) dapat menyebabkan keterlambatan dalam skrining klinis. Hasil penelitian menunjukkan bahwa respons terhadap CCN bergantung pada jenis pengklasifikasi. Random Forest menunjukkan peningkatan recall secara numerik (ΔRecall = +0,0059), tetapi perbedaannya tidak signifikan secara statistik (p = 0,6434). MLP menghasilkan peningkatan recall numerik terbesar (ΔRecall = +0,0143) dan menghasilkan 10 prediksi true-positive agregat lebih banyak, sementara false-positive berkurang sebanyak dua, meskipun perbedaan recall-nya juga tidak signifikan secara statistik (p = 0,2195). Sebaliknya, Logistic Regression mengalami penurunan recall yang signifikan secara statistik*

VOL. 12. NO. 1 AUGUST 2026
P-ISSN: 2685-8223 | E-ISSN: 2527-4864
DOI: 10.33480/jitk.v12i1.8299

*(ΔRecall = −0,0115; p = 0,0186), sedangkan Naïve Bayes menunjukkan penurunan signifikan terbesar (ΔRecall = −0,0333; p < 0,001). Temuan ini menunjukkan bahwa representasi fitur berbasis informasi klinis tidak meningkatkan kinerja prediktif secara seragam, melainkan menghasilkan pengaruh yang bergantung pada karakteristik pengklasifikasi. Validasi lebih lanjut menggunakan dataset yang lebih besar dan beragam diperlukan sebelum pendekatan ini diterapkan dalam lingkungan klinis.*

***Kata Kunci**: Normalisasi Klinis, Fitur Kontekstual, Klasifikasi Penyakit Jantung, Optimasi Recall, Representasi Berbasis Risiko.*

## INTRODUCTION

Cardiovascular disease remains a major global health burden, and early identification of high risk patients is central to screening-oriented clinical decision support [1]. Machine learning has been widely used to classify heart disease from structured clinical variables, particularly using benchmark datasets such as Cleveland, Framingham, and electronic health record derived datasets [2]. Recent reviews indicate that the field has progressed from single classifiers toward ensemble, deep learning, and explainable AI frameworks; however, clinical translation is still constrained by dataset quality, model transparency, and the way clinical variables are represented before learning [2].

Recent empirical studies have evaluated heterogeneous machine-learning approaches for cardiovascular prediction, including classifier comparison, feature-selection strategies, ensemble learning, and explainable prediction frameworks [3]–[7]. Recent studies have also incorporated feature selection, ensemble learning, and explainability techniques into heart-disease prediction pipelines [6]–[9]. Based on the reviews we referenced, there is a tendency that heart disease prediction studies more often report model performance, feature selection, or XAI rather than separately evaluating the influence of preprocessing strategies and feature representation; therefore, this aspect is more appropriately positioned as an area that is still relatively underexplored [2].

A key methodological gap is therefore located at the feature representation stage. Standard normalization, such as z-score scaling, transforms variables according to statistical distribution but does not encode clinically meaningful boundaries. In cardiovascular screening, variables such as resting blood pressure, total cholesterol, maximum heart rate, and exercise-induced ST depression can be interpreted using clinically meaningful reference points, physiological relationships, or derived representations. Based on the review of existing sources, ML studies tend to emphasize model performance, feature selection, or interpretability, while controlled evaluations of normalization alternatives with the same model and validation are rare; therefore, this research is more appropriately positioned as a follow up effort that specifically highlights the role of feature representation.

This distinction is important because feature representation determines how clinical information is presented to a learning algorithm before model training. Clinically informed feature representation may improve alignment between model inputs and medical reasoning, but it may also alter feature distributions, numerical relationships, or the amount of continuous information available to particular learning algorithms. Consequently, its benefit must be tested empirically across different model families rather than assumed to be universally positive.

This study evaluates the effect of CCN on heart disease classification using the Cleveland Heart Disease dataset. The contribution of this study is threefold: first, it formulates CCN as a clinically informed feature-representation strategy for selected cardiovascular variables; second, it compares this strategy with conventional numerical representation across five classifier families under the same repeated cross-validation framework; and third, it evaluates both statistical reliability and practical screening trade-offs through recall, effect-size analysis, and confusion-matrix interpretation. Thus, the study contributes not merely another classifier comparison, but a focused evaluation of how clinically informed feature representation interacts with different learning algorithms.

### Research Gap And Novel Contribution

Recent literature indicates that heart disease and cardiovascular prediction research has largely concentrated on classifier comparison, feature selection, feature engineering and preprocessing optimization, ensemble and deep-learning approaches, and predictive-performance optimization [2], [4], [6], [10]–[14]. Noroozi et al. [4] systematically examined multiple feature-selection strategies for heart disease prediction across several machine-learning algorithms, while Praveen et al. [6] combined hybrid feature selection with an adaptive boosted ensemble approach for

cardiovascular disease prediction. Similarly, Sumwiza et al. [15] demonstrated the effectiveness of Random Forest-based models for cardiovascular disease prediction. Collectively, these studies have contributed to algorithmic improvement and feature optimization; however, limited attention has been given to evaluating how clinically informed feature representation influences model behavior.

Recent studies have explored statistical or computational normalization, feature engineering, data harmonization, and feature-transformation strategies for cardiovascular prediction [12], [16]–[21]; however, these approaches do not specifically evaluate clinically informed feature representations derived from cardiovascular reference points and physiological relationships. As a result, the effect of embedding cardiovascular risk semantics into the preprocessing stage remains insufficiently explored.

The novelty of this study is not the development of a new classification algorithm, but the systematic evaluation of CCN as a clinically informed feature-representation strategy. Unlike previous studies that primarily focus on model optimization, feature selection, or post-hoc explainability, this research isolates the feature-representation stage and evaluates its effects across heterogeneous classifier families under the same validation framework.

The methodological contribution lies in representing selected cardiovascular variables using clinically motivated reference points and physiological relationships, thereby introducing domain-specific medical knowledge into the feature representation process. The empirical contribution is demonstrated through comparative evaluation across five model families, statistical significance testing, effect size analysis, and screening-oriented trade-off assessment. This design allows the study to determine not only whether CCN changes performance, but also whether such changes are statistically reliable and relevant to screening-oriented trade-offs.

Evidence supporting this contribution is reflected in the heterogeneous responses observed across classifier families under identical validation conditions. CCN produced different directions and magnitudes of recall change across the evaluated models, with statistically significant reductions observed for some classifiers and non-significant numerical improvements for others. These findings provide empirical evidence that clinically informed feature representation can influence classifier behavior differently even when the underlying clinical variables and validation framework are held constant.

Table 1. State of the Art Comparison and Research Gap Analysis in Heart Disease Prediction Studies

| Study | Algorithm Comparison | Feature Selection | XAI | Clinical Feature Representation | Statistical Validation | Screening Trade-off |
|---|---|---|---|---|---|---|
| Noroozi et al. [4] | ✓ | ✓ | ✗ | ✗ | ✗ | ✗ |
| Praveen et al. [6] | ✓ | ✓ | ✗ | ✗ | ✗ | ✗ |
| Sumwiza et al. [15] | ✓ | ✓ | ✗ | ✗ | ✗ | ✗ |
| Shimizu et al. [7] | ✓ | ✗ | ✓ | ✗ | ✓ | ✗ |
| Proposed Study | ✓ | ✗ | ✗ | ✓ | ✓ | ✓ |

Source: (Research Results, 2026)

The classification of prior studies shown in Table 1 was derived from the methodological descriptions reported in the respective publications. Noroozi et al. [4] evaluated multiple feature-selection strategies across several machine-learning algorithms, while Praveen et al. [6] combined hybrid feature selection with an adaptive boosted ensemble approach for cardiovascular disease prediction. Similarly, Sumwiza et al. [15] demonstrated the effectiveness of Random Forest-based models for cardiovascular disease prediction. Shimizu et al. [7] compared multiple machine-learning algorithms for cardiovascular risk prediction and incorporated LIME and Shapley values to support model interpretability, including evaluation on an external cohort. Among the representative studies summarized in Table 1, none explicitly evaluated clinically informed feature representation as a controlled methodological factor under the same classifier and validation framework.

Table 1 summarizes the positioning of the proposed study relative to previous heart disease prediction research. Existing studies predominantly focus on classifier comparison, ensemble learning, feature selection, dimensionality reduction, or explainable AI techniques. In contrast, the present study explicitly evaluates CCN as a feature representation strategy and assesses its impact through statistical significance testing, effect size analysis, and screening-oriented trade-off evaluation. This comparison highlights the


Accredited Rank 2 (Sinta 2) based on the Decree of the Dirjen Penguatan RisBang Kemenristekdikti No.225/E/KPT/2022, December 07, 2022. Published by LPPM Universitas Nusa Mandiri

VOL. 12. NO. 1 AUGUST 2026
P-ISSN: 2685-8223 | E-ISSN: 2527-4864
DOI: 10.33480/jitk.v12i1.8299

methodological gap addressed by the study and provides empirical support for its novelty contribution.

## MATERIALS AND METHODS

This study employed a quantitative experimental research design to evaluate the effect of CCN on heart disease classification performance. The experiment was structured as a comparative analysis between two feature-representation schemes: conventional numerical representation and CCN based on clinically meaningful reference points and physiological relationships. The objective was to determine whether embedding clinical semantics into feature representation influences classification reliability, particularly in terms of recall performance.

The dataset used in this research was the Cleveland Heart Disease dataset obtained from the UCI Machine Learning Repository [22]. The dataset consists of 303 patient records with 13 clinical attributes. For the present classification experiment, the diagnostic target was represented as a binary outcome indicating the absence or presence of heart disease. As shown in Table 2, Class 0, indicating no heart disease, contains 164 records (54.13%), while Class 1, indicating heart disease, contains 139 records (45.87%). This distribution indicates that the dataset is relatively balanced, with no severe class imbalance that could strongly bias the classification results.

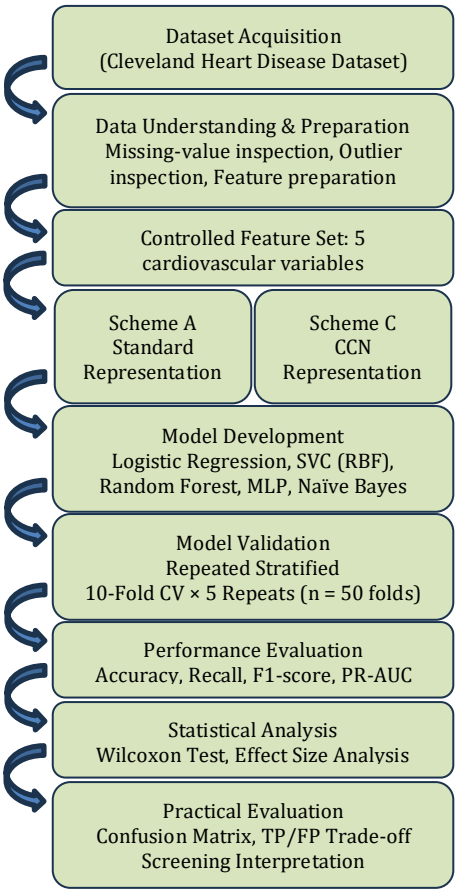
Table 2. Class Distribution of the Heart Disease Dataset

| Target Class | Clinical Meaning | Frequency | Percentage |
|---|---|---|---|
| **0** | No heart disease | 164 | 54.13% |
| **1** | Heart disease | 139 | 45.87% |
| **Total** | | **303** | **100.00%** |

Source: (Research Results, 2026)

The features include demographic and physiological variables such as age, resting blood pressure (trestbps), serum cholesterol (chol), maximum heart rate achieved (thalach), ST depression induced by exercise (oldpeak), and several categorical clinical indicators. Although the Cleveland dataset contains 13 predictor attributes, the controlled feature-representation experiment in this study focused on five continuous cardiovascular variables: age, resting blood pressure (trestbps), serum cholesterol (chol), maximum heart rate achieved (thalach), and exercise-induced ST depression (oldpeak). These variables were selected because they were directly targeted by the CCN transformations. Accordingly, the Standard and CCN schemes were compared using corresponding representations derived from the same five underlying clinical variables, allowing the effect of feature representation to be evaluated without changing the underlying clinical information being compared. The dataset was selected because of its standardized structure and continued use as a benchmark in contemporary heart-disease classification studies [4].

To ensure methodological transparency and reproducibility, the research process was structured into a series of sequential stages, beginning from data acquisition and ending with statistical and practical evaluation. The complete workflow is presented in Figure 1.




Source: (Research Results, 2026)

Figure 1. Experimental Pipeline for Comparing Standard and CCN

The study started with the acquisition of the Cleveland Heart Disease dataset from the UCI Machine Learning Repository. This dataset was selected because it is one of the most widely used benchmark datasets in cardiovascular prediction research, allowing direct comparison with previous studies. However, its relatively small sample size

and single-center origin should be considered when interpreting the generalizability of the findings.

After data acquisition, an initial data understanding stage was conducted to examine class distribution, feature characteristics, and potential data quality issues. This stage was necessary because preprocessing decisions may influence downstream model behavior and evaluation outcomes. The initial data preparation stage included examination of missing values, potential outliers, and feature characteristics. Following this inspection, the controlled experiment constructed the Standard and CCN representations from the five selected cardiovascular variables before classifier evaluation. This separation allowed the subsequent analysis to focus specifically on differences arising from feature representation.

The study then implemented two alternative feature representation strategies. The first strategy retained the selected cardiovascular variables in their conventional numerical representation, with z-score standardization incorporated within the pipelines of scale-sensitive classifiers. The second strategy applied CCN, in which the same underlying variables were represented using clinically meaningful reference points, contextual ratios, physiological relationships, and retained continuous values. This design was intended to isolate the effect of feature representation while maintaining identical experimental conditions across models.

The model development stage involved five classification algorithms representing different learning paradigms: Logistic Regression (linear model), Support Vector Machine with RBF kernel (kernel-based model), Random Forest (tree-based ensemble model), Multilayer Perceptron (neural network model), and Naïve Bayes (probabilistic model). The selection was deliberately designed to investigate whether clinically informed feature representation affects different model families in distinct ways rather than assuming uniform benefits across classifiers.

To reduce evaluation bias and improve robustness, model validation was performed using repeated stratified 10-fold cross-validation with five repetitions, resulting in 50 evaluation folds. Stratification preserved class proportions in each fold, while repetition reduced variance caused by a single random partition. For scale-sensitive classifiers, statistical scaling was incorporated within the corresponding model pipeline, with scaling parameters estimated from the training partition and subsequently applied to the corresponding validation partition.

The evaluation stage consisted of two complementary perspectives. First, predictive performance was assessed using Accuracy, Recall, F1-score, and PR-AUC. Recall was designated as the primary metric because false negative predictions may delay the identification of patients requiring further cardiovascular assessment. Second, statistical significance was examined using the Wilcoxon signed-rank test, complemented by effect size analysis to distinguish between statistical significance and practical relevance.

Finally, a practical screening-oriented analysis was conducted through confusion matrix aggregation and trade-off evaluation. This stage was included because statistically significant improvements do not automatically imply clinical usefulness. The analysis therefore examined whether increases in true positive detection were accompanied by increases in false positives, providing a more realistic assessment of the potential impact of CCN in screening-oriented decision support systems.

**Normalization Schemes**

To systematically evaluate the influence of feature representation on classification performance, two normalization scenarios were implemented in this study: **Scheme A (Standard Statistical Normalization)** and **Scheme C (Clinical CCN)**. Both schemes were applied within identical cross-validation pipelines to ensure that performance differences reflect representation effects rather than algorithmic variation.

**A. Standard Statistical Normalization (Scheme A)**

Scheme A represents the five selected cardiovascular variables—age, trestbps, chol, thalach, and oldpeak—in their conventional continuous numerical form. For scale-sensitive classifiers, z-score standardization was incorporated within the corresponding model pipeline, with the scaling parameters estimated from the training partition. Random Forest was evaluated using the same conventional numerical feature representation without scale-dependent standardization. In this approach, z-score normalization is used to standardize each feature according to its distribution within the training data. The transformation is defined as:

$$x^{!} = \frac{x-\mu}{\sigma} \quad (1)$$

where $x$ represents the original feature value, $\mu$ is the mean, and σ is the standard deviation estimated


Accredited Rank 2 (Sinta 2) based on the Decree of the Dirjen Penguatan RisBang Kemenristekdikti No.225/E/KPT/2022, December 07, 2022. Published by LPPM Universitas Nusa Mandiri

VOL. 12. NO. 1 AUGUST 2026
P-ISSN: 2685-8223 | E-ISSN: 2527-4864
DOI: 10.33480/jitk.v12i1.8299

from the corresponding training partition. This transformation was applied within the pipelines of the scale-sensitive classifiers to reduce differences in numerical scale while preserving the continuous representation of the original variables. This normalization ensures zero mean and unit variance, thereby preventing scale dominance and improving convergence behavior for distance based and gradient based models.

The objective of Scheme A is to preserve the continuous statistical structure of the data while ensuring numerical stability during model training. The following numeric attributes were normalized under Scheme A: age, resting blood pressure (trestbps), serum cholesterol (chol), maximum heart rate achieved (thalach), ST depression (oldpeak). Table 3 summarizes the conventional numerical representation used in Scheme A. The five selected cardiovascular variables were retained as continuous measurements, while z-score standardization was applied within the pipelines of scale-sensitive classifiers during model evaluation.

Table 3. Conventional Numerical Representation Used in Scheme A

| Variable | Standard Representation |
|---|---|
| age | Continuous numerical value |
| trestbps | Continuous numerical value |
| chol | Continuous numerical value |
| thalach | Continuous numerical value |
| oldpeak | Continuous numerical value |

Source: (Research Results, 2026)

**B. CCN (Scheme C)**

Scheme C applies Clinical Contextual Normalization (CCN), a clinically informed feature-representation strategy that incorporates clinically meaningful reference points and physiological relationships into selected cardiovascular variables. Unlike Scheme A, which retains the conventional continuous numerical representation, CCN combines clinically anchored categorical, ratio-based, physiologically derived, and continuous representations. The transformations were defined as follows.

**Age (age).** Age was retained as a continuous variable:

$$age_h = age$$

**Resting Blood Pressure (trestbps).** Resting blood pressure was represented using three operational categories inspired by commonly used clinical blood-pressure reference thresholds [23]. In the implemented transformation, values were mapped as >0–120 mmHg, >120–140 mmHg, and >140–300 mmHg, corresponding to CCN category values of 0, 1, and 2, respectively.

$$trestbps_{cat} = \begin{cases} 0, trestbps \leq 120 \\ 1, 120 < trestbps \leq 140 \\ 2, trestbps > 140 \end{cases}$$

**Serum Cholesterol (chol).** Rather than converting cholesterol into discrete categories, serum cholesterol was represented as a ratio relative to the clinically established reference value of 200 mg/dL [24]

$$chol_{ratio} = \frac{chol}{200}$$

**Maximum Heart Rate (thalach).** Age-predicted maximum heart rate was first calculated using the equation proposed by Tanaka et al. [25]:

$$hrmax_{pred} = 208 - 0.7(age)$$

The achieved maximum heart rate was then represented relative to this age-predicted value:

$$thalac_{ratio} = \frac{thalach}{hrmax_{pred}} \quad (2)$$

**ST Depression (oldpeak).** Exercise-induced ST depression was retained as a continuous variable:

$$oldpeak_h = oldpeak$$

Thus, the CCN model input consisted of five representations corresponding to the same five underlying cardiovascular variables used in Scheme A: $age_h$, $trestbps_{cat}$, $chol_{ratio}$, $thalach_{ratio}$, and $oldpeak_h$. The variable $hrmax_{pred}$ served only as an intermediate physiological quantity for calculating $thalach_{ratio}$ and was not included as an additional model predictor. To illustrate the contextual feature transformation applied in this study, a sample of the dataset after clinical normalization is presented in Table 4. The table illustrates how the selected clinical variables are represented using a combination of operational categories, contextual ratios, physiological relationships, and retained continuous values.

Table 4. Example of Clinical Contextual Feature Representation

| age_h | trestbps_cat | chol_ratio | hrmax_pred | thalach_ratio | oldpeak_h |
|---|---|---|---|---|---|
| 63 | 2 | 1.165 | 163.9 | 0.915 | 2.3 |
| 67 | 2 | 1.430 | 161.1 | 0.670 | 1.5 |
| 67 | 0 | 1.145 | 161.1 | 0.801 | 2.6 |

| age_h | trestbps_cat | chol_ratio | hrmax_pred | thalach_ratio | oldpeak_h |
|---|---|---|---|---|---|
| 37 | 1 | 1.250 | 182.1 | 1.027 | 3.5 |
| 41 | 1 | 1.020 | 179.3 | 0.959 | 1.4 |

Source: (Research Results, 2026)

**Note:** *$hrmax_{pred}$ is presented to illustrate the physiological derivation of $thalach_{ratio}$. It is an intermediate quantity and was not included as an additional predictor during model training. Therefore, both Scheme A and Scheme C used five corresponding model-input features.*

The contextual transformation was designed to incorporate clinically meaningful reference points and physiological relationships directly into feature representation. The transformations were heterogeneous rather than uniformly categorical: resting blood pressure was represented categorically, cholesterol and maximum heart rate were expressed through contextual ratios, while age and oldpeak retained their continuous values. Consequently, CCN modifies the representation of selected variables without uniformly discretizing the entire feature space. This design preserves correspondence with the five underlying cardiovascular variables while introducing clinical context into selected representations.

Tables 3 and 4 illustrate two different feature-representation strategies applied to the same underlying cardiovascular variables. Scheme A retains the variables in their conventional continuous numerical form, with z-score standardization incorporated for scale-sensitive classifiers. In contrast, CCN introduces clinically motivated categorical, ratio-based, and physiological representations while retaining selected variables continuously. Consequently, performance differences between the two schemes should be interpreted as effects associated with alternative representations of the same underlying clinical information rather than as the consequence of adding different clinical predictors.

### Conceptual Comparison Between Scheme A and Scheme C

Scheme A retains the conventional continuous numerical representation of the five selected cardiovascular variables and applies statistical scaling where required by the classifier. In contrast, Scheme C introduces clinical context through a combination of categorical, ratio-based, physiologically derived, and continuous representations. Therefore, the principal distinction between the two schemes lies in how the same underlying clinical information is represented for machine learning rather than in the inclusion of additional predictor variables. Both schemes were evaluated using the same classifier configurations and repeated stratified cross-validation framework. Accordingly, the observed performance differences were examined primarily in relation to the alternative feature-representation strategies rather than differences in classifier configuration. For model evaluation, five classification algorithms were implemented: Logistic Regression, Support Vector Machine with RBF kernel, Random Forest, Multilayer Perceptron (MLP), and Naïve Bayes. The experimental evaluation employed repeated stratified 10-fold cross-validation with five repetitions, resulting in 50 evaluation folds.

While existing studies predominantly emphasize algorithmic optimization, model ensembling, and hyperparameter tuning, this study focuses on the role of feature representation in medical classification. CCN introduces clinically informed representations through a combination of reference-based categorization, contextual ratios, and physiological relationships while preserving selected variables in continuous form. The approach is evaluated across heterogeneous classifier families using a common repeated cross-validation framework, paired statistical testing, effect-size analysis, and screening-oriented confusion-matrix interpretation. This design enables the study to examine whether alternative representations of the same underlying cardiovascular information produce different responses across learning algorithms.

## RESULTS AND DISCUSSION

To evaluate the impact of normalization strategy, two experimental pipelines were implemented. Pipeline A retained the five selected variables in their conventional numerical representation, with z-score standardization incorporated within the pipelines of scale-sensitive classifiers, whereas Pipeline C utilized CCN through clinically motivated categorical, ratio-based, physiological, and continuous representations. All classifiers were trained and evaluated under identical cross-validation conditions to ensure fair comparison. The performance differences reported in this section therefore reflect the influence of feature representation rather than algorithmic tuning. This section presents the experimental results and provides a comprehensive discussion of their implications in relation to the research objective, namely evaluating the impact of CCN on heart disease classification performance.


Accredited Rank 2 (Sinta 2) based on the Decree of the Dirjen Penguatan RisBang Kemenristekdikti No.225/E/KPT/2022, December 07, 2022. Published by LPPM Universitas Nusa Mandiri

VOL. 12. NO. 1 AUGUST 2026
P-ISSN: 2685-8223 | E-ISSN: 2527-4864
DOI: 10.33480/jitk.v12i1.8299

**Performance Comparison Between Normalization Schemes**

The first analysis compares classification performance between the standard statistical normalization scheme and the CCN scheme. The average results across 50 cross-validation folds are summarized in Table 5.

Table 5. Average Classification Performance Across Normalization Schemes

| Dataset | Model | Accuracy | Recall | F1-score | PR-AUC |
|---|---|---|---|---|---|
| Clinical | LR | 0.7089 | 0.6589 | 0.6720 | 0.7835 |
| Clinical | MLP | 0.6876 | 0.6012 | 0.6357 | 0.7643 |
| Clinical | NB | 0.7194 | 0.5982 | 0.6575 | 0.7936 |
| Clinical | RF | 0.7097 | 0.6585 | 0.6733 | 0.7791 |
| Clinical | SVC | 0.6852 | 0.5992 | 0.6349 | 0.7291 |
| Standard | LR | 0.7162 | 0.6704 | 0.6819 | 0.7893 |
| Standard | MLP | 0.6798 | 0.5869 | 0.6226 | 0.7591 |
| Standard | NB | 0.7168 | 0.6315 | 0.6686 | 0.7900 |
| Standard | RF | 0.7044 | 0.6525 | 0.6675 | 0.7737 |
| Standard | SVC | 0.7247 | 0.6138 | 0.6684 | 0.7686 |

Source: (Research Results, 2026)

Table 5 demonstrates that the effect of CCN varies across classifier families. RF and MLP showed numerical increases in mean recall under CCN, whereas LR, SVC, and NB showed lower mean recall than under the Standard representation. These differences indicate that alternative representations of the same underlying cardiovascular variables may interact differently with the learning characteristics of individual classifiers. However, differences in mean performance alone do not establish statistical superiority; therefore, the reliability of these changes was further evaluated using paired fold-level statistical testing. Accordingly, the effectiveness of CCN should be interpreted as a classifier-dependent phenomenon rather than as a general preprocessing improvement applicable to all machine-learning algorithms. To formally represent the primary evaluation metric used in this study, recall is defined as:

$$Recall = \frac{TP}{TP+FN} \quad (3)$$

where TP represents the number of correctly classified positive cases (true positives), and FN represents the number of false negatives. In this study, recall was prioritized because false-negative predictions correspond to positive cases that are not identified by the classifier. The numerical decreases observed for SVC and NB further illustrate that clinically informed representation is not necessarily beneficial for every classifier. One possible explanation is that the CCN transformations modify feature distributions and numerical relationships in ways that interact differently with classifier assumptions. Kernel-based methods such as SVC depend on the geometry of the input feature space, whereas probabilistic classifiers such as NB are sensitive to the distributional characteristics of the represented variables. Importantly, however, these mechanisms should be interpreted as plausible explanations rather than established causal effects. The subsequent paired statistical analysis distinguishes which observed recall changes were sufficiently consistent across folds to reach statistical significance. Additionally, the average recall across repeated cross-validation folds is computed as:

$$\bar{R} = \frac{1}{k}\sum_{i=1}^{k} R_i \quad (4)$$

where $R_i$ denotes the recall value obtained in fold $i$, and $k$ is the total number of evaluation folds (k = 50 in this study). This formulation ensures that performance estimation reflects stability across multiple data partitions rather than a single split.

**Statistical Significance Analysis**

To determine whether the performance differences between the Standard and CCN schemes were statistically significant, the Wilcoxon signed-rank test was applied to paired recall values obtained from the repeated cross-validation folds. The test was used to assess whether the paired recall differences were systematically centered around zero without assuming normality of the paired differences. Let $d_i = R_{CCN,i} - R_{standard,i}$ represent the difference in recall for fold $i$, where $i$ = 1,2,...,$n$ and $n$ = 50. The signed-rank quantity can be expressed as:

$$W = \sum_{i=1}^{n} sgn(d_i).R_i \quad (5)$$

where:
$d_i$ = paired difference between schemes
$sgn(d_i)$ = sign of the difference (+1 or −1)
$R_i$ = rank of the absolute difference $|d_i|$

The test evaluates whether the paired differences are systematically centered around zero based on their signed ranks. Under the null hypothesis, the distribution of positive and negative ranks is symmetric. A p-value below 0.05 indicates statistically significant performance differences between normalization schemes. The results are summarized in Table 6.

Table 6. Statistical Comparison Using Wilcoxon Signed-Rank Test

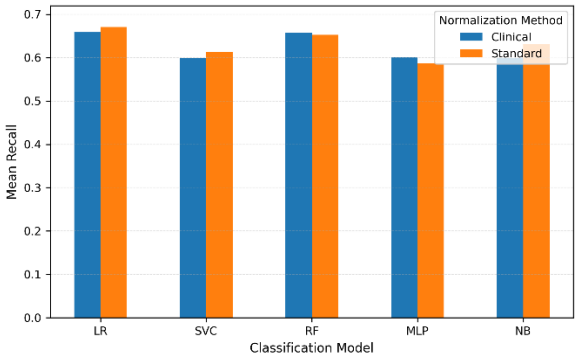
| Model | ΔRecall | p-value (Recall) | Effect size r | Effect magnitude | n | Interpretation |
|---|---|---|---|---|---|---|
| LR | −0.0115 | **0.0186** | −0.333 | Moderate | 50 | Significant Decrease |
| SVC | −0.0146 | 0.1222 | −0.219 | Small | 50 | Not Significant |
| RF | +0.0059 | 0.6434 | +0.065 | Negligible | 50 | Not Significant |
| MLP | +0.0143 | 0.2195 | +0.174 | Small | 50 | Not Significant |
| NB | −0.0333 | **<0.001** | −0.560 | Large | 50 | Significant Decrease |

Source: (Research Results, 2026)

Table 6 presents the paired Wilcoxon signed-rank analysis of recall differences between Standard representation and CCN across the 50 matched cross-validation folds. The results demonstrate a heterogeneous and classifier-dependent response to CCN. Logistic Regression showed a statistically significant decrease in recall (ΔRecall = −0.0115, p = 0.0186, r = −0.333), corresponding to a moderate effect. Naïve Bayes exhibited the largest statistically significant recall reduction (ΔRecall = −0.0333, p < 0.001, r = −0.560), with a large effect size. In contrast, Random Forest showed a small numerical increase in recall (ΔRecall = +0.0059), but the difference was not statistically significant (p = 0.6434, r = 0.065). MLP similarly exhibited a numerical recall increase (ΔRecall = +0.0143) without statistical significance (p = 0.2195, r = 0.174). SVC showed a numerical recall reduction (ΔRecall = −0.0146), although the paired difference was also not statistically significant (p = 0.1222, r = −0.219).

The Wilcoxon signed-rank test was conducted using 50 paired fold-level recall scores obtained from the repeated cross-validation scheme (10 folds × 5 repeats). Effect-size analysis was used alongside the p-value to characterize the magnitude and direction of the paired differences. Only LR and NB showed statistically significant differences, and both effects represented decreases in recall under CCN. RF and MLP showed positive numerical recall changes, but these differences were not statistically significant, while the numerical decrease observed for SVC was also not statistically significant. These findings indicate that CCN does not produce a uniform performance advantage across classifier families. Instead, its effect depends on the interaction between the alternative feature representation and the learning characteristics of each classifier.

To further illustrate the impact of CCN, a comparative visualization of recall performance across models is presented in Figure 2. The graph displays the average recall obtained under both normalization schemes for each classification algorithm.




Source: (Research Results, 2026)

Figure 2. Comparison of Recall Performance Between Standard and Clinical Normalization

Figure 2 illustrates the classifier-dependent recall responses to CCN. RF and MLP showed numerical increases in mean recall, whereas LR, SVC, and NB showed numerical decreases relative to the Standard representation. Among the evaluated classifiers, MLP exhibited the largest positive numerical recall change (ΔRecall = +0.0143), followed by RF (ΔRecall = +0.0059). However, as demonstrated by the paired statistical analysis in Table 6, neither increase was statistically significant. Therefore, Figure 2 should be interpreted as a descriptive visualization of model-specific recall changes rather than evidence that a particular classifier family consistently benefits from CCN.

**Trade-Off Analysis of Detection and False Alarm**

Beyond statistical averages, practical screening systems must balance sensitivity and false alarms. Table 7 summarizes the aggregated confusion matrix comparison.

Table 7. Aggregated Confusion Matrix Comparison for MLP

| Scheme | TP | FP | FN | TN |
|---|---|---|---|---|
| Standard | 408 | 198 | 287 | 622 |
| CCN | 418 | 196 | 277 | 624 |
| Change | +10 | -2 | -10 | +2 |

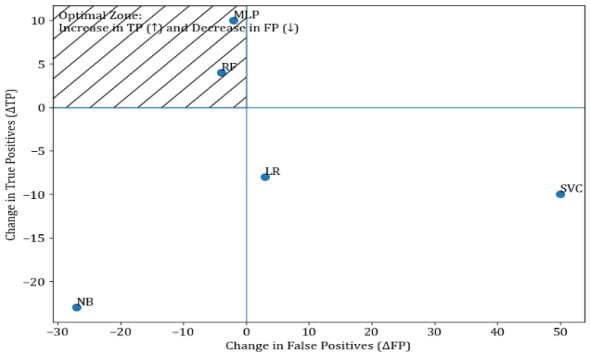
Source: (Research Results, 2026)

Table 7 presents the aggregated confusion-matrix results for MLP across the repeated cross-validation evaluations. Under CCN, the aggregated number of true-positive predictions increased from 408 to 418, while false-negative predictions


Accredited Rank 2 (Sinta 2) based on the Decree of the Dirjen Penguatan RisBang Kemenristekdikti No.225/E/KPT/2022, December 07, 2022. Published by LPPM Universitas Nusa Mandiri

VOL. 12. NO. 1 AUGUST 2026
P-ISSN: 2685-8223 | E-ISSN: 2527-4864
DOI: 10.33480/jitk.v12i1.8299

decreased from 287 to 277. False-positive predictions also decreased slightly from 198 to 196, accompanied by an increase in true negatives from 622 to 624. Thus, the MLP results show a favorable descriptive screening pattern, with more positive predictions correctly identified without an accompanying increase in false-positive predictions. Because these counts are aggregated across repeated cross-validation folds, they should be interpreted as cumulative prediction outcomes rather than as counts of unique patients. To analyze detection reliability and false alarm control simultaneously, a trade-off visualization is presented in Figure 3. The graph plots the change in true positives (ΔTP) on the vertical axis and change in false positives (ΔFP) on the horizontal axis.




Source: (Research Results, 2026)
Figure 3. Descriptive Trade-Off of Changes in True-Positive and False-Positive Predictions

Figure 3 provides a descriptive visualization of changes in true-positive and false-positive predictions under CCN relative to the Standard representation. The upper-left quadrant (ΔTP > 0 and ΔFP < 0) represents a potentially favorable screening pattern because additional positive cases are correctly detected without an increase in false-positive predictions. MLP and RF are positioned in this quadrant, indicating favorable descriptive trade-offs in their aggregated prediction outcomes. However, these patterns should be interpreted together with the inferential results in Table 6. Neither the recall increase for MLP nor that for RF was statistically significant. Therefore, their position in the favorable quadrant indicates a descriptive screening pattern rather than evidence of statistically reliable superiority of CCN. Conversely, the responses of the other classifiers further demonstrate that the direction of the CCN effect is model dependent.

Reducing false-negative predictions while avoiding unnecessary increases in false-positive predictions is an important practical consideration in the screening-oriented interpretation of the present results. From a screening-oriented perspective, the MLP results are noteworthy because the aggregated prediction outcomes under CCN included 10 more true-positive predictions and 10 fewer false-negative predictions, while false positives decreased by two. This pattern suggests that, for MLP in the present experiment, the alternative feature representation produced a favorable descriptive balance between positive-case detection and false-alarm control. However, these values represent aggregated prediction outcomes across repeated cross-validation folds and should not be interpreted as 10 additional unique patients detected.

The magnitude and reliability of this pattern should also not be overstated. The paired recall difference for MLP was not statistically significant ($p = 0.2195$, $r = 0.174$), and comparable benefits were not consistently observed across the other classifiers. Accordingly, the MLP trade-off provides complementary descriptive information to the fold-level statistical analysis rather than evidence of a general clinical benefit of CCN. These findings motivate further investigation of CCN in screening-oriented applications, particularly to determine whether the favorable descriptive pattern observed for MLP can be reproduced in larger and externally validated clinical cohorts.

**Descriptive Analysis of Feature Representation Scale**

An additional descriptive analysis examined how the numerical scale of selected features changed after CCN transformation. Because CCN includes categorical and ratio-based representations, the resulting features are expressed on scales that differ from those of the original measurements. Table 8 therefore reports the standard deviations of the corresponding representations as a descriptive illustration of these scale changes rather than as a direct measure of improved statistical stability.

Table 8. Descriptive Standard Deviation of Standard and CCN Feature Representations

| Underlying Variable | SD of Standard Representation | SD of CCN Representation |
|---|---|---|
| trestbps | 17.5997 | 0.7275 |
| chol | 51.7769 | 0.2589 |
| thalach | 22.8750 | 0.1258 |
| oldpeak | 1.1611 | 1.1611 |
| age | 9.0387 | 9.0387 |

Source: (Research Results, 2026)
**Note:** *Standard deviations across the two representation schemes are descriptive and are not directly comparable as measures of variance*

*reduction because several CCN features are expressed on transformed numerical scales.*

Table 8 illustrates that CCN changes the numerical scale of several feature representations. The standard deviations of age and oldpeak remain unchanged because these variables retain their original continuous values. In contrast, cholesterol is expressed relative to a reference value of 200 mg/dL, maximum heart rate is represented as a ratio to age-predicted maximum heart rate, and resting blood pressure is represented categorically. Consequently, their numerical dispersion differs from that of the original measurements.

These differences should not be interpreted as evidence that CCN inherently reduces noise or improves feature stability, because the transformed variables are expressed on different numerical scales. Instead, the table provides a descriptive view of how clinically informed transformations alter the statistical representation presented to the learning algorithms. Such changes in representation may contribute to the classifier-dependent responses observed in Tables 5 and 6, but the present experiment does not isolate the specific mechanism responsible for each model's response. Therefore, relationships between transformed feature distributions and classifier behavior should be regarded as hypotheses for further investigation rather than causal explanations established by the current results.

**Clinical Implication**

The potential clinical relevance of CCN lies in incorporating clinically meaningful reference points and physiological relationships into selected feature representations rather than relying exclusively on conventional numerical representation. In the present implementation, resting blood pressure was represented categorically, cholesterol was expressed relative to a clinical reference value, maximum heart rate was contextualized using an age-predicted physiological relationship, while age and exercise-induced ST depression retained their continuous values. This approach is intended to reduce the semantic gap between selected clinical measurements and their representation for machine-learning analysis.

The empirical findings do not demonstrate a uniform screening advantage of CCN. RF and MLP showed positive numerical recall changes, but neither difference was statistically significant. Nevertheless, MLP exhibited a favorable descriptive screening pattern in the aggregated repeated cross-validation predictions, with 10 more true-positive predictions and 10 fewer false-negative predictions while false positives decreased by two. Conversely, LR and NB showed statistically significant recall reductions under CCN. These heterogeneous findings indicate that clinically informed representation should not be assumed to improve screening performance across classifier families. From a future implementation perspective, CCN could be investigated as a feature-representation component within an offline clinical decision-support pipeline. Routine cardiovascular measurements could first be transformed into their corresponding CCN representations before being processed by a validated classification model. However, such integration should only be considered after external validation on larger, multi-center, and clinically representative cohorts. The present results should therefore be viewed as methodological evidence regarding feature representation rather than as validation of a deployable clinical decision-support system.

The observed MLP trade-off warrants careful interpretation. In screening-oriented applications, reducing false-negative predictions is generally desirable because missed positive cases may delay further clinical assessment. In the present experiment, the aggregated MLP predictions under CCN showed more true positives and fewer false negatives without an increase in false positives. However, these counts were accumulated across repeated cross-validation folds and do not represent additional unique patients detected. Moreover, the paired recall difference for MLP was not statistically significant. Accordingly, this pattern should be interpreted as a descriptive screening signal that warrants external investigation rather than as evidence of established clinical benefit.

Several limitations should be considered before translating these findings into clinical practice. The study was conducted using the Cleveland Heart Disease dataset, which contains only 303 records and represents a single benchmark population. Consequently, the observed classifier-dependent effects may not necessarily generalize to larger, multi-center, or demographically diverse patient cohorts. Furthermore, the proposed normalization strategy was evaluated in an offline experimental setting rather than within an operational clinical workflow. Future studies should therefore examine its robustness using external validation datasets, electronic health record data, and prospective clinical decision-support environments.


Accredited Rank 2 (Sinta 2) based on the Decree of the Dirjen Penguatan RisBang Kemenristekdikti No.225/E/KPT/2022, December 07, 2022. Published by LPPM Universitas Nusa Mandiri

VOL. 12. NO. 1 AUGUST 2026
P-ISSN: 2685-8223 | E-ISSN: 2527-4864
DOI: 10.33480/jitk.v12i1.8299

**Overall Discussion**

The findings of this study demonstrate that feature representation is an important methodological factor in heart disease classification. Previous studies have predominantly emphasized classifier comparison, feature selection, ensemble learning, or explainability-oriented modeling [4], [6], [7]. In contrast, the present results show that modifying the representation of selected clinical variables can alter model behavior, recall performance, and screening-oriented trade-offs even when the classifier configurations and validation procedures are maintained.

This finding supports the view that feature representation should be evaluated as part of model design rather than treated merely as a routine preprocessing step. A notable observation is that the effect of CCN is strongly classifier dependent. RF and MLP showed positive numerical changes in recall under CCN, but neither difference was statistically significant. RF exhibited a small numerical recall increase ($\Delta$Recall = +0.0059, $p = 0.6434$, $r = 0.065$), whereas MLP showed the largest positive numerical change ($\Delta$Recall = +0.0143, $p = 0.2195$, $r = 0.174$). Although these results do not establish statistical superiority, the aggregated MLP predictions showed a favorable descriptive screening pattern, with more true-positive and fewer false-negative predictions without an increase in false positives. Therefore, the positive responses observed for RF and MLP should be interpreted as model-specific descriptive patterns rather than evidence that CCN consistently improves predictive performance.

A different response was observed for LR, SVC, and NB. Logistic Regression experienced a statistically significant reduction in recall ($\Delta$Recall = −0.0115, $p = 0.0186$, $r = −0.333$), corresponding to a moderate effect, while Naïve Bayes showed the largest significant reduction ($\Delta$Recall = −0.0333, $p < 0.001$, $r = −0.560$), corresponding to a large effect. SVC also showed a numerical recall decrease ($\Delta$Recall = −0.0146), although the difference was not statistically significant ($p = 0.1222$, $r = −0.219$). These heterogeneous responses demonstrate that clinically informed feature representation can produce positive, negligible, or adverse performance changes depending on the interaction between the transformed feature space and the learning characteristics of the classifier.

These findings highlight a broader methodological issue in medical artificial intelligence: clinical meaning and statistical information are not necessarily equivalent. Clinically motivated transformations can make selected representations more closely aligned with medical reference points or physiological relationships, but they can also modify the numerical structure available to a learning algorithm. In the present CCN implementation, these changes are heterogeneous: resting blood pressure is represented categorically, cholesterol and maximum heart rate are expressed through contextual ratios, while age and oldpeak remain continuous. Consequently, the effect cannot be attributed solely to discretization. Rather, each classifier is exposed to an alternative representation of the same underlying clinical information, and its response depends on how that representation interacts with its learning characteristics.

From a research perspective, the study also relates to the broader development of interpretable cardiovascular machine learning. Recent cardiovascular prediction research has combined predictive modeling with post-hoc interpretability methods such as LIME and Shapley values [7], [8]. In contrast, clinically informed feature representation provides one possible avenue for incorporating domain knowledge before model training. However, the present study evaluates its effects on classifier behavior and does not establish that CCN itself improves model explainability. Future studies should therefore directly evaluate whether clinically informed representations improve clinician understanding, explanation quality, or trust in model outputs.

The novelty of this study therefore lies not in proposing a new classifier or optimization algorithm, but in systematically evaluating CCN as a clinically informed feature-representation strategy under controlled experimental conditions. Unlike previous studies that primarily compare algorithms, feature-selection approaches, or ensemble architectures [4], [6], [15], [26], this work isolates the feature-representation stage by comparing alternative representations of the same five underlying cardiovascular variables across heterogeneous classifier families under a common repeated cross-validation framework. The resulting heterogeneous responses provide empirical evidence that the influence of clinically informed representation is classifier dependent rather than universally beneficial.

Given these limitations, future studies should validate these findings using larger multi-center datasets, electronic health record data, and prospective clinical implementations. The analysis was conducted exclusively on the Cleveland Heart Disease dataset, which contains only 303 observations and represents a single benchmark

population. Consequently, the observed effects may not generalize to broader patient populations, different healthcare environments, or real-world clinical settings. Furthermore, the study evaluated CCN within an offline experimental framework rather than an operational clinical decision-support system.

## CONCLUSION

This study investigated the effect of Clinical Contextual Normalization (CCN) on heart disease classification by comparing clinically informed feature representations with conventional numerical representations across five machine-learning classifiers. The results demonstrate that the influence of CCN is classifier dependent rather than uniformly beneficial. Random Forest showed a small numerical increase in recall (ΔRecall = +0.0059), but the paired difference was not statistically significant ($p = 0.6434$, $r = 0.065$). MLP exhibited the largest positive numerical recall change (ΔRecall = +0.0143), although the difference was also not statistically significant ($p = 0.2195$, $r = 0.174$). Nevertheless, its aggregated repeated cross-validation predictions showed a favorable descriptive screening pattern, with 10 more true-positive and 10 fewer false-negative predictions while false positives decreased by two. In contrast, Logistic Regression showed a statistically significant recall decrease (ΔRecall = −0.0115, $p = 0.0186$, $r = −0.333$), while Naïve Bayes exhibited the largest significant reduction (ΔRecall = −0.0333, $p < 0.001$, $r = −0.560$). SVC also showed a numerical recall reduction (ΔRecall = −0.0146), but the difference was not statistically significant ($p = 0.1222$, $r = −0.219$).

These heterogeneous responses indicate that incorporating clinical context into feature representation can alter classifier behavior, but its effect depends on the characteristics of the learning algorithm. Therefore, CCN should be evaluated jointly with classifier selection rather than assumed to provide universal predictive improvement. The findings should be interpreted within the limitations of the relatively small, single-cohort Cleveland Heart Disease dataset. In addition, the favorable MLP confusion-matrix pattern represents aggregated repeated cross-validation predictions rather than additional unique patients and does not establish clinical benefit. Future research should therefore validate CCN using larger external and multi-center cohorts and investigate which clinically informed transformations are beneficial for particular classifier families before considering implementation in clinical decision-support environments.

## REFERENCE

[1] World Health Organization, 'Cardiovascular diseases (CVDs)'. 2025.

[2] T. Banerjee and İ. Paçal, "A systematic review of machine learning in heart disease prediction," Turkish Journal of Biology, vol. 49, no. 5, pp. 600–634, 2025, doi: 10.55730/1300-0152.2766.

[3] H. A. Al-Alshaikh et al., "Comprehensive evaluation and performance analysis of machine learning in heart disease prediction," Scientific Reports, vol. 14, no. 1, p. 7819, 2024, doi: 10.1038/s41598-024-58489-7.

[4] Z. Noroozi, A. Orooji, and L. Erfannia, "Analyzing the impact of feature selection methods on machine learning algorithms for heart disease prediction," Scientific Reports, vol. 13, no. 1, p. 22588, 2023, doi: 10.1038/s41598-023-49962-w.

[5] S. Srinivasan et al., "An active learning machine technique based prediction of cardiovascular heart disease from UCI-repository database," Scientific Reports, vol. 13, no. 1, p. 13588, 2023, doi: 10.1038/s41598-023-40717-1.

[6] S. P. Praveen et al., "Enhanced feature selection and ensemble learning for cardiovascular disease prediction: hybrid GOL2-2T and adaptive boosted decision fusion with babysitting refinement," Frontiers in Medicine, vol. 11, p. 1407376, 2024, doi: 10.3389/fmed.2024.1407376.

[7] G. Y. Shimizu et al., "Machine learning-based risk prediction for major adverse cardiovascular events in a Brazilian hospital: Development, external validation, and interpretability," PLOS ONE, vol. 19, no. 10, p. e0311719, 2024, doi: 10.1371/journal.pone.0311719.

[8] J. Wang, Q. Xue, C. W. J. Zhang, K. K. L. Wong, and Z. Liu, "Explainable coronary artery disease prediction model based on AutoGluon from AutoML framework," Frontiers in Cardiovascular Medicine, vol. 11, p. 1360548, 2024, doi: 10.3389/fcvm.2024.1360548.

[9] H. El-Sofany, B. Bouallegue, and Y. M. A. El-Latif, "A Proposed Technique for Predicting Heart Disease Using Machine Learning Algorithms and an Explainable AI Method," Scientific Reports, vol. 14, no. 1, p. 23277,

VOL. 12. NO. 1 AUGUST 2026
P-ISSN: 2685-8223 | E-ISSN: 2527-4864
DOI: 10.33480/jitk.v12i1.8299

2024, doi: 10.1038/s41598-024-74656-2.
[10] V. Chang, V. Bhavani, A. Xu, and M. Hossain, "An artificial intelligence model for heart disease detection using machine learning algorithms," Healthcare Analytics, vol. 2, p. 100016, 2022, doi: 10.1016/j.health.2022.100016.
[11] S. Subramani et al., "Cardiovascular diseases prediction by machine learning incorporation with deep learning," Frontiers in Medicine, vol. 10, p. 1150933, 2023, doi: 10.3389/fmed.2023.1150933.
[12] M. Bouqentar et al., "Early heart disease prediction using feature engineering and machine learning algorithms," Heliyon, vol. 10, no. 19, p. e38731, 2024, doi: 10.1016/j.heliyon.2024.e38731.
[13] M. A. Islam, M. Z. H. Majumder, M. S. Miah, and S. Jannaty, "Precision healthcare: A deep dive into machine learning algorithms and feature selection strategies for accurate heart disease prediction," Computers in Biology and Medicine, vol. 176, p. 108432, 2024, doi: 10.1016/j.compbiomed.2024.108432.
[14] V. R. Kukkala, S. P. Praveen, N. S. K. M. K. Tirumanadham, and P. N. Srinivasu, "A Study on Outlier Detection and Feature Engineering Strategies in Machine Learning for Heart Disease Prediction," Computer Systems Science and Engineering, vol. 48, no. 5, pp. 1085–1112, 2024, doi: 10.32604/csse.2024.053603.
[15] K. Sumwiza, C. Twizere, G. Rushingabigwi, P. Bakunzibake, and P. Bamurigire, "Enhanced cardiovascular disease prediction model using random forest algorithm," Informatics in Medicine Unlocked, vol. 41, p. 101316, 2023, doi: 10.1016/j.imu.2023.101316.
[16] P. R. Kumar, P. Chakrabarti, T. Chakrabarti, B. Unhelkar, and M. Margala, "Heart disease prediction using spark architecture with fused feature set and hybrid Squeezenet-Linknet model," Biomedical Signal Processing and Control, vol. 100, p. 107070, 2025, doi: 10.1016/j.bspc.2024.107070.
[17] M. Hosseini Chagahi, S. Mohammadi Dashtaki, B. Moshiri, and M. D. J. Piran, "Cardiovascular disease detection using a novel stack-based ensemble classifier with aggregation layer, DOWA operator, and feature transformation," Computers in Biology and Medicine, vol. 173, p. 108345, 2024, doi: 10.1016/j.compbiomed.2024.108345.
[18] S. G. Taj and K. Kalaivani, "Hybrid prediction model with improved score level fusion for heart disease diagnosis," Computational Biology and Chemistry, vol. 113, p. 108278, 2024, doi: 10.1016/j.compbiolchem.2024.108278.
[19] R. Amin et al., "Harmonization of Heart Disease Dataset for Accurate Diagnosis: A Machine Learning Approach Enhanced by Feature Engineering," Computers, Materials & Continua, vol. 82, no. 3, pp. 3907–3919, 2025, doi: 10.32604/cmc.2025.059942.
[20] S. U. Değer, "A study on heart data analysis and prediction using advanced machine learning methods," Computers in Biology and Medicine, vol. 192, no. B, p. 110308, 2025, doi: 10.1016/j.compbiomed.2025.110308.
[21] D. Yadav, D. Rani, and O. P. Verma, "Impact of feature selection and feature engineering in prediction of cardiovascular diseases," Computers in Biology and Medicine, vol. 197, no. B, p. 111027, 2025, doi: 10.1016/j.compbiomed.2025.111027.
[22] A. Janosi, W. Steinbrunn, M. Pfisterer, and R. Detrano, 'Heart Disease'. UCI Machine Learning Repository, 1989.
[23] McEvoy, J. W., McCarthy, C. P., Bruno, R. M., Brouwers, S., Canavan, M. D., Ceconi, C., et al. (2024). 2024 ESC Guidelines for the management of elevated blood pressure and hypertension. European Heart Journal, 45(38), 3912–4018.
[24] Cao, J., Donato, L., El-Khoury, J. M., Goldberg, A., Meeusen, J. W., & Remaley, A. T. (2024). ADLM Guidance Document on the Measurement and Reporting of Lipids and Lipoproteins. The Journal of Applied Laboratory Medicine, 9(5), 1040–1056. https://doi.org/10.1093/jalm/jfae057.
[25] Almaadawy, O., Uretsky, B. F., Krittanawong, C., & Birnbaum, Y. (2024). Target Heart Rate Formulas for Exercise Stress Testing: What Is the Evidence? Journal of Clinical Medicine, 13(18), 5562. https://doi.org/10.3390/jcm13185562.
[26] W. Jian et al., "Feature elimination and stacking framework for accurate heart disease detection in IoT healthcare systems using clinical data," Frontiers in Medicine, vol. 11, p. 1362397, 2024, doi: 10.3389/fmed.2024.1362397.