

Cleveland Heart Disease Dataset: From Primary Data to an AI-Ready Resource.

Gennady Chuiko

henadiy.chuiko@chmnu.edu.ua

Petro Mohyla Black Sea National University <https://orcid.org/0000-0001-5590-9404>

Olga Yaremchuk

Petro Mohyla Black Sea National University

Research Article

Keywords: data engineering, AI-ready dataset, Cleveland Heart Disease Dataset, feature selection, label-noise removal, clinical AI, biomedical data quality.

Posted Date: October 22nd, 2026

DOI: <https://doi.org/10.21203/rs.3.rs-11455193/v1>

License:   This work is licensed under a Creative Commons Attribution 4.0 International License.

[Read Full License](#)

Additional Declarations: The authors declare no competing interests.

CLEVELAND HEART DISEASE DATASET: FROM PRIMARY DATA TO AN AI-READY RESOURCE.

Chuiko G.P. – Dr. Sc., Professor, Professor of the Department of Computer Engineering, Petro Mohyla Black Sea National University, Mykolaiv, Ukraine

Yaremchuk O.M. - deputy director of the Medical Institute, Senior Lecturer of the Department of Medical Biology and Physics, Microbiology, Histology, Physiology, and Pathophysiology, Petro Mohyla Black Sea National University, Mykolaiv, Ukraine

ABSTRACT

Background. The Cleveland Heart Disease Dataset is one of the most widely used benchmark resources for cardiovascular risk prediction. Despite its extensive use in clinical artificial intelligence research, relatively little attention has been paid to dataset quality, feature redundancy, and label noise, which may affect the reliability and reproducibility of predictive models.

Objective. This study aims to transform the Cleveland Heart Disease Dataset from primary clinical data into an AI-ready resource through systematic data engineering.

Methods. An integrated data-engineering workflow was developed using WEKA. Ensemble feature selection based on eight independent attribute evaluators was applied to identify informative clinical variables and reduce dataset dimensionality. Label noise was detected using the CAIRAD algorithm. Potentially mislabeled records were removed, resulting in a reduced and cleansed dataset. The utility of the resulting resource was evaluated using three neural-network-based classifiers and repeated cross-validation.

Results. Feature selection reduced the number of predictive attributes from 13 to 7, while label-noise detection identified 49 potentially mislabeled records. The resulting AI-ready resource contained 254 instances. Across all evaluated classifiers, the engineered dataset demonstrated statistically significant improvements in Accuracy, Precision, Recall, F1-score, Matthews Correlation Coefficient, and Precision-Recall AUC compared with the original dataset. Differences among classifiers were substantially smaller than those associated with dataset engineering.

Conclusions. Systematic feature reduction and label-noise removal transformed the Cleveland Heart Disease Dataset into an AI-ready resource with improved predictivity and reduced data complexity. The proposed workflow highlights the importance of health data quality and reproducible data-engineering practices for clinical AI and decision-support applications.

Keywords: data engineering, AI-ready dataset, Cleveland Heart Disease Dataset, feature selection, label-noise removal, clinical AI, biomedical data quality.

ABBREVIATION

AI is Artificial Intelligence

ANN is the artificial neural networks

CAIRAD is a co-appearance-based analysis for incorrect records and attribute-value detection

DL is Deep Learning

MCC is the Matthews correlation coefficient

ML is Machine Learning

MLP is a Multi-layer Perceptron

PRC-AUC is Area Under Precision-Recall Curve

ROC-AUC is the Area Under the Receiver Operating Characteristic Curve

SVM is the Support Vector machine

UCI is the University of California, Irvine

WEKA is the Waikato Environment for Knowledge Analysis

NOMENCLATURE

ds0 – original dataset

ds1 – reduced and cleansed AI-ready dataset

F1-score is a known measure of predictive performance

1 INTRODUCTION

1.1 Background

Cardiovascular diseases remain one of the leading causes of mortality worldwide and continue to represent a major challenge for healthcare systems. Consequently, machine learning (ML) and artificial intelligence (AI) methods are increasingly being investigated as tools for cardiovascular risk assessment and clinical decision support [1].

Among publicly available benchmark datasets, the Cleveland Heart Disease Dataset [2] occupies a special position. Since its introduction, this dataset has been extensively used for the development, validation, and comparison of machine learning models for heart disease prediction [1]. As a result, it has become one of the most frequently studied datasets in clinical AI research.

Traditionally, research on the Cleveland dataset has focused on improving predictive performance through the development of new classification algorithms, neural-network architectures, ensemble methods, and hyperparameter optimization strategies [1]. However, predictive performance depends not only on the choice of classifier but also on the quality, consistency, and dimensionality of the underlying data.

Recent studies have emphasized the importance of comprehensive model evaluation using multiple performance metrics rather than relying solely on classification accuracy [3]. This broader perspective highlights the need to examine not only classifier design but also the contribution of data engineering processes, including feature selection, dimensionality reduction, data cleansing, and label-noise detection.

Therefore, systematic data engineering may represent an important yet comparatively underexplored direction for improving the quality, reproducibility, and practical utility of clinical datasets used in artificial intelligence research. The present study investigates this possibility by transforming the Cleveland Heart Disease Dataset into an AI-ready resource through feature reduction, label-noise detection, and dataset cleansing.

1.2 Related Work

1.2.1 Classifier-Centered Studies

The Cleveland Heart Disease Dataset has been extensively investigated as a benchmark for cardiovascular risk prediction. Most published studies have focused on improving classification performance through algorithmic innovations. For example, gradient boosting models achieved an accuracy of approximately 86% [1], ensemble-learning approaches reported results close to 93% [4], neural-network-based methods reached accuracies between 90% and 94% [5,8,9], while optimization-assisted feature-selection approaches combined with support vector machines produced comparable performance [6].

These studies demonstrate the sustained interest of the research community in classifier development and optimization. However, the reported performance improvements are typically achieved through modifications of learning algorithms, ensemble strategies, or hyperparameter tuning rather than through systematic improvement of the underlying dataset.

1.2.2 Dataset Limitations

In parallel with advances in classifier design, several limitations of the Cleveland dataset have been reported in the literature. The review presented in [7] identifies concerns related to the relatively small number of records, demographic imbalance, and the presence of inaccurate observations. The same review notes that only 282 of the 303 records were verified as accurate, suggesting that data quality issues may affect predictive performance.

These issues also affect the readiness of healthcare datasets for reproducible AI development and clinical decision-support applications.

1.2.3 Research Gap and Motivation

Despite extensive efforts devoted to classifier optimization, comparatively little attention has been paid to systematic data engineering of the Cleveland dataset. In particular, the combined application of feature reduction, label-noise identification, and dataset cleansing has received substantially less attention than the development of new predictive models.

This imbalance is noteworthy because improvements in data quality may potentially influence predictive performance as strongly as, or even more strongly than, further modifications of classification algorithms. Therefore, the Cleveland dataset provides a suitable case study for investigating how systematic data engineering can transform a widely used clinical dataset into an AI-ready resource suitable for robust, reproducible, and clinically relevant AI research .

1.3 Study Contributions

This study makes three principal contributions to the analysis of the Cleveland Heart Disease Dataset.

1. An ensemble-based feature-selection strategy combining eight independent attribute evaluators is applied to identify the most informative clinical variables and reduce dataset dimensionality from 13 to 7 attributes.
2. A data-cleansing procedure based on the CAIRAD label-noise detection algorithm is employed to identify and remove potentially mislabeled records, resulting in a reduced and cleansed AI-ready resource containing 254 instances.
3. Through extensive experimental validation using three neural-network-based classifiers and repeated cross-validation, we demonstrate that data engineering improvements produce statistically significant gains across multiple performance metrics. Moreover, the observed improvements exceed the performance differences among the evaluated classifiers, highlighting the critical role of dataset quality in clinical AI.

.This work therefore demonstrates how feature reduction and label-noise removal can transform the Cleveland Heart Disease Dataset into an AI-ready resource for reproducible clinical AI and health-information research.

2 DATASET AND DATA ENGINEERING PIPELINE

2.1 Original Dataset

The study was based on the Cleveland Heart Disease Dataset obtained from the UCI Machine Learning Repository [2]. Since its publication, this dataset has become one of the most widely used benchmark resources for cardiovascular risk prediction and clinical AI research.

The dataset contains 303 patient records described by 13 predictive attributes and one binary target variable indicating the estimated risk of heart disease. The attributes comprise both numerical and categorical clinical variables, including demographic information, physiological measurements, electrocardiographic findings, exercise-test results, and established cardiovascular risk indicators. A detailed description of the variables is presented in Table 1.

Among the 303 instances, 138 belong to the low-risk class and 165 to the high-risk class. The dataset contains eight categorical and five numerical attributes with substantially different scales and value ranges, making preprocessing an important preliminary step before machine-learning analysis.

Table 1. Description of the Cleveland dataset [2]

ID	Attribute name	Type	Description	Unit	Range
1	age	integer	Age of patients	years	29 - 77
2	sex	nominal	Gender		0, 1 = female, male
3	cp	nominal	Chest pain type		0 = Typical Angina, 1 = Atypical Angina, 2 = Non-anginal Pain, 3 = Asymptomatic
4	trtbps	real	Resting blood pressure	mmHg	94 -200

5	chol	real	Cholesterol via a BMI sensor	mg/dl	126 - 564
6	fbg	nominal	fasting blood sugar > 120 mg/dl		0, 1 = false, true
7	restecg	nominal	Resting ECG results		0, 1,3 = Normal, ST,?
8	thalachh	integer	Maximum heart rate achieved	bpm	71 - 202
9	oldpeak	real	ST depression induced by exercise relative to rest	?	0 – 6.2
10	slp	nominal	the slope of the peak exercise ST segment		0,1,2 = Up, Flat, Down
11	caa	integer	Number of major vessels colored by fluoroscopy		0 - 4
12	thall	nominal	Thallium Stress Test result		0 - 3
13	exng	nominal	Exercise-induced angina		0,1 = No, Yes
14	output	nominal	Target variable		0 or 1 = less chance of heart attack, or more chance of heart attack

Several limitations of the dataset have previously been reported in the literature [7]. These include the relatively small sample size, demographic imbalance, and the possible presence of inaccurate records. Such characteristics make the dataset an appropriate case study for investigating the effect of data engineering techniques aimed at improving data quality and predictive performance.

The Cleveland dataset served as the starting point for the data engineering pipeline proposed in this study. The subsequent stages involved feature selection, label-noise identification, dataset cleansing, and the construction of a reduced AI-ready dataset suitable for experimental validation.

2.2 Ensemble Feature Selection

Feature selection is a fundamental data engineering procedure aimed at reducing dataset dimensionality while preserving the information relevant for predictive modeling [12]. In clinical datasets, redundant or weakly informative variables may increase model complexity, reduce interpretability, and introduce additional sources of variability. Therefore, identifying a compact subset of informative features represents an important step in constructing AI-ready datasets.

To mitigate the inherent bias of individual ranking methods, an ensemble feature-selection strategy was adopted. Eight independent attribute evaluators available in WEKA were applied to the standardized dataset: Correlation Ranking Filter, Gain Ratio Attribute Evaluation, Information Gain Ranking Filter, OneRule Attribute Evaluation, Relief Ranking Filter, PairwiseCorrelationAttributeEval, Symmetrical Uncertainty Ranking Filter, and J48 Attribute Evaluation. The resulting rankings for all 13 predictive attributes are presented in Table 2.

Table 2. Attributes (features) ranking

No	WEKA evaluator	Attribute ranks (best first)
1	Correlation Ranking Filter	13,9,10,8,3,12,11,2,1,7,4,5,6
2	Gain Ratio feature evaluator	13,9,8,3,12,10,11,2,1,7,6,4,5
3	Information Gain Ranking Filter	13,3,12,10,9,8,11,1,2,7,6,5,4
4	OneRule feature evaluator	13,3,12,9,11,10,8,1,2,7,4,5,6

5	RELIEF Ranking Filter	12,3,13,2,11,9,7,10,8,1,6,4,5
6	PairwiseCorrelationAttributeEval	3,3,9,12,8,10,11,2,1,7,6,5,4
7	Symmetrical Uncertainty Ranking Filter	13,3,9,12,8,10,11,2,1,7,6,4,5
8	J48 feature evaluator	13,3,12,9,8,11,10,1,2,7,6,5,4

For each attribute a_i , eight rank values were obtained from the eight evaluators:

$$R_i = \{r_{i1}, r_{i2}, \dots, r_{i8}\},$$

where r_{ij} denotes the rank assigned to attribute a_i by evaluator j .

The ensemble rank of attribute a_i was defined as the median rank

$$\tilde{r}_i = \text{median}(R_i).$$

To quantify the agreement among evaluators, the median absolute deviation (MAD) was computed as

$$\text{MAD}_i = \text{median}\{[r_{ij} - \tilde{r}_i] : j = 1, \dots, 8\}$$

Attributes were subsequently ordered according to increasing values of $\tilde{r}_i$, with lower median ranks indicating higher estimated importance. The corresponding MAD values were used as a robust measure of ranking stability across evaluators. Small MAD values indicate strong agreement among evaluators, whereas larger values reflect greater variability in the assigned ranks. Since ranks are ordinal quantities, median ranks and MAD were preferred over the mean and standard deviation because they provide more robust estimates in the presence of evaluator disagreement. .

The ranking results are summarized in Table 2 . The ranking results revealed a consistent separation of variables into groups of high-, intermediate-, and low-importance features. The attribute-ranking matrix is visualized in Figure 1.

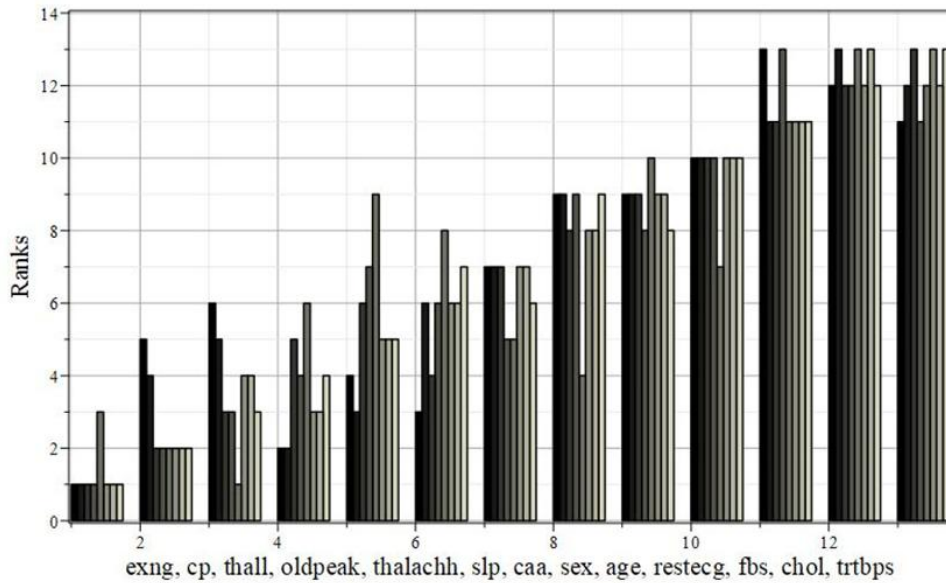


Figure 1. Feature-ranking matrix obtained from eight independent WEKA attribute evaluators. Each column corresponds to one dataset attribute, while each row represents an individual ranking method. Attributes are ordered according to their median ensemble ranks.

Table 2 shows that the attribute exng (exercise-induced angina) most frequently occupied the highest-ranking positions across the evaluated methods. In contrast, trtbps, chol, and fbs consistently appeared among the lowest-ranked attributes.

The matrix of ranks shown in Figure 1 was further analyzed to identify similarities among attributes. Because rank values form ordinal data, Manhattan distances were calculated between all pairs of attribute-ranking vectors, resulting in a symmetric 13×13 distance matrix. Manhattan distance was used because it directly measures cumulative rank differences between attributes. The corresponding heatmap is presented in Figure 2.

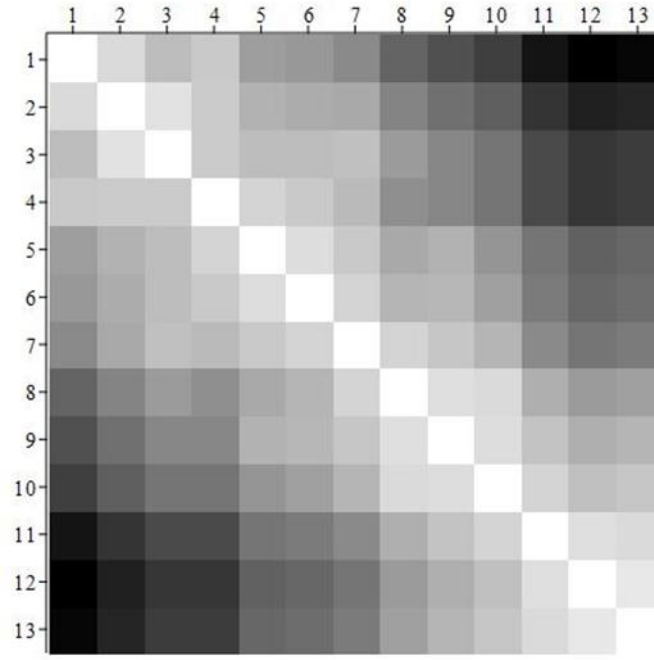


Figure 2. Heatmap of Manhattan distances derived from the feature-ranking matrix, showing natural groupings of attributes according to their ensemble ranks.

The heatmap revealed a natural grouping of attributes into four subsets:

1. Highest-ranked subset: $\langle exng, cp, thall, oldpeak \rangle$ (median ranks 1-4);
2. Upper intermediate subset: $\langle thalachh, slp, caa \rangle$ (median ranks 5-7);
3. Lower intermediate subset: $\langle sex, age, restecg \rangle$ (median ranks 8-10);
4. Lowest-ranked subset: $\langle fbs, chal, trtbps \rangle$ (median ranks 11-13).

Although minor permutations within individual groups are possible because of overlapping MAD values, the separation between these subsets proved stable across evaluators.

Based on the ensemble analysis, the lower intermediate and lowest-ranked subsets were excluded from further consideration. The resulting reduced dataset retained seven predictive attributes (Table 1):

$\{exng, cp, thall, oldpeak, thalachh, slp, caa\}$.

Notably, this subset coincides with the feature subset independently suggested by the WEKA CfsSubsetEval procedure with the BestFirst search strategy, providing additional support for its relevance.

The resulting seven-variable representation preserves the most informative clinical attributes while reducing dataset dimensionality by almost twice: from 13 to 7 predictive variables. This reduced feature space served as the basis for the subsequent label-noise identification and dataset-cleansing stage of the proposed data engineering pipeline.

2.3 Label-Noise Detection and Dataset Cleansing

The medical datasets that are relatively old, like the one we are looking at here, are often quite noisy. That means they may contain outliers or extreme values for specific attributes, record inconsistencies, and even incorrect target variable values (misabeled instances, class noise), which can be especially damaging in machine learning. Instances may be mislabeled when assigned the wrong class label [13]. If one considers medical diagnostics, different classes might have similar symptoms. In contrast to class noise, attribute noise reflects erroneous values for one or more attributes [13].

To identify mislabeled data (class noise), we used the "co-appearance-based analysis for incorrect records and attribute-value detection (CAIRAD)" [14] algorithm. This algorithm has three tuning hyperparameters: 1) L - co-appearance score threshold (default 0.3), 2) T - co-appearance threshold (default 0.8), and 3) M (convert detected noise into missing values), which was set to "true."

CAIRAD detected 49 mislabeled instances out of 303 concerning the target variable in the dataset [2]. These instances were removed, so the reduced dataset comprises only 254. Of course, the records of the reduced dataset still contain missing values (from 5% to 12 %, depending on the attribute), which appeared in the places of noised attributes but not for the target variable after CAIRAD filtering.

There are three known ways of dealing with missing values: 1) removing instances, 2) imputing with means, medians, modes, or other methods, and 3) disregarding if the classifiers can work with missing values. The authors preferred the third method for the following reasons:

- The relatively small number of instances did not allow for the first method, as it may result in insufficient data;
- Imputation is a subjective process that introduces initially absent values into the data.
- The deep learning algorithms of WEKA, which were mentioned earlier, work well with missing data. We believe that missing data has less impact on their performance than noisy data.

2.4 Resulting AI-Ready Dataset

The combined application of ensemble feature selection and label-noise detection resulted in a reduced and cleansed version of the Cleveland Heart Disease Dataset. The resulting dataset contains 254 instances and 7 predictive attributes selected from the original set of 13 variables. A comparison of the original dataset (*ds0*) and the resulting dataset (*ds1*) is presented in Table 3.

Table 3. Comparison of the original and engineered datasets

Dataset	Instance number	Attributes number	Classes	Missing values
ds0	303	13	138 - low risk 165 - high risk	absent
ds1	254	7	104 - low risk 150 - high risk	present

Compared with the original dataset, *ds1* exhibits substantially lower dimensionality and excludes records identified as potentially mislabeled by the CAIRAD procedure. Although the cleansing process introduced some missing attribute values, the resulting dataset preserves the most informative clinical variables while reducing feature redundancy and class-label uncertainty.

In the context of this study, the term AI-ready dataset refers to a dataset that has undergone systematic preprocessing, feature reduction, and data-quality improvement before model development. The constructed dataset *ds1* satisfies these requirements and therefore represents an AI-ready version of the Cleveland Heart Disease Dataset.

The engineered dataset *ds1* served as the input for all subsequent validation experiments described in Section 3.

3 EXPERIMENTAL VALIDATION

3.1 Classifiers and Experimental Setup

To evaluate the impact of the proposed data engineering pipeline, experimental validation was performed using three neural-network-based classifiers available in WEKA. The comparison involved both the original dataset (*ds0*) and the resulting AI-ready dataset (*ds1*) under identical validation conditions.

The experiment was set to "Simple" configuration mode, and the type of experiment was "Classification" with 10-fold cross-validation. Two sets of data, which were described in the previous section, were used. Three DL algorithms of classification from WEKA were used. The iterative

control process included 25 replicates for each dataset and classifier and ten cross-validations. This resulted in 250 classifications for six possible pairs comprising two datasets and three classifiers. This approach allowed for estimating performance indicators, calculating their standard deviations, and statistically testing hypotheses. The "Data sets first" option was selected (Fig.3).

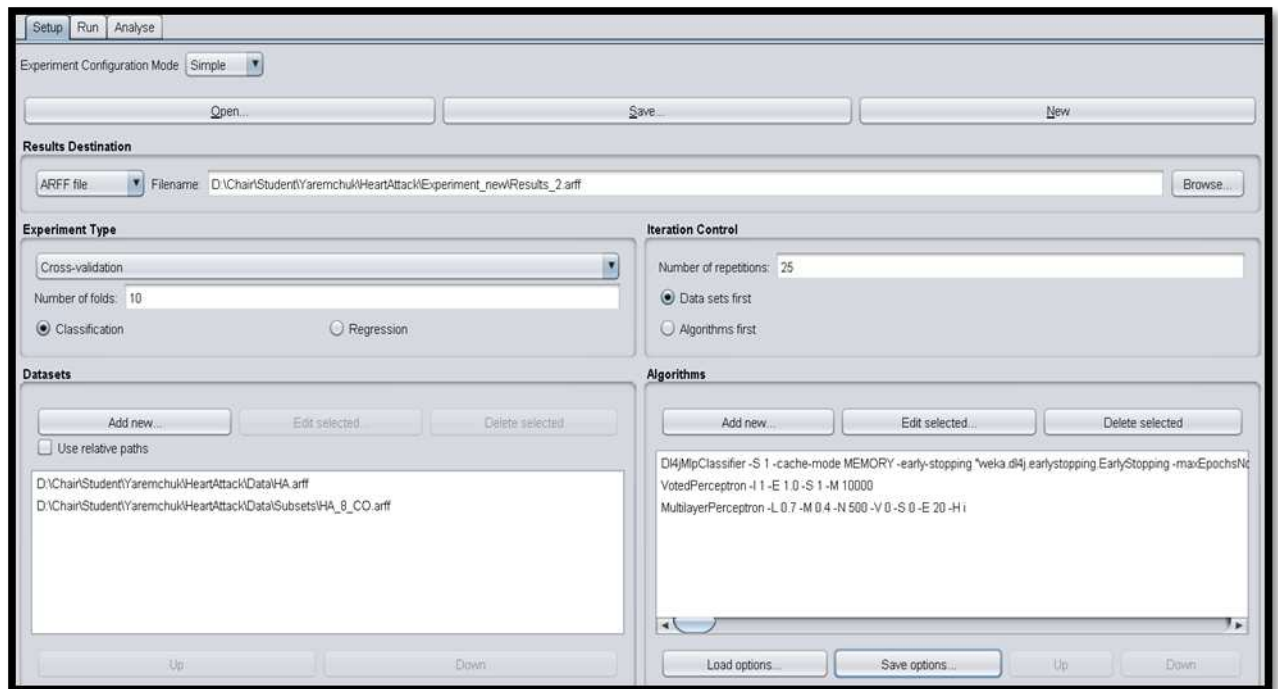


Figure 3. Experimental setup implemented in WEKA for comparative evaluation of the original dataset (*ds0*) and the AI-ready dataset (*ds1*).

Figure 3 presents three WEKA classifiers based on neural networks that utilize deep learning methods. These classifiers are D14jMlp, Voted Perceptron (VP), and Multi-layer Perceptron (MLP). The first two were used with default hyperparameters without any tuning. However, for MLP, we modified three default parameters:

- L (learning rate) was changed to 0.7 from the default value of 0.3;
- M (momentum) was set to 0.4 instead of 0.2;
- H (hidden layers) was adjusted to "i" from "a."

In the above context, "i" signifies that the number of hidden layers corresponds to the number of attributes. At the same time, default "a" means a semi-sum of the number of attributes and classes.

The Experimenter can save, repeat, or modify the experiment's options using an "*.EXP" file. The experiment's average time is about 30 minutes on a standard computer. All results were saved in an ARFF-type file (refer to Fig. 3). Every performance index for the six paired datasets and classifiers was computed 250 times since we chose 10-fold cross-validation and 25 times repeating. That permitted testing the statistical hypotheses and estimating the standard deviations of these indices.

3.2 Statistical Evaluation

The predictive performance of the evaluated models was assessed using multiple complementary metrics, including Accuracy, Precision, Recall, F1-score, Matthews Correlation Coefficient (MCC), and Precision-Recall Area Under Curve (PRC-AUC). The use of multiple metrics was motivated by the moderate class imbalance present in both datasets and the limitations of relying exclusively on classification accuracy [3].

For each classifier-dataset combination, performance indicators were estimated using 10-fold cross-validation repeated 25 times. This procedure generated 250 performance estimates for each evaluated metric, allowing the calculation of average values and variability measures.

Statistical significance was assessed using the corrected paired Student's t-test implemented in WEKA [15]. The null hypothesis stated that no significant difference exists between the predictive

performance obtained with the original dataset (*ds0*) and the engineered dataset (*ds1*). The alternative hypothesis stated that the observed differences are statistically significant.

The significance level was fixed at $\alpha = 0.05$. Consequently, differences with $p \leq 0.05$ were considered statistically significant.

4 RESULTS AND DISCUSSION

4.1 Performance Improvement Through Data Engineering

The primary outcome of the study is the successful transformation of the Cleveland Heart Disease Dataset into an AI-ready resource through systematic data engineering. The observed improvements in predictive performance provide quantitative evidence of the effectiveness of this transformation.. Table 4 compares the performance of three classifiers on the original dataset (*ds0*) and the resulting AI-ready dataset (*ds1*).

We rejected the null hypothesis in the experiment and found support for the alternative hypothesis across all performance indices. The results showed that the reduced dataset (*ds1*) has better predictive power than the initial dataset (*ds0*), which is statistically significant. To support this conclusion, we included four key performance indicators - Accuracy, F1-score, Matthews correlation coefficient (MCC), and Precision-Recall Curve area (PRC AUC) - in Table 4. These indicators are particularly relevant for imbalanced datasets.

Table 4. Performance indicators comparison of two datasets for three classifiers

Classifier	Accuracy		F1-score		MCC		PRC AUC	
	ds0	ds1	ds0	ds1	ds0	ds1	ds0	ds1
DL4jMlp	84.6	95.1	0.83	0.94	0.70	0.90	0.91	0.99
VP	83.2	93.9	0.81	0.93	0.67	0.88	0.85	0.95
MLP	81.5	97.8	0.89	0.97	0.63	0.96	0.90	1.00

The results indicate that systematic data engineering improved the quality and utility of the underlying clinical data resource. The resulting AI-ready resource consistently outperformed the original dataset across all evaluated metrics while requiring fewer attributes and containing fewer potentially erroneous records. The substantial growth of the Matthews correlation coefficient (MCC) is impressive. The MCC test is simple and intuitive. The classifier must accurately predict the majority of negative and positive cases, regardless of their ratio (balance) in the overall data set, to obtain a high-quality score closer to one [16].

The Area Under the Curve of Precision and Recall (PRC-AUC) for a perfect classifier equals 1. In a "baseline" classifier, the AUC-PR depends on the ratio of instances belonging to the "positive" class. For example, the "baseline" PRC-AUC = 0.59 for the class with a high risk of *ds1* and 0.54 for *ds0* (see Table 3,4). A classifier with real predictive value will fall between the "baseline" and perfect classifiers. Table 4 shows that the DL classifiers of WEKA look nearly perfect if dealing with the reduced dataset (*ds1*). In contrast to MCC and Kappa statistics, this indicator is sensitive to the class being considered.

The simple tuning of hyperparameters made to the MLP algorithm, which can be likened to "make-up" and was described in the previous section, along with the data in Table 4, may lead one to think that the MLP is the top performer among these three classifiers. However, this would be wishful thinking. The issue is that the null hypothesis stands, indicating no significant disparity in performance among these classifiers. Therefore, the matching variations in performance indicators in Table 4 are not actual and statistically insignificant, contrasting the disparity between datasets. In other words, each of the three classifiers performs similarly to the other two.

If one wants to see performance indicators for both classes (0 for low risk, 1 for high risk), using the Explorer or Workbench modules of WEKA is handier. Table 4 shows the classification results for both datasets (*ds0*, *ds1*) and each of the two classes using the MLP classifier. It is worth mentioning that this classifier may not be the best option, as we found that all three deep learning algorithms are equivalent in this regard.

Table 5. Performance indicators comparison of two datasets for MLP classifier

Dataset	Confusion matrix	Precision	Recall	F1-score	MCC	PRC AUC	Class
ds0	$\begin{pmatrix} 111 & 27 \\ 25 & 140 \end{pmatrix}$	0.816	0.804	0.810	0.654	0.877	0
		0.838	0.848	0.843	0.654	0.877	1
ds1	$\begin{pmatrix} 101 & 3 \\ 4 & 146 \end{pmatrix}$	0.962	0.971	0.967	0.943	0.996	0
		0.980	0.973	0.977	0.943	0.998	1

The slight differences between Tables 4 and 5 are lucid, resulting from the averaging in Table 3. However, both tables demonstrate that the reduced dataset (ds1) outperforms the full one (ds0) across all indices. Table 4 is also informative as it illustrates the direct impact of imbalances between classes: all performance indicators for the smaller class (those with low risk of heart attack) are slightly lower than for the larger one (high risk). Imbalanced classes, typically binary, are standard in medical diagnostics. They require performance indicators that are dimly sensitive to imbalance (MCC, Kappa statistics, PRC AUC) if the imbalance is relatively tiny, as in this case, or even special corrections of instance weights otherwise.

Let us delve deeper into hyperparameter tuning. This tuning may improve performance by two, or three percent compared to the already promising results of the smaller dataset (ds1). However, this improvement does not surpass the statistical errors estimated in the experiment and, therefore, cannot be deemed significant. Thus, it may not be worth the effort.

4.2 Transforming Primary Data into an AI-Ready Resource

It is instructive to compare the performance of the proposed data-engineering approach with the best ensemble-learning results reported in the literature [4]. In that study, the dataset was partitioned into smaller subsets using a mean-based splitting strategy and subsequently modeled using a Classification and Regression Tree (CART). From a health-informatics perspective, these findings are important because they demonstrate that improving the quality of a clinical data resource may be as valuable as increasing model complexity.

Table 6. Comparison of performance indicators between the proposed approach and the ensemble-learning technique of [4] (values in parentheses denote standard deviations when available)

	Accuracy	Precision	Recall	F1-score
Our results, MLP	0.98 (0.03)	0.97 (0.05)	0.98 (0.05)	0.97 (0.04)
Results [4], CART	0.93	0.96	0.91	0.93

The results presented in Table 6 indicate that both approaches achieve comparable predictive performance. Given the reported variability of the performance metrics, the observed differences are unlikely to be statistically significant. Consequently, both methods demonstrate similar diagnostic capabilities.

However, an important distinction should be emphasized. The ensemble-learning approach of [4] achieves competitive results by increasing model complexity, whereas the present study relies primarily on data engineering through feature reduction and label-noise removal. As a result, the proposed workflow produces a smaller and more compact dataset while maintaining predictive performance comparable to that of a more complex ensemble model.

More importantly, the observed performance gains associated with data engineering demonstrate that substantial improvements can be achieved through systematic enhancement of clinical data resources. The resulting AI-ready resource provides a more reliable foundation for future clinical AI studies than the original dataset.. This finding suggests that, for the Cleveland Heart Disease Dataset, improvements in data quality have a greater impact on predictive performance than further optimization of classifier architecture. The result supports a data-centric perspective in which systematic dataset engineering complements, and in some cases outweighs classifier optimization.

4.3 AI-Ready Resources for Reproducible Clinical AI

The results obtained suggest that improvements in clinical AI systems should not be sought exclusively through the development of increasingly sophisticated prediction models. In health information systems, data quality, data availability, and reproducibility are fundamental requirements for trustworthy AI deployment. The present study addresses these requirements through a transparent and reproducible data-engineering workflow. In the Cleveland dataset, systematic data engineering produced substantially larger improvements than changes in classifier architecture.

From a practical perspective, feature reduction decreases data acquisition requirements, while label-noise removal improves the reliability of predictive models. Both aspects are important for real-world clinical deployment. These findings support a data-centric approach to clinical AI development, where dataset quality is considered as important as algorithm design.

The combined application of feature selection and dataset cleansing resulted in an AI-ready resource suitable for reproducible experimentation, methodological benchmarking, and future development of clinical decision-support systems. The resulting dataset contains fewer variables, excludes records identified as potentially mislabeled, and demonstrates improved predictive performance while maintaining clinical relevance.

To facilitate reproducibility, the AI-ready dataset *dsI* is publicly available through the Mendeley Data repository [17].

Because all preprocessing, feature-selection, and validation procedures are fully described and rely on publicly available tools, the proposed workflow can be independently reproduced and further extended by other researchers. Beyond heart disease prediction, the proposed workflow may serve as a general framework for preparing legacy healthcare datasets for contemporary AI applications. This perspective shifts attention from model-centered optimization toward the engineering of reliable and reusable clinical data resources.

5 CONCLUSIONS

This study presented a systematic data-engineering workflow for transforming the Cleveland Heart Disease Dataset from primary clinical data into an AI-ready resource. The proposed approach combined ensemble feature selection and label-noise detection, reducing the number of predictive attributes from 13 to 7 and identifying 49 potentially mislabeled records.

The resulting AI-ready resource demonstrated consistently superior predictive performance compared with the original dataset across all evaluated metrics, including Accuracy, Precision, Recall, F1-score, Matthews Correlation Coefficient, and Precision-Recall AUC. The observed improvements were statistically significant for all evaluated classifiers.

Beyond predictive performance, the study highlights the importance of data quality as a fundamental component of clinical AI development. The results show that substantial gains can be achieved through systematic improvement of healthcare data resources before additional model optimization is considered. Feature reduction decreased data complexity, while label-noise removal improved the reliability of the underlying information resource.

The transformed Cleveland dataset provides a compact, reproducible, and publicly accessible AI-ready resource for future research in cardiovascular risk prediction, clinical decision support, and health information systems. More broadly, the proposed workflow may serve as a practical framework for preparing legacy healthcare datasets for contemporary AI applications, emphasizing the role of data engineering in the development of trustworthy and reusable clinical data resources..

Data Availability

The AI-ready resource developed in this study is publicly available through the Mendeley Data repository [17]. The original Cleveland Heart Disease Dataset is available from the UCI Machine Learning Repository [2].

Funding

The authors received no specific funding for this work.

Competing Interests

The authors declare that they have no competing interests.

Authors' Contributions

G.P. Chuiko conceived the study, designed the methodology, performed the data analysis, and drafted the manuscript. O.M. Yaremchuk contributed to the clinical interpretation of the results, validation of the study design, visualizing and manuscript revision. Both authors read and approved the final manuscript.

Ethics Approval and Consent to Participate

This study used a publicly available de-identified dataset and did not involve direct participation of human subjects. Therefore, ethics approval and informed consent were not required.

Consent for Publication

Not applicable.

Declaration of Generative AI and AI-assisted technologies in the writing process

While preparing this manuscript, the authors used Grammarly for language editing and proofreading. The authors reviewed and edited the content as necessary and take full responsibility for the final version of the manuscript..

REFERENCES

- [1] Gupta, S.K., Shrivastava, A., Upadhyay, S.P., and Chaurasia, P.K. (2021). A Machine Learning Approach for Heart Attack Prediction. *International Journal of Engineering and Advanced Technology*.10,(6). <https://doi.org/10.35940/ijeat.f3043.0810621>
- [2] Janosi A., William S., Pfisterer M., and Detrano R., (1988). Heart Disease. UCI Machine Learning Repository. <https://doi.org/10.24432/C52P4X>
- [3] Naidu G., ZuvaT., Sibanda EM. (2023). A Review of Evaluation Metrics in Machine Learning Algorithms. In: Silhavy R., Silhavy P. (eds) *Artificial Intelligence Application in Networks and Systems*. CSOC 2023. *Lecture Notes in Networks and Systems*, vol 724. Springer, Cham. https://doi.org/10.1007/978-3-031-35314-7_2
- [4] Mienye ID, Sun Y, Wang Z. An improved ensemble learning approach for the prediction of heart disease risk. *Informatics Med Unlocked*. 2020;20:100402. <https://doi.org/10.1016/j.imu.2020.100402>
- [5] Oumaima Terrada, Soufiane Hamida, Bouchaib Cherradi, Abdelhadi Raihani, Omar Bouattan. 2020. Supervised Machine Learning Based Medical Diagnosis Support System for Prediction of Patients with Heart Disease. *Advances in Science, Technology and Engineering Systems Journal*. 5. 269-277; <http://dx.doi.org/10.25046/aj050533>
- [6] Al-Tashi Q., Rais H., Jadid S. (2019). Feature Selection Method Based on Grey Wolf Optimization for Coronary Artery Disease Classification. In: Saeed F., Gazem N., Mohammed F., Busalima. (eds) *Recent Trends in Data Science and Soft Computing*. IRICT 2018. *Advances in Intelligent Systems and Computing*, vol 843. Springer, Cham. https://doi.org/10.1007/978-3-319-99007-1_25
- [7] Ruqiya, Muhammad Tayyab Yaqoob, Abid Muhammad Khan, Aftab Ahmed Shaikh1, Noman Khan. Review on Cleveland Heart Disease Dataset using Machine Learning. *Quaid-e-Awam Univ Res J Eng Sci Technol*. 2023;21:87-98. <http://dx.doi.org/10.52584/QRJ.2101.11>
- [8] Fatima Fayeza, Jaiswal Arunima, and Sachdeva Nitin, Heart Disease Prediction Using Supervised Classifiers (2022). *Proceedings of the International Conference on Innovative Computing & Communication (ICICC) 2022*, Available at SSRN: <https://ssrn.com/abstract=4121817> or <http://dx.doi.org/10.2139/ssrn.4121817>
- [9] Akella A, Akella S. Machine learning algorithms for predicting coronary artery disease: efforts toward an open source solution. *Future Sci OA*, 2021;7(6); <https://doi.org/10.2144%2Ffsoa-2020-0206>

- [10] Desai F, Chowdhury D, Kaur R, et al. HealthCloud: A system for monitoring health status of heart patients using machine learning and cloud computing. Internet of Things. 2022;17:100485. <https://doi.org/10.1016/j.iot.2021.100485>
- [11] Bouckaert RR, Frank E, Kirkby R, Reutemann P, Seewald A, Scuse D. WEKA Manual for Version 3-9-5. The University of Waikato; 2020. <https://sourceforge.net/projects/weka/>
- [12] Cai J, Luo J, Wang S, Yang S. Feature selection in machine learning: A new perspective. Neurocomputing. 2018;300:70-79. <https://doi.org/10.1016/j.neucom.2017.11.077>
- [13] Gupta S, Gupta A. Dealing with Noise Problem in Machine Learning Datasets: A Systematic Review. Procedia Comput Sci. 2019;161:466-474. <https://doi.org/10.1016/j.procs.2019.11.146>
- [14] Rahman MG, Islam MZ, Bossomaier T, Gao J. CAIRAD: A co-appearance-based analysis for incorrect records and attribute-values detection. The 2012 International Joint Conference on Neural Networks (IJCNN), Brisbane, QLD, Australia, 2012, pp. 1-10, <https://doi.org/10.1109/IJCNN.2012.6252669>
- [15] Paired Student's T-Test BT - The Concise Encyclopedia of Statistics. In: New York, NY: Springer New York; 2008:409-410. https://doi.org/10.1007/978-0-387-32833-1_304
- [16] Chicco D, Jurman G. The advantages of the Matthews correlation coefficient (MCC) over F1 score and accuracy in binary classification evaluation. BMC Genomics 21, 6 (2020). <https://doi.org/10.1186/s12864-019-6413-7>
- [17] Chuiko G., Yaremchuk O., "Heart attack risk prediction: a more productive, reduced, and cleansed ARFF dataset for machine learning," Mendeley Data, V1, 2024, <https://data.mendeley.com/datasets/n6293b4ggc/1>