MethodsX 15 (2025) 103510

Contents lists available at ScienceDirect

MethodsX

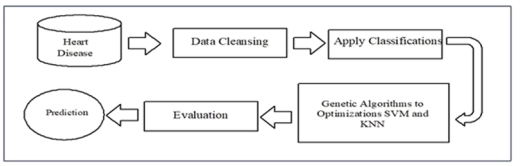
journal homepage: www.elsevier.com/locate/methodsx

Hybrid genetic algorithms-driven optimization of machine learning models for heart disease prediction

Sherko H. Murad [a,*], Noor Bahjat Tayfor [b], Nozad H. Mahmood [c], Lawson Arman [d]

[a] Computer Science Department, Cihan University of Sulaimaniya, Sulaymaniyah, Kurdistan, Iraq
[b] Institute of Informatic, Faculty of Science and Informatics, University of Szeged, Szeged, Hungary
[c] Cihan University Sulaimaniya Research Center (CUSRC), Cihan University Sulaimaniya, Sulaymaniyah City, Kurdistan Region, Iraq
[d] Department of facility Therapy named after professor. V.A. Valdman, St Peterburg state Pediatric Medical university, Saint Petersburg, Russia

GRAPHICAL ABSTRACT



ARTICLE INFO

Keywords:
Cardiovascular disease prediction
Machine learning (ML)
Support vector machine (SVM)
K-nearest neighbour (KNN)
Genetic algorithm (GA)

ABSTRACT

Machine learning (ML) models, such as K-Nearest Neighbor (KNN) and Support Vector Machine (SVM), play a vital role in predicting heart disease. However, their performance is often limited by poor hyperparameter selection. This study presents a novel hybrid approach that uses a Genetic Algorithm (GA) to systematically optimize the hyperparameters of KNN and SVM models, leading to improved classification outcomes. The optimization driven by the GA resulted in significant performance improvements, increasing the classification accuracy of KNN to 95.38 % and SVM to 90 %. Furthermore, there were significant improvements in precision, recall, and F-score. Our findings demonstrate that GA-based hyperparameter tuning is an effective strategy for improving the predictive power and clinical relevance of ML models used for heart disease classification.

- **GA Driven Optimization:** Genetic algorithm was utilized to fine-tune the hyperparameters of K-Nearest Neighbour (KNN) and Support Vector Machine (SVM) classifiers for improved performance.

Related research article: None.**For a published article**: None.

* Corresponding author.
E-mail address: sherko.murad@sulicihan.edu.krd (S.H. Murad).

https://doi.org/10.1016/j.mex.2025.103510
Received 28 March 2025; Accepted 14 July 2025
Available online 16 July 2025

- **Significant Performance Gains:** The optimization process aimed to maximize classification across metrics such as accuracy, precision, recall, and F-score.
- **Improved Accuracy:** The GA-based tuning significantly increased the classification accuracy, improving KNN to 95.38 % and SVM to 90 %.

Specifications table

Subject area	Computer Science
More specific subject area	Machine Learning
Name of your method	Genetic SVM, Genetic KNN, Cardiovascular Disease Prediction
Name and reference of original method	Machine Learning
Resource availability	The dataset consists of 1190 records of patients from US, UK, Switzerland and Hungary https://www.kaggle.com/datasets/sid321axn/heart-statlog-cleveland-hungary-final Python

Background

Cardiovascular diseases (CVDs), including coronary heart disease and stroke, are the leading causes of morbidity and mortality worldwide [1]. Their prevalence has surged from 271 million in 1990 to 523 million in 2019, with related deaths increasing from 12.1 million to 18.6 million in the same period. According to the World Health Organization (WHO), CVDs account for 32 % of global deaths, totaling around 20.5 million annually. Without significant advancements in management, CVD-related mortality could surpass 24.2 million by 2030, with myocardial infarction and stroke being the primary contributors [16,18]. Artificial Intelligence (AI) offers promising solutions to enhance the detection and management of CVD. Machine learning (ML), a subdomain of AI, can analyze vast datasets and detect complex patterns indicative of heart abnormalities [2]. Machine learning (ML) models have been widely applied to cardiovascular disease (CVD) prediction, with performance measured through various metrics [14]. Among these, k-nearest neighbors (KNN) and support vector machines (SVM) are commonly used, each with its strengths and limitations. KNN, though computationally simpler, often delivers lower accuracy unless improved with kernel functions or distance-weighted techniques. SVM, while highly accurate, is computationally intensive and struggles with high-dimensional data, such as gene expression datasets [6,15]. Both KNN and SVM face challenges related to data complexity and dimensionality, which can lead to reduced accuracy. Feature selection techniques help enhance model performance by eliminating irrelevant variables [8]. Additionally, optimizing ML models through parameter tuning is crucial for achieving better accuracy. Genetic Algorithm (GA), inspired by biological evolution, serves as a powerful optimization technique for complex machine learning models [17]. They operate through iterative selection, crossover, and mutation processes to refine solutions, making them effective for both constrained and unconstrained optimization problems [10]. In the study [22,24], the authors propose a novel hybrid model that integrates a Hybrid Deep Belief Network (HDBN) with a Context-Aware Edge Network (CAEN) to enhance disease prediction accuracy in healthcare systems. The framework was evaluated using four diverse datasets. It demonstrated strong performance, achieving an accuracy of 93 %, a precision of 87 %, a specificity of 95 %, and a recall of 91 %, indicating its effectiveness in processing large-scale medical data with improved predictive capability. In their study [25], the authors present a hybrid feature selection method that combines Ant Colony Optimization (ACO) and Ant Lion Optimization (ALO) to enhance leukemia prediction accuracy. Applied to microarray gene expression data, this hybrid ACO-ALO approach achieved a classification accuracy of 93.94 %, outperforming traditional ACO and ALO methods, which attained accuracies of 93.94 % and 90.91 %, respectively. GA has proven successful in various domains, including structural optimization, by efficiently searching for optimal design solutions [5]. Their integration with machine learning (ML) can enhance model efficiency by fine-tuning parameters [13] for improved predictive performance. This study aims to leverage GA for optimizing KNN and SVM classifiers in heart disease prediction, proposing a novel approach that enhances accuracy and reliability.

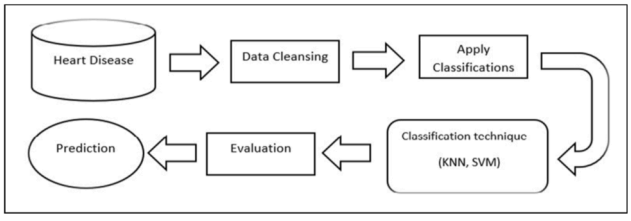


Fig. 1. Experiment workflow for proposal method.

Method details

The overall steps of the proposed method are illustrated in the following data flowchart shown in Fig. 1.

Data acquisition

The study uses the publicly available (Heart Disease Dataset) from the UCI Machine Learning Repository, which combines data from four different medical centers in the US, UK, Switzerland, and Hungary. A compact version available on Kaggle data was used for this study, which contains 1192 patients for heart disease prediction. Each patient record consists of 11 attributes (features) related to demographic, clinical, and physiological measurements, along with a target variable indicating whether the patient has heart disease or not. The dataset includes a mix of numerical, binary, and categorical features, representing essential health indicators and diagnostic test results. These attributes are commonly used in cardiovascular risk assessment and heart disease prediction studies. The target variable is binary, with a value of 1 indicating the presence of heart disease and 0 indicating its absence. The selected attributes are based on their recognized clinical significance in cardiovascular risk assessment and prevalence in heart disease prediction research, encompassing demographic, symptomatic, physiological, and diagnostic test variables that collectively offer a comprehensive profile for disease prediction Then, 5-fold cross-validation was employed exclusively within the 80 % training set. The remaining 20 % was used for testing the predictive power of the trained machine learning models, as shown in Table 1.

Data cleansing and preprocessing

The data undergoes pre-processing to remove any missing or noisy records. The data undergoes a cleansing phase that involves eliminating records with excessive noise and outliers and recognizing missing values as absent to guarantee accurate outcomes. Furthermore, security concerns have eliminated several features, including the patient's name, personal identification number, passport number, date of birth, place of birth, resident address, and cell phone number. To ensure that there is no multicollinearity issue, the correlation coefficients were checked among the features. The highest correlation is between Old Peak and ST Slope, which is 0.52. This indicates that there is no multicollinearity problem.

Applying classifiers

K-Nearest Neighbors (KNN) and Support Vector Machine (SVM) were chosen because they are widely established, effective, and interpretable classifiers in heart disease prediction tasks, as supported by recent literature. Additionally, these models benefit significantly from hyperparameter optimization, making them suitable candidates for enhancement using Genetic Algorithms (GA) to improve accuracy and predictive reliability. To achieve the primary objective of this article, the following ML techniques are used:

K-Nearest neighbor (KNN)

The K-nearest neighbors (KNN) technique is a supervised machine learning algorithm used for classification and regression

Table 1
Dataset attribute details and values.

Attributes	Meaning	Value
Age	Patient age	Integer value
Sex	Male or Female	0 = Male 1 = Female
Chest Pain Type	the kind of Chest pain	1 = Typical angina 2 = Atypical angina 3 = non-anginal pain 4 = Asymptomatic
Resting Pb S	diastolic blood pressure in rest	Integer value
Cholesterol	Cholesterol level	Integer value
Fasting Blood Sugar	Blood sugar in level fasting	0 = No 1 = Yes
Resting ECG	ECG in resting	0 = Minor 1 = Intermediate 2 = Major/Complex
Max Heart Rate	Maximum heart rate	Integer value between 60 and 202
Exercise Angina	Angina observed during exercise	0 = No 1 = Yes
Old Peak	ST Segment depression experienced by exercise relative to rest	Integer value measured in depression
ST Slope	The slope of the peak exercise	1 = Upsloping 2 = Flat 3 = Down sloping
Target	whether the patient has a heart disease or not	0 = Without Heart Disease 1 = With Heart Disease

prediction tasks [20]. The classification of new data points in KNN is based on assigning them the class labels of their closest neighbor in the feature space [9]. The KNN algorithm contains multiple hyperparameters that must be adjusted to maximize the model's performance [12]. The following are the hyperparameters:

1. Number of Neighbors (*K*): It determines the quantity of closest neighbors to consider while producing a prediction. The ideal value of K depends upon the problem's complexity and the amount of available training data.
2. *Distance* Metric: It is used to compute the distance between the query and training points. Several popular measures are Euclidean distance, Manhattan distance, Minkowski distance, and Cosine distance. The selection of a distance metric can significantly impact the effectiveness of the KNN algorithm, depending on the specific problem being addressed.
3. *Weights*: It combines the labels of the closest neighbors. Two commonly used options are uniform Weights and Distance-Weighted Weights.

Support vector machines (SVM)

Support Vector Machines (SVM) is a beneficial and flexible supervised machine learning algorithm for classification and regression [3]. The primary goal of the SVM method is to identify the best hyperplane in an N-dimensional space that can effectively separate the data points belonging to distinct classes in the feature space [7]. SVM needs to select a hyperplane that best divides data instances belonging to separate categories to maximize the margin between any two nearest data points from different classes [4]. The hyperplane is chosen to maximize the distance between the nearest data points from different categories, or, in other words, the margin [15]. The equation for the optimal separating hyperplane can be given as Eq. (1):

$$f(x) = w^T x + b \tag{1}$$

Where:

w - weight vector,
x - attribute values,
b - scalar (i.e., bias term).

SVM has several hyperparameters to be tuned to optimize the model's performance. Those hyperparameters are:

1. The *kernel functions* determine the nature of the decision boundary that the SVM will learn. The popular kernel functions are linear, polynomial, radial basis function (RBF), and sigmoid.
2. The *regularization parameter*, denoted as *C*, determines the balance between maximizing the margin and minimizing classification error.
3. The *gamma* parameter is utilized in the RBF, Polynomial, and Sigmoid kernels. It regulates the impact of an individual training instance on the decision boundary.

Genetic algorithm

A Genetic Algorithm (GA) will be applied to enhance the performance of ML techniques, yielding improved accuracy and classification reports. GA, classified under the evolutionary algorithms category, replicates Darwin's theory of evolution [9]. GA is utilized to address challenging optimization issues, frequently characterized as NP-hard [21]. The genetic encoding and evaluation function are contingent upon the specific problem and domain being addressed. Once these parameters are established, a GA systematically progresses through rounds of selection, crossover, and mutation to enhance a population of individuals representing potential solutions to the problem [11]. A Genetic Algorithm (GA) was implemented to optimize the SVM and KNN models for reproducibility, ensuring that all experiments were repeated 10 times with different random seeds. The GA's fitness function was the mean accuracy score from 5-fold cross-validation on the training set. Each chromosome represented specific hyperparameters: KNN included the number of neighbors (K) and distance metric, while SVM contained the C parameter, kernel type, and gamma value. The process involved 50 individuals over 100 generations, utilizing tournament selection, one-point crossover with a probability of 0.8, and uniform mutation with a probability of 0.1 to identify optimal hyperparameters. The computational aspects of the study were built using Python, chosen for its wide range of machine-learning libraries. The Pandas library was used for data leading and cleaning, and the Scikit-learn library was used to implement the KNN and SVM classifiers and to conduct model evaluations. The pseudocode is in detail, clearly illustrating how each stage of the algorithm contributes to optimizing the performance of the machine learning model.

Pseudocode for Genetic Algorithm

```
Start
For seed = 1 to 10 do
    Set random seed
    Initialize population P of 50 individuals (random hyperparameters)
For generation = 1 to 100 do
        Evaluate fitness of each individual via 5-fold cross-validation
```

(*continued on next page*)

(*continued*)

```
    Select parents using Tournament Selection
    Apply One-point Crossover (prob = 0.8)
    Apply Uniform Mutation (prob = 0.1)
    Replace old population with new offspring
End For
Record best solution for this seed
End For
Select overall best hyperparameters across all seeds
Stop
```

Proposed algorithm (SVM and KNN- genetic optimization)

SVM and KNN with Genetic Optimization refer to approaches where Support Vector Machines (SVM) and K-Nearest Neighbors (KNN) are combined with a Genetic Algorithm (GA) to improve their performance**,** typically by optimizing parameters or feature selection.

Genetic-SVM

Genetic-SVM is a hybrid approach that combines Genetic Algorithm (GA) and Support Vector Machine (SVM). In this method, a Genetic Algorithm is used to optimize the hyperparameters of the SVM, aiming to improve its performance in classification tasks. SVMs require proper tuning of hyperparameters, such as the regularization parameter (C) and kernel parameters (e.g., gamma in the RBF kernel), to achieve the best results. Genetic-SVM automates this process through an evolutionary approach. When it used SVM, it obtained the 'poly' as the best kernel function, 1.0 as the optimal gamma value, and 1.0 as the best C regularization value.

Genetic-KNN

Genetic-KNN is a hybrid approach that combines K-Nearest Neighbors (KNN) with a Genetic Algorithm (GA) to optimize KNN's performance, typically by tuning the hyperparameters and performing feature selection. The goal of using GA with KNN is to enhance the accuracy and efficiency of the KNN classifier by automatically finding the optimal values for its parameters, such as the number of neighbors **(**K**)** and the most relevant feature subset.

Evaluation

To assess the effectiveness of SVM and KNN algorithms, the confusion matrix, accuracy, precision, recall, and F1-score measures can be used.

Confusion matrix

The confusion matrix is employed to assess the precision of the machine learning technique. TP, FP, TN, and FN are acronyms for true positive, false positive, true negative, and false negative, respectively. The confusion matrix in Table 2 typically assesses these four metrics. The subsequent pertains to the standard confusion matrix [23]:

Where

- TP: True Positive refers to heart disease patients correctly identified by the ML model.
- TN: True Negative refers to individuals without heart abnormalities correctly classified by the ML model.
- FP: False Positive refers to patients with heart disease who the ML model wrongly classifies.
- FN: False Negative refers to individuals who do not have heart disease but are wrongly classified as having it by the ML model.

Accuracy

The accuracy can be measured by calculating the percentage of correct observations relative to the total number of observations. Eq. (2) shows how the accuracy is measured:

$$Accuracy = \frac{TP + TN}{TP + FP + FN + TN} \tag{2}$$

Table 2
Confusion matrix.

Confusion matrix		Predictive class	
		0	1
Actual Class	**0**	TP	FN
	1	FP	TN

Precision

It is the ratio of accurately anticipated positive observations to the total number of positive observations. The precision is measured in Eq. (3).

$$Precision = \frac{TP}{TP + FP} \tag{3}$$

Recall/ sensitivity

The procedure involves calculating the ratio of accurately predicted positive outcomes to the total number of observations. Sensitivity and response capacity are often used interchangeably. The recall is measured in Eq. (4).

$$Recall = \frac{TP}{TP + FN} \tag{4}$$

F1-Score

It denotes the calculated average of the precision and recall values, taking into account their respective weights. The F1-score is calculated in Eq. (5).

$$F1 - Score = 2 \times \left[\frac{precision \times recall}{(precision + recall)}\right] \tag{5}$$

Results and discussion

First of all, we will apply SVM and KNN models without tuning hyperparameters:

Table 3 presents the evaluation metrics for two machine learning models, K-Nearest Neighbors (KNN) and Support Vector Machine (SVM), before hyperparameter optimization. The metrics include Accuracy, Recall, Precision, and F1 Score, which are common performance indicators for classification models. According to the results, the Accuracy of both models is around 89 %, showing comparable performance. However, the Recall and Precision are slightly different, with SVM having a better performance in identifying positives (Recall: 0.81 for SVM vs. 0.78 for KNN) and correctly predicting positive outcomes (Precision: 0.87 for SVM vs. 0.84 for KNN). The F1-score is similarly higher for SVM, reflecting a better balance between Precision and Recall for this model. In other words, it illustrates that SVM performs marginally better than KNN in its default form, particularly when it comes to Precision and Recall, making it more effective at minimizing false positives and false negatives in heart disease prediction.

Fig. 2 displays the Receiver Operating Characteristic (ROC) curves for the KNN and SVM models without Genetic Algorithm (GA) optimization. The ROC curve plots the True Positive Rate (Sensitivity) against the False Positive Rate (1-Specificity) at various threshold settings. The ROC curves for both KNN and SVM before optimization provide a visual representation of the models' performance. The higher the curve, the better the model is at distinguishing between the classes. If the curve is closer to the top-left corner, it indicates better performance. As a result, without optimization, the ROC curves show that both models perform well, with 0.92 for KNN and 0.94 for SVM, suggesting that SVM may have a slight edge. The ROC curve demonstrates that the models are effective but leave room for improvement in classification performance.

Table 4 shows the evaluation metrics after applying Genetic Algorithm (GA) optimization to both models. The same metrics (Accuracy, Recall, Precision, F1-Score) are shown, reflecting the improvement due to hyperparameter tuning. For KNN, the Accuracy increased to 95.38 % (± 47), and all other metrics (Recall, Precision, and F1) reached a value of 0.95 (± 0.25), indicating a strong balance between Precision and Recall. For SVM, the Accuracy improved slightly to 90 % (± 0.19), and Precision increased significantly to 0.92 (± 0.38), though the Recall decreased to 0.79 $\pm$ 0.54). The table demonstrates that the Genetic Algorithm (GA) significantly improved the KNN model's performance, with all metrics showing a notable increase. The improvement for SVM is less pronounced, though it achieved a much higher Precision, making it more reliable at predicting positive cases. In other words, KNN benefited more from GA optimization, with a higher overall performance. This suggests that KNN's hyperparameters were more amenable to optimization using GA. SVM also improved, but the trade-off between Recall and Precision indicates that it became better at predicting positives while potentially missing some actual positive cases (lower Recall).

Fig. 3 shows the ROC curves after Genetic Algorithm (GA) optimization for KNN and SVM. The curves offer an updated view of the models' ability to distinguish between classes after tuning their hyperparameters. The ROC curves after optimization for both models are likely higher (closer to the top-left corner), indicating improved model performance. The AUC values (area under the curve) would now reflect higher accuracy in distinguishing between positive and negative cases, especially for KNN, which showed significant improvements (increased from 0.92 to 0.99) across all metrics. The KNN ROC curve would likely exhibit a significant increase in AUC, reflecting the improvement in the model's ability to classify both positive and negative cases correctly. The SVM ROC curve would also

Table 3
Evaluation of the traditional KNN and SVM methods.

Method without optimization	Accuracy	Recall	Precision	F1
KNN	89.08 %	0.78	0.84	0.81
SVM	89.92 %	0.81	0.87	0.84

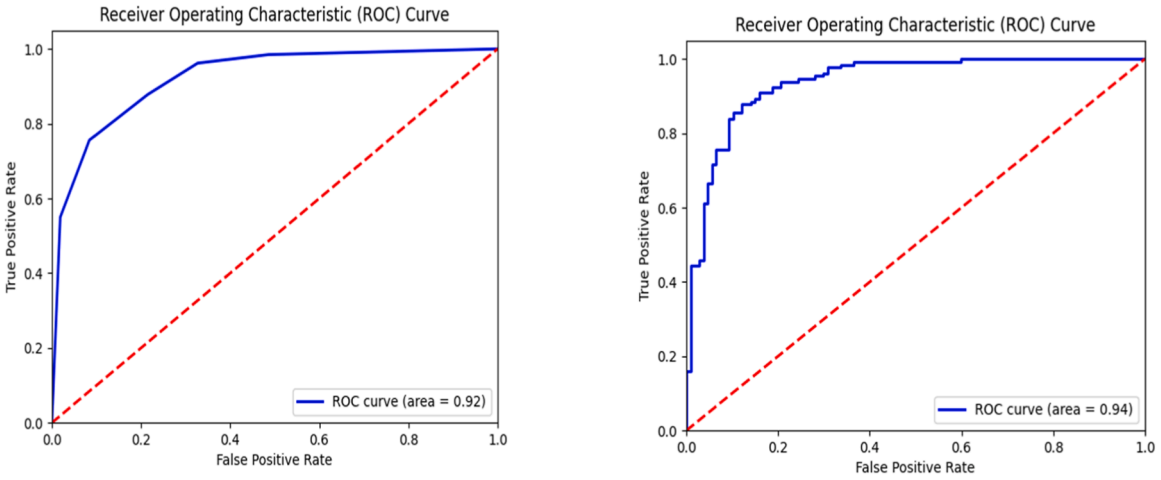


Fig. 2. ROC under the curve of the KNN and SVM without GA.

Table 4
Evaluation of the KNN and SVM methods with GA optimization.

Method with Optimization	Accuracy	Recall	Precision	F1
KNN	0.95.38 % ±0.47	0.95 % ± 0.25	0.95 %±0.19	0.95 %±0.39
SVM	90 %±0.82	0.79 %±0.54	0.92 %±0.38	0.85 %±0.28

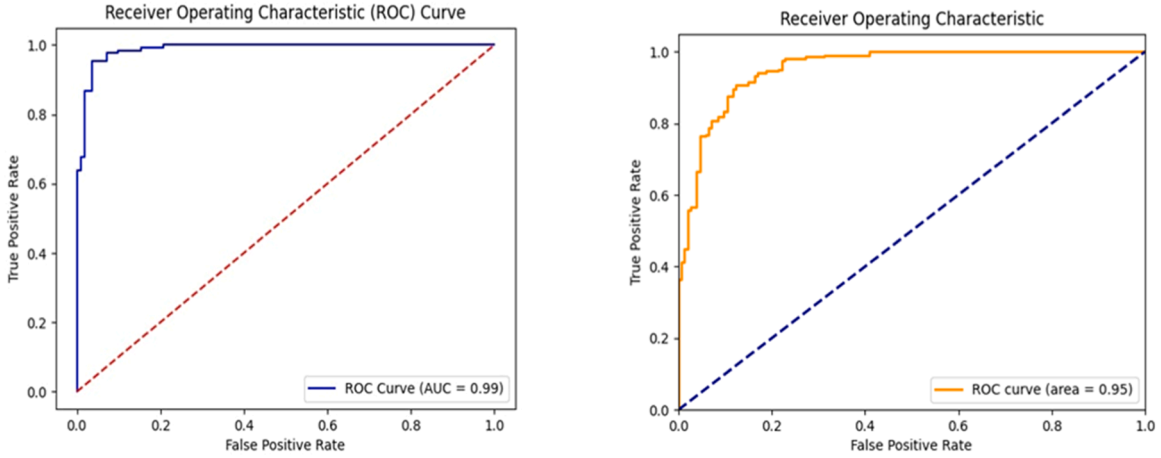


Fig. 3. ROC under the curve of the KNN and SVM with GA.

show some improvement (increased from 0.94 to 0.95, though the benefits are more pronounced for KNN. These curves indicate that the Genetic Algorithm (GA) optimization successfully enhanced both models, particularly the KNN model.

Fig. 4 illustrates the performance of the KNN and SVM algorithms before and after optimization using a Genetic Algorithm (GA). There are the Accuracy, Recall, Precision, and F1 scores. For KNN, it appears that all the metrics (Accuracy, Recall, Precision, F1) show significant improvement after GA optimization. For SVM, improvements are evident, but they may not be as dramatic as those for KNN. Notably, Precision has increased substantially, while Recall has decreased slightly. Overall, Fig. 5 likely demonstrates that GA optimization effectively enhances the performance of both KNN and SVM, with KNN potentially benefiting more.

Fig. 5 shows a horizontal bar chart titled "Feature Importance using KNN with GA." It ranks the importance of features for a model that employs K-Nearest Neighbors (KNN) combined with a Genetic Algorithm (GA). The most important feature is "old peak," followed by "cholesterol" and "resting BPS." The chart uses gradient colors to represent the importance, with the least essential features being "ST slope_3," "chest pain type_1," and "ST slope_0."

This study investigated the application of Genetic Algorithm to optimize machine-learning models for heart disease prediction. The study focused on K-Nearest Neighbors and Support Vector Machine classifiers, aiming to improve performance by tuning

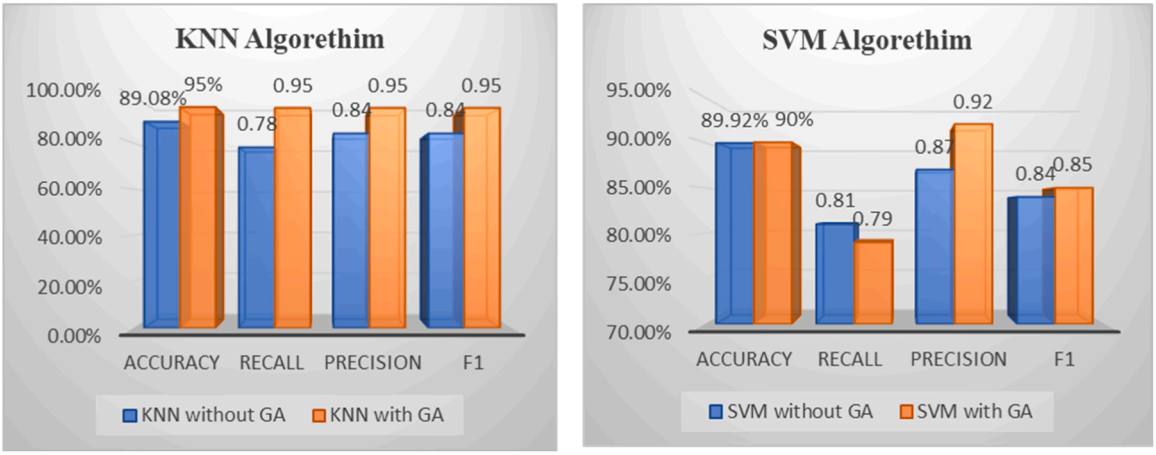


Fig. 4. Comparison of the performance of the KNN and SVM algorithms before and after optimization.

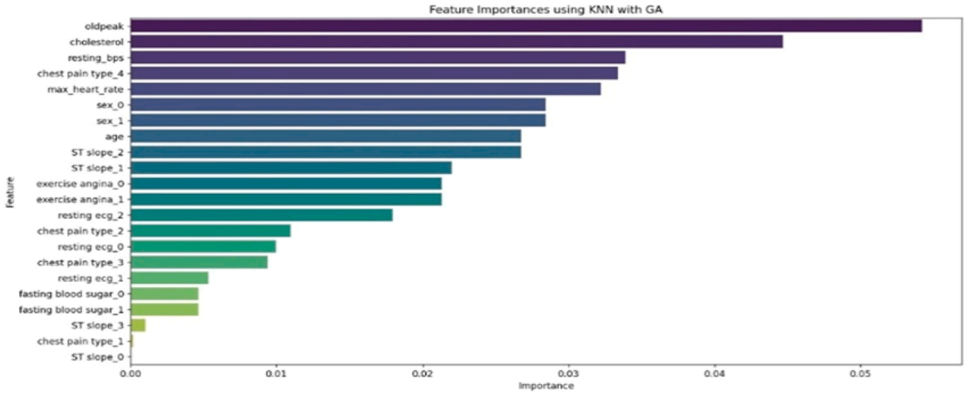


Fig. 5. Feature importance using KNN with GA.

hyperparameters using a genetic algorithm (GA). The initial evaluation of the traditional KNN and SVM models without hyperparameter tuning revealed comparable performance, with SVM exhibiting slightly better precision and recall. This observation aligns with the general understanding of these algorithms, where SVM often shows higher accuracy but can be computationally more complex [15]. However, the integration of GA optimization led to significant improvements in both models. KNN, in particular, demonstrated a substantial increase in accuracy, recall, precision, and F1-score, achieving 95.38 % accuracy and 0.95 for all other metrics. This result highlights the effectiveness of GA in optimizing KNN's hyperparameters, specifically the number of neighbors (K) and the chosen distance metric, resulting in more accurate classification. SVM also benefited from GA optimization, with a notable increase in precision, although a slight decrease in recall was observed. This trade-off suggests that GA tuning might have prioritized minimizing false positives, making the SVM model more reliable in predicting positive cases but potentially missing some actual positives. The ROC curves visually confirmed the performance enhancement after GA optimization. The AUC values for both models increased, indicating improved accuracy in distinguishing between positive and negative cases, especially for KNN. These findings reinforce the idea that the Genetic Algorithm successfully optimized the models' hyperparameters, enhancing their ability to distinguish between patients with and without heart disease. The feature importance analysis using KNN with GA identified the most influential features in predicting heart disease, including age, peak, cholesterol, and resting blood pressure. This information aligns with existing medical knowledge, which recognizes these factors as significant contributors to cardiovascular diseases [19]. This identification of crucial features is highly valuable as it provides insights for developing targeted interventions and risk assessments in medical practice. The study's findings are consistent with previous research demonstrating the efficacy of GA in optimizing machine learning models for various applications, including disease prediction.

Comparative GA optimization method with grid search and random search

Table 5 compares the performance of the KNN algorithm when optimized using three different hyperparameter optimization approaches: Genetic Algorithm, Grid Search, and Random Search. Among these strategies, the Genetic Algorithm achieved the highest accuracy (0.95), precision (0.95), recall (0.95), and F1-score (0.95), indicating outstanding model performance and a strong balance

Table 5
Optimization of KNN with different optimization methods.

Optimizations algorithms	Accuracy	Precision	Recall	F1-Score
Genetic Algorithm	0.95	0.95	0.95	0.95
Grid search	0.86	0.86	0.89	0.87
Random search	0.87	0.88	0.90	0.89

between sensitivity and specificity. In contrast, Grid Search resulted in lower metrics, with an accuracy of 0.86 and an F1-score of 0.87, reflecting its limited effectiveness due to its exhaustive yet rigid search approach. Random Search performed slightly better than Grid Search, achieving an accuracy of 0.87 and an F1-score of 0.89, benefiting from its ability to explore a broader search space randomly. These results demonstrate that the Genetic Algorithm (GA) is more effective and efficient for optimizing KNN hyperparameters in heart disease prediction, surpassing traditional tuning methods in all evaluation metrics.

The limitations of this study present several possibilities for future research. Our findings are based on a single public dataset, so validating the model's generalizability with larger and more diverse datasets is crucial. While this study focused on K-Nearest Neighbors (KNN) and Support Vector Machines (SVM), future research could investigate employing a genetic algorithm (GA) to optimize other machine learning models, such as ensemble methods, neural networks, or deep learning algorithms. This could further improve the accuracy and efficiency of heart disease predictions.

Conclusion

This study investigated the potential of Genetic Algorithm (GA) in optimizing machine learning models for heart disease prediction. Focusing on KNN and SVM classifiers, the research successfully demonstrated that GA effectively enhances the performance of these models. The Key findings of this study include: GA significantly improved the accuracy and other performance metrics of both KNN and SVM models. KNN exhibited more significant improvement compared to SVM, suggesting greater benefit from GA optimization. Furthermore, feature importance analysis highlighted key risk factors for heart disease, aligning with existing medical knowledge. These results highlight the potential of integrating Genetic Algorithms in machine learning.

Limitation

The genetic algorithm optimization has no additional effect on other algorithms, such as Decision Tree and Naïve Bayes, after being tested in this experiment.

Ethics statements

Not applicable.

Credit author statement

Sherko H. Murad: Conceptualization, methodology, formal analysis, investigation, and writing—original draft. **Noor Bahjt Tayfor:** Methodology, Writing—Original Draft **Nozad H. Mahmood:** Conceptualization, Formal Analysis, Investigation, Writing—Original Draft. **Arman Lawson**: Conceptualization, Supervision, Writing—Review & editing

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

Data availability

The analysis code used to generate the findings is available from the corresponding author upon reasonable request.

References

[1] M. Amini, F. Zayeri, M. Salehi, Trend analysis of cardiovascular disease mortality, incidence, and mortality-to-incidence ratio: results from global burden of disease study, BMC Public Health 21 (1) (2021), https://doi.org/10.1186/s12889-021-10429-0.
[2] S.I. Ayon, M.M. Islam, M.R. Hossain, Coronary artery heart disease prediction: a comparative study of computational intelligence techniques, IETE J. Res. 68 (4) (2021) 2488–2507, https://doi.org/10.1080/03772063.2020.1713916.

[3] J. Cervantes, F. Garcia-Lamont, L. Rodríguez-Mazahua, A. Lopez, A comprehensive survey on support vector machine classification: applications, challenges and trends, Neurocomputing 408 (2020) 189–215.
[4] R.-C. Chen, C. Dewi, S.-W. Huang, R.E. Caraka, Selecting critical features for data classification based on machine learning methods, J. Big Data 7 (1) (2020) 52.
[5] R. DESAI, Structural optimization using genetic algorithm. Handbook of AI-based Metaheuristics, CRC Press, 2021.
[6] K.S.S. DHANUSH, Gene expression analysis using SVM and KNN classifiers on various datasets, in: 2023 5th International Conference on Smart Systems and Inventive Technology (ICSSIT), IEEE, 2023, pp. 1325–1332.
[7] S. Ghosh, A. Dasgupta, A. Sweta Padma, A study on support vector machine based linear and non-linear pattern classification, in: 2019 International Conference on Intelligent Sustainable Systems (ICISS), 2019, pp. 24–28x.
[8] A. Hamdard, H. Lodin, Effect of feature selection on the accuracy of machine learning model, Int. J. Multidiscip. Res. Anal. 6 (2023), https://doi.org/10.47191/ijmra/v6-i9-66.
[9] S. Huang, M. Huang, Y. Lyu, A novel approach for sand liquefaction prediction via local mean-based pseudo nearest neighbor algorithm and its engineering application, Adv. Eng. Inform. 41 (2019) 100918.
[10] S.D.C. IMMANUEL, U. K, Genetic algorithm: an approach on optimization, in: international conference on communication and electronics systems (ICCES), IEEE, 2019, pp. 701–708.
[11] K. Le, T. Wei, J. Tagalog on, Genetic Algorithm and Application, Science Open Preprints, 2023.
[12] M. Loukili, F. Messaoudi, M. El Ghazi, Supervised learning algorithms for predicting customer churn with hyperparameter optimization, Int. J. Adv. Soft Comput. Appl. 14 (3) (2022).
[13] D. Muhajir, M. Akbar, A. Bagaskara, R. Vinarti, Improving classification algorithm on education dataset using hyperparameter tuning, Procedia Comput. Sci. 197 (2022) 538–544.
[14] M.A. Naser, A.A. Majeed, M. Alsabah, T.R. Al-Shaikhli, K.M. Kaky, A Review of Machine Learning's Role in Cardiovascular Disease Prediction: Recent Advances and Future Challenges. In Algorithms, 17, Multidisciplinary Digital Publishing Institute (MDPI), 2024, https://doi.org/10.3390/a17020078.
[15] S. RAY, An analysis of computational complexity and accuracy of two supervised machine learning algorithms—K-nearest neighbor and support vector machine, in: Data Management, Analytics and Innovation: Proceedings of ICDMAI 1, Springer, 2021, pp. 335–347.
[16] G.A. Roth, G.A. Mensah, C.O. Johnson, G. Addolorato, E. Ammirati, L.M. Baddour, N.C. Barengo, A. Beaton, E.J. Benjamin, C.P. Benziger, A. Bonny, M. Brauer, M. Brodmann, T.J. Cahill, J.R. Carapetis, A.L. Catapano, S. Chugh, L.T. Cooper, J. Coresh, V. Fuster, Global burden of cardiovascular diseases and risk factors, J. Am. Coll. Cardiol. 76 (25) (2020) 2982–3021, https://doi.org/10.1016/j.jacc.2020.11.010.
[17] A.K. SARMA, "Introduction to genetic algorithm with a simple analogy. Nature-inspired methods for metaheuristics optimization," Algorithms and Applications in Science and Engineering, pp. 27–34, 2020.
[18] J. Sekar, P. Aruchamy, H. Sulaima Labbe Abdul, A.S. Mohammed, S. Khamuruddeen, An efficient clinical support system for heart disease prediction using TANFIS classifier, Comput. Intell. 38 (2) (2022) 610–640.
[19] A. N. Sinha, T. Jangid, M. Joshi, P. Mohanty, A Machine Learning Based Smart Healthcare Framework For Cardiovascular Disease Prediction, S., iCardo, 2022.
[20] M. Srivenkatesh, Prediction of cardiovascular disease using machine learning algorithms, Int. J. Eng. Adv. Technol. 9 (3) (2020) 2404–2414.
[21] X. Yang, Z. Wang, H. Zhang, N. Ma, N. Yang, H. Liu, H. Zhang, L. Yang, A review: machine learning for combinatorial optimization problems in energy areas, Algorithms 15 (6) (2022) 205.
[22] I.K.A. Sugitha, A. Triayudi, E.T.E. Handayani, Classification of heart disease using the K-nearest neighbor algorithm and logistic regression, J. Pilar Nusa Mandiri 20 (2) (2024) 183–190.
[23] S.H. Murad, A.H. Awlla, B.T. Moahmmed, Prediction Lung Cancer based critical factors using machine learning, Sci. J. Univ. Zakho 11 (3) (2023) 447–452.
[24] G. Prabaharan, S.U. Sankar, V. Anusuya, K.J. Deepthi, R. Lotus, R. Sugumar, Optimized disease prediction in healthcare systems using HDBN and CAEN framework, MethodsX (2025) 103338.
[25] D. Santhakumar, G. Rajaram, R. Elankavi, J. Viswanath, I. Govindharaj, J. Raja, Enhanced leukemia prediction using hybrid ant colony and ant lion optimization for gene selection and classification, MethodsX 14 (2025) 103239.