İkram HALICI

M.Sc. Thesis

AGU 2024

DIAGNOSIS OF CORONARY ARTERY DISEASE WITH MACHINE LEARNING APPROACHES

M.Sc. THESIS
SUBMITTED TO THE DEPARTMENT OF ELECTRICAL AND COMPUTER ENGINEERING
AND THE GRADUATE SCHOOL OF ENGINEERING AND SCIENCE
OF ABDULLAH GUL UNIVERSITY
IN PARTIAL FULFILLMENT OF THE REQUIREMENTS
FOR THE DEGREE OF
MASTER OF SCIENCE

By
İkram HALICI
January 2024

DIAGNOSIS OF CORONARY ARTERY DISEASE WITH MACHINE LEARNING APPROACHES

A THESIS
SUBMITTED TO THE DEPARTMENT OF ELECTRICAL AND COMPUTER ENGINEERING
AND THE GRADUATE SCHOOL OF ENGINEERING AND SCIENCE OF ABDULLAH GUL UNIVERSITY
IN PARTIAL FULFILLMENT OF THE REQUIREMENTS
FOR THE DEGREE OF
MASTER OF SCIENCE

By
İkram HALICI
January 2024

SCIENTIFIC ETHICS COMPLIANCE

I hereby declare that all information in this document has been obtained in accordance with academic rules and ethical conduct. I also declare that, as required by these rules and conduct, I have fully cited and referenced all materials and results that are not original to this work.

Name-Surname: İkram HALICI
Signature :

REGULATORY COMPLIANCE

M.Sc. thesis titled **"DIAGNOSIS OF CORONARY ARTERY DISEASE WITH MACHINE LEARNING APPROACHES"** has been prepared in accordance with the Thesis Writing Guidelines of the Abdullah Gül University, Graduate School of Engineering & Science.

Prepared By	Advisor
İkram HALICI	Prof. V. Çağrı GÜNGÖR

Head of the Electrical and Computer Engineering Graduate Program
Asst. Prof. Samet GÜLER

ACCEPTANCE AND APPROVAL

M.Sc. thesis titled **"DIAGNOSIS OF CORONARY ARTERY DISEASE WITH MACHINE LEARNING APPROACHES"** and prepared by Ikram Halici has been accepted by the jury in the Electrical and Computer Engineering Graduate Program at Abdullah Gül University, Graduate School of Engineering & Science.

24 /01 / 2024

(Thesis Defense Exam Date)

JURY:

Advisor : Prof. V. Çağrı GÜNGÖR

Member : Assoc. Prof. Özkan Ufuk NALBANTOĞLU

Member : Assist. Prof. Rıfat KURBAN

APPROVAL:

The acceptance of this M.Sc. thesis has been approved by the decision of the Abdullah Gül University, Graduate School of Engineering & Science, Executive Board dated ….. /….. / …….. and numbered ……..……. .

…….. /…….. / ……..

(Date)

Graduate School Dean
Prof. İrfan ALAN

ABSTRACT

DIAGNOSIS OF CORONARY ARTERY DISEASE WITH MACHINE LEARNING APPROACHES

İkram HALICI
MSc. in Electrical and Computer Engineering
Advisor: Prof. V. Çağrı GÜNGÖR
January 2024

The World Health Organization states that Coronary Artery Disease (CAD) ranks as a primary cause of recorded fatalities. CAD occurs as a result of the blockage of coronary artery vessels, which are located on the surface of the heart and supply the blood that the heart needs. Diagnosing the disease using traditional methods is challenging and requires costly tests. In recent years, the use of machine learning-based methods has increased as an alternative diagnostic approach. However, existing studies in the literature suffer from low detection rates and long training times. Therefore, there is still a need for reliable and low-cost diagnostic methods. In this thesis, a new model, CSA-PSO-ANN, is proposed for the diagnosis of coronary artery disease. The aim is to reduce the training time of the machine learning model and achieve a higher accuracy in diagnosing the disease. Experiments have been conducted on two publicly available datasets. Parallelization, feature selection, and hyperparameter optimization have been performed to shorten the model's training time. The performance of the model has been compared with well-known machine-learning algorithms and previous studies. The experiments showed that the proposed model effectively diagnoses the disease and outperforms other methods in terms of accuracy and F1 score performance metrics.

Keywords: Coronary Artery Disease Diagnosis, Artificial Neural Network, Hybrid Optimization Algorithms, Clonal Selection Algorithm, Particle Swarm Optimization

ÖZET

KORONER ARTER HASTALIĞININ MAKİNE ÖĞRENİMİ YAKLAŞIMLARI İLE TEŞHİSİ

İkram HALICI
Elektrik ve Bilgisayar Mühendisliği Anabilim Dalı Yüksek Lisans
Tez Danışmanı: Prof. V. Çağrı GÜNGÖR
January 2024

Dünya Sağlık Örgütü'nün verilerine göre, koroner arter hastalığı(KAH), bilinen ölüm nedenlerinin önde gelen sebeplerinden biridir. KAH, kalp yüzeyinde bulunan ve kalbin kan ihtiyacını karşılayan koroner arter damarlarının tıkanması sonucunda oluşmaktadır. Hastalığın geleneksel yöntemlerle teşhis edilmesi zordur ve maliyetli testlerin yapılmasını gerektirmektedir. Son yıllarda alternatif teşhis yöntemi olarak makine öğrenimi tabanlı yöntemlerin kullanımı artmıştır. Ancak, mevcut literatürdeki çalışmalar düşük tespit oranları ve uzun eğitim sürelerinden muzdariptir. Bu nedenle, güvenilir ve düşük maliyetli teşhis yöntemlerine olan ihtiyaç devam etmektedir. Bu tez çalışmasında, koroner arter hastalığının teşhisi için yeni bir model, CSA-PSO-ANN, önerilmektedir. Önerilen yöntem ile makine öğrenimi modelinin eğitim süresinin kısaltılması ve hastalığın daha yüksek doğruluk oranı ile teşhis edilebilmesi amaçlanmaktadır. Bu tez çalışmasında deneyler halka açık iki veriseti üzerinde gerçekleştirilmiştir. Model eğitim süresini kısaltmak için paralleştirme, öznitelik seçimi ve hiperparametre optimizasyonu yapılmıştır. Model performansı literatürde bilinen makine öğrenimi algoritmaları ve geçmiş çalışmalarla karşılaştırılmıştır. Yapılan deneyler sonucunda, önerilen modelin hastalığın teşhisi konusunda etkili çalıştığı ve diğer yöntemlere göre doğruluk ve F1 skoru performans ölçümlerinde daha iyi sonuçlar elde ettiği görülmüştür.

Anahtar kelimeler: Koroner Arter Hastalığı Teşhisi, Yapay Sinir Ağı, Hibrid Optimizasyon Algoritmaları, Klonal Seçim Algoritması, Parçacık Sürü Optimizasyonu

Acknowledgements

i. I want to express my appreciation to my supervisor, Professor V. Çağrı GÜNGÖR, for his guidance and support throughout this research. His expertise and encouragement significantly contributed to the completion of this thesis.

ii. My cordial thanks also extend to Dr. Özkan Ufuk NALBANTOĞLU, and Dr. Rıfat KURBAN for serving on my dissertation defense committee.

iii. My sincere thanks are Dr. Bilge Kağan DEDETÜRK and Dr. Burak KOLUKISA for their support, valuable contributions and encouragement, which played a crucial role in shaping the outcome of this thesis.

iv. I express deep appreciation to my family for their support, sacrifices and belief in my capabilities.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

LIST OF ABBREVIATIONS

ACC	Accuracy
ANN	Artificial Neural Network
CAD	Coronary Artery Disease
CSA	Clonal Selection Algorithm
CV	Cross Validation
DT	Decision Tree
F1	F1 Score
FS	Feature Selection
HD	Heart Disease
Hybrid Opt.	Hybrid Optimization
KNN	k-Nearest Neighbors
LR	Logistic Regression
MLP	Multi Layer Perceptron
Opt.	Optimization
PP	Preprocessing
PSO	Particle Swarm Optimization
RT	Running Time
SVM	Support Vector Machine

To my first teacher, my dear aunt,
Hatice Halıcı

Chapter 1

Introduction

Cardiovascular diseases are the foremost reasons for mortality across the global. A 2021 World Health Organization report indicates that these disease are account for 32% of all deaths worldwide [1]. While there are various types of this condition, one of the commonly encountered forms is referred to as coronary artery disease(CAD). The heart meets its blood needs through vessels known as coronary arteries on the heart's surface. Coronary Artery Disease (CAD) develops due to the obstruction of these vessels. Deposits of cholesterol and other materials accumulate in the artery, resulting in the formation of plaque, as depicted in Figure 1.1. Eventually, this plaque fills the artery and restricts blood flow in the vessel. In the initial stages of this condition, chest pain may be noticed due to the constriction of these vessels and limited blood flow. Diagnosing coronary artery disease (CAD) presents a challenge. As a result, the primary symptoms in patients often include severe conditions like heart attack and heart failure.

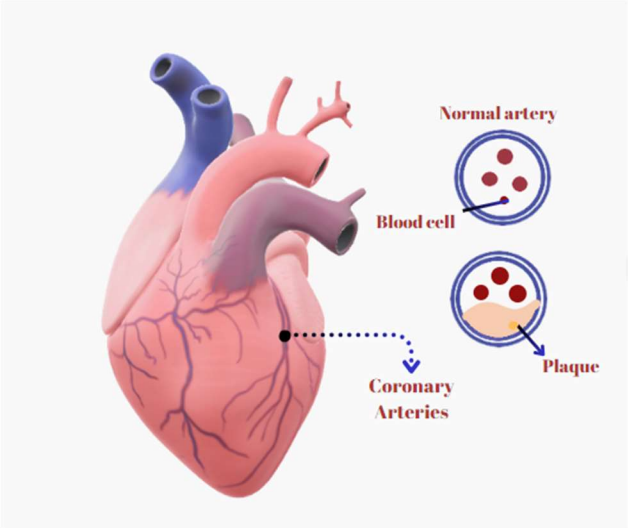

Figure 1.1 Coranary Artery Disease

The initial step in CAD diagnosis includes assessing whether a patient belongs to the risk group. If determined to be at risk, various costly tests, including blood tests, chest X-ray, coronary angiogram, electrocardiogram, and echocardiogram, are conducted [2].

Since considering the diagnosis of the disease presents a challenge and the traditional methods of the diagnosis is costly, alternative ways are being explored. Given the increasing momentum of machine-learning solutions in addressing complex issues in recent decades, there is a potential application of machine-learning methods in the diagnosis of CAD. In [3], numerous studies have been conducted to classify patients as having or not having CAD using different ML methods on various datasets. Observations indicate that one of the most commonly employed methods is Artificial Neural Network (ANN) for diagnosing this disease, as illustrated in Figure 1.2.

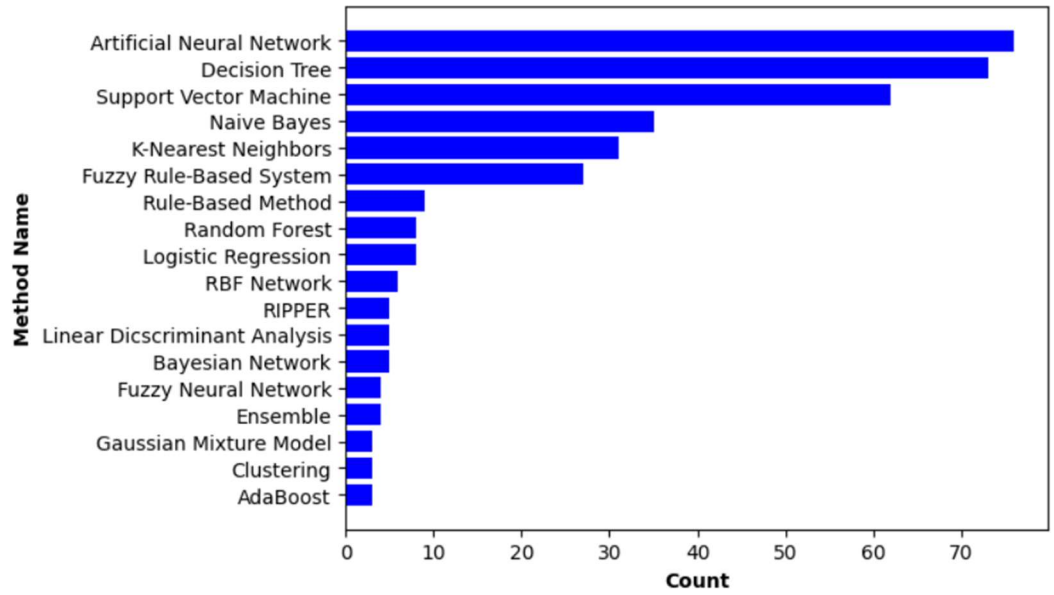

Figure 1.2 Most frequently utilized machine-learning methods for diagnosis CAD

ANN, which takes inspiration from the human brain, is composed of computationally modelled cells (neurons) and layers [4]. The learning process within ANN can be seen as the problem of adjusting the network structure and weights between layers. Since the aim is to obtain the most suitable weights and biases for the given tasks, it indicate an optimization problem. Traditional methods, such as gradient-descent-based algorithms, are used for the optimization or training of ANNs. However, these methods have limitations due to their insufficient exploration capabilities [5]. This limitation can result in being stuck in local optima, ultimately reducing the performance of ANN. Considering that ANN is one of the most frequently employed methods in Coronary Artery Disease (CAD) diagnosis and that optimization algorithms used in ANN training may constrain the model's exploration capabilities, enhancing the capabilities of ANN in CAD diagnosis represents a research gap in this field.

In this thesis, a novel approach named CSA-PSO-ANN is introduced to fill this research gap in diagnosing coronary artery disease. Clonal Selection Algorithm (CSA) is an Artificial Immune Optimization algorithm that operates based on generating antibodies with high affinity, adhering to the affinity maturation principle. This algorithm is suited for optimization problems as it conducts a local search through a hypermutation process and a global search via the reseptor editing process within a solution space. Nevertheless, when dealing with complicated optimization problems the hypermutation process in CSA may impose limitations on its search capabilities [6]. Conversely, Particle Swarm Optimization (PSO) stands out as a meta-heuristic optimization characterized by simple principles and low computational complexity [7]. The proposed method utilizes the velocity component of PSO to expand the local search capacity of CSA, and this hybrid approach is employed during the training phase of the Artificial Neural Network (ANN). This aims to enhance the exploration capabilities of the ANN, thereby achieving a more accurate diagnosis of the disease.

Experiments are conducted on two heart disease datasets that are publicly available. Data preprocessing and feature selection are applied. To accurately assess the model's performance, hyperparameter optimization and 10-fold cross-validation are employed. Additionaly, CPU/GPU parallelization is integrated to reduce training time of the proposed method.

Chapter 2

Literature Review

Numerous studies have been conducted to explore the diagnosis of heart disease, each with distinct objectives, encompassing the identification of optimal feature selection methods, the development of innovative classification models, and the formulation of effective classification methodologies.

Batanieh et al. [8] proposed the hybrid MLP-PSO algorithm and employed the Cleveland dataset, comprising 13 features and 303 samples. The authors addressed missing values by substituting them with feature-specific mean values. Additionally, they applied categorical data encoding and feature scaling to the dataset. The MLP model was executed with weights and biases determined by PSO. The evaluation was conducted through 5-fold cross-validation, utilizing the grid search method for hyperparameter tuning. By comparing the outcomes with ten different ML algorithms, the proposed method yielded the highest accuracy at 84.60%.

In a separate study, Dulhare [9] utilized the Statlog dataset, consisting of 13 features and 270 samples, employing PSO for feature selection and Naive Bayes for heart disease classification, resulting in an accuracy of 87.91%.

To enhance classification performance, researchers in [10] focused on feature extraction and the training phase of Neural Network (NN). The Statlog dataset was employed for performance testing, incorporating statistical and higher-order statistical features with PCA. In the NN training phase, a hybrid approach, named PM-LU, combining PSO and Lion Algorithm (LA), was introduced, achieving an accuracy of 87.09% and an F1 score of 84.61%. Notably, cross-validation was not implemented in this particular study.

In [11], the authors introduced an emotional neural network (EmNN) based on particle swarm optimization (PSO) and conducted a comparative performance analysis with a hybrid artificial neural network incorporating fuzzy logic (PSO-ANFIS). This investigation employed EmNNs rooted in brain-based emotional learning, a specific category associated with higher accuracy. PSO was employed for the optimization of this

neural network. The experiments were conducted on the Z-Alizadeh Sani, Cleveland, and Statlog datasets. While data preprocessing was omitted, feature selection was implemented across all datasets. Notably, 8 features were selected for the Statlog dataset and 7 for the Cleveland dataset, resulting in an accuracy of 85.2% for Statlog and 84% for Cleveland.

In the investigation detailed in [12], the author employed the Z-Alizadehsani, Cleveland, and South African datasets, introducing a novel hybrid feature selection methodology. Remarkably, they achieved an accuracy of 85.4% on the Cleveland dataset. In a subsequent study [13], the scope of feature selection methodologies was broadened to enhance overall performance outcomes. Additionally, Fisher linear discriminant analysis was implemented to streamline computational processes by reducing the number of features in coronary artery disease diagnosis, striving to create a model that performs optimally across diverse datasets. Notably, the MLP classifier yielded results of 82.5% accuracy and 83.80% F-measure for the Cleveland dataset. Another investigation, as outlined in [14], introduced an innovative, self-optimizing and adaptive ensemble algorithm for machine learning. The system autonomously selects the most suitable machine learning models, with the goal of creating an optimized, adaptive ensemble machine learning strategy that can achive high accuracy on diverse coronary artery disease datasets. Impressively, even when utilizing raw datasets, they attained an accuracy of 83.43% on the Cleveland dataset. Finally, in the study documented in [15], the Z-Alizadehsani, Cleveland, and Statlog datasets were employed to introduce an exhaustive ensemble feature selection method and a probabilistic ensemble feature selection method. This novel approach underwent evaluation using six distinct classification algorithms and four variants of voting algorithms. The outcomes revealed accuracy values of 85.47% and 85.55% for the Cleveland and Statlog datasets, respectively.

Gupta et al. [16] introduced a Clonal Selection Algorithm (CSA) integrated with k-Nearest Neighbour (KNN). They adapted the clonal selection algorithm to work in conjunction with KNN and conducted a comparative analysis with alternative machine learning algorithms for cardiovascular disease detection. The method they proposed underwent testing on a dataset comprising 12 features and 70,000 samples, yielding an accuracy of 78.40%, along with MCC and RocAuc values of 48.80% and 76%, respectively.

In their investigation [17], the authors seek to improve the efficiency of Artificial Neural Networks (ANNs) through the utilization of a hybrid optimization algorithm. They employed three standard datasets (Wisconsin Breast Cancer, Pima Indian Diabetes, and Cleveland) and adopted a combined approach of differential evolution (DE) and particle swarm optimization (PSO) for global search, coupled with backpropagation (BP) for local search during ANN training. To standardize the datasets, min-max normalization was applied before constructing the model. The evaluation involved 10-fold cross-validation, and performance comparisons were made between the proposed Differential Evolution with Global Information and Back Propagation (DEGI-BP) and the DE-BP, PSO-BP alternatives. Notably, experiments conducted on the Cleveland dataset demonstrated that the suggested approach surpassed other hybrid optimization algorithms, achieving an accuracy value of 86.66%.

The primary objective of the investigation outlined in [18] is to showcase the efficacy of deep learning algorithms, particularly Artificial Neural Networks (ANN), in predicting Coronary Artery Disease (CAD) at an early stage. The Cleveland dataset served as the basis for applying various algorithms, including Support Vector Machines (SVM), k-Nearest Neighbors (KNN), Decision Trees (DT), Random Forest (RF), and ANN. Among these, ANN demonstrated higher performance, achieving an accuracy rate of 84.44% and an F1-score of 82.05%. Although the authors provided detailed information on the architecture, loss function, and activation functions employed in the ANN, there was a lack of disclosure regarding the optimizer type and evaluation method used.

In a recent investigation [19], a novel hybrid approach was introduced, incorporating KNN, SDA, NCA, and Fuzzy ANN. The data preprocessing was performed, involving the utilization of KNN (K-Nearest Neighbors) and SDA (Self Diagnosis Algorithm) to address missing values. For enhanced precision, NCA (Neighborhood Component Analysis) and PCA (Principal Component Analysis) were implemented. Disease diagnosis was conducted using Fuzzy ANN, and the Cleveland and Statlog datasets were employed. However, it is noteworthy that the evaluation method employed for the model was hold-out, and the article lacked specific details, including information on the optimizer type and hold-out rate.

Barfungpa et al. [20] utilized Heart Disease(HD) datasets, encompassing Statlog, Hungary, Cleveland, Switzerland, and Long Beach VA datasets. Data preprocessing involved the removal of noisy data, the substitution of missing values with mean

computation, and normalization using z-score or zero mean normalization. Feature selection was conducted through the Enhanced Sparrow Search Algorithm (E-SSA). The classifier chosen was the Deep Dense Residual Attention Aquila Convolutional Network (Deep-DenseAquilaNet). The proposed model demonstrated high performance in terms of accuracy, f-measure, sensitivity, specificity, and precision. While the authors provided comprehensive details about the classifier network employed, the model evaluation was performed using the hold-out method.

The literature includes various studies that aim to develop a robust solution for diagnosing CAD. Table 2.1 displays the summary of relevant research used int the literature for diagnosing CAD. These studies explore different approaches, including feature selection, feature extraction, preprocessing steps, and optimization techniques to enhance model performance. Despite these efforts, challenges such as low detection rates and high computational times still exist. Moreover, while the utilization of hybrid optimization algorithms is rare in studies, hyperparameter optimization has not been applied in most cases.

In this study, we introduce a new hybrid classifier named CSA-PSO-ANN, which leverages two optimization algorithms to use the strengths of both techniques. These algorithms are applied during the training phase of the Artificial Neural Network (ANN) to achieve enhanced performance outcomes. The hyperparameter of the CSA-PSO-ANN model is fine-tuned using Bayesian optimization, and the performance of the model is assessed through ten-fold cross-validation. Additionally, we reduce the computational time of our proposed hybrid method by employing CPU/GPU parallelization. This strategy improves diagnostic rates and also reduces the computational time of the model.

Table 2.1 A summary of relevant research used in the literature for diagnosing CAD.

Paper	Method	PP	FS	CV	Opt.	Hybrid Opt.	Dataset	Year
[17]	NN-DEGI-BP	✓	-	10	✓	✓	Cleveland	2016
[9]	FS with PSO and NB	-	✓	-	-	-	Statlog	2018
[12]	MLP	✓	✓	20	✓	-	Cleveland	2018
[13]	MLP	✓	✓	10	✓	-	Cleveland	2019
[14]	Ensemble	✓	✓	10	✓	-	Cleveland	2019
[11]	PSO-EmNN	-	✓	10	✓	-	Statlog	2020
[11]	PSO-EmNN	-	✓	10	✓	-	Cleveland	2020
[10]	PM-LU-NN	✓	-	-	✓	✓	Statlog	2020
[16]	CSA-KNN	-	-	-	✓	-	Private	2021
[8]	MLP-PSO	✓	-	5	✓	-	Cleveland	2022
[15]	MLP	✓	✓	10	✓	-	Statlog	2023
[15]	MLP	✓	✓	10	✓	-	Cleveland	2023
[18]	ANN	✓	-	-	✓	-	Cleveland	2023
[19]	KNN-SDA-NCA-Fuzzy ANN	✓	✓	-	-	-	Cleveland	2023
[19]	KNN-SDA-NCA-Fuzzy ANN	✓	✓	-	-	-	Statlog	2023
[20]	Deep-DenseAquilaNet	✓	✓	-	✓	-	HD Datasets	2023
PM	CSA-PSO-ANN	✓	✓	10	✓	✓	Cleveland	2023
PM	CSA-PSO-ANN	✓	✓	10	✓	✓	Statlog	2023

Chapter 3

Material and Methods

3.1 Datasets

This study utilized two commonly found datasets, namely Cleveland [21] and Statlog [22], which are frequently encountered in existing research literature. Both datasets classify each sample as either having Coronary Artery Disease (CAD) or being healthy. The Cleveland and Statlog datasets are widely utilized as popular datasets for heart disease research.

The Cleveland dataset has 303 samples and 14 features, with a detailed feature description provided in Table 3.1. It contains six missing samples, which are excluded instead of filled to avoid any adverse impact on the model's performance. Similarly, the Statlog dataset, with 270 samples and 14 features, closely resembles the Cleveland dataset. However, the Statlog dataset does not have any missing samples.

The datasets are thoroughly examined regarding their complexity, imbalanced ratio, Fisher's discriminant ratio, and sparsity. The findings are depicted in the Figure 3.1. A low F1 value indicates a lack of distinguishing features between different classes, suggesting intricate issues. The F1 results suggest that Statlog presents a more linear problem compared to Cleveland. N1 denotes complexity and measures the separability of class distributions [15]. Typically, an inversely proportional relationship is expected between F1 and N1 values, but the figure illustrates that the Statlog dataset is more complex than the Cleveland dataset. Lastly, the Statlog dataset demonstrates higher sparsity compared to the Cleveland dataset, and both datasets exhibit balance.

Table 3.1 The feature description in the Cleveland dataset

Feature name	Feature type	Description
Age	Numeric	Patients' age
Sex	Categorical	Gender of the patient (M: Male, F: Female)
Exang	Categorical	Exercise induced angina (Yes, No)
Ca	Numeric	The count of major vessels (0-3)
Cp	Categorical	Chest pain type (Typical angina, Atypical Angina, Non-Anginal pain, Asymptomatic)
Tretbps	Numeric	Resting blood pressure (in mm Hg)
Chol	Numeric	Cholestoral in mg/dl
Fbs	Categorical	Fasting blood sugar levels exceeding 120 mg/dl (1 indicates true; 0 indicates false)
Restecg	Categorical	Results of resting electrocardiogram
Thalach	Numeric	Peak heart rate reached
Slope	Categorical	Incline of the ST segment during peak exercise (Upsloping, Flat, Downsloping)
Thall	Categorical	Normal, Fixed defect, Reversable defect
Oldpeak	Float	ST segment depression during exercise compared to at rest
Output	Integer	1 indicates a higher likelihood of CAD 0 indicates a lower likelihood of CAD

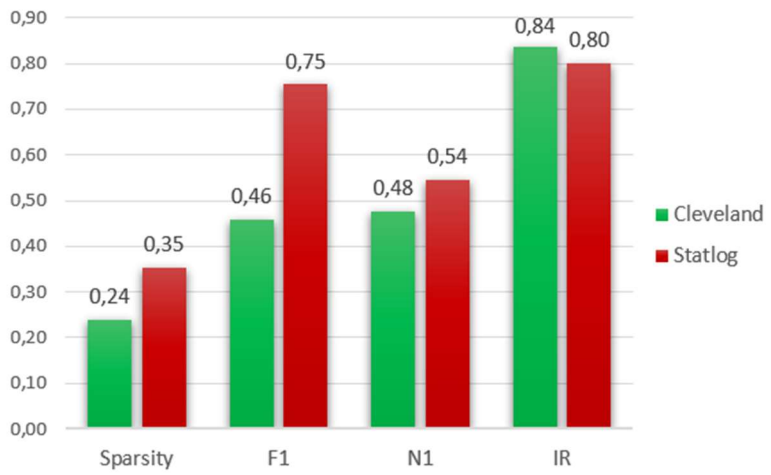

Figure 3.1 A comprehensive analysis of datasets

3.2 Data Preprocessing

In this thesis, for the attainment of more accurate results with machine learning algorithms, data preprocessing is applied after a comprehensive dataset analysis is conducted. This process involves the conversion of categorical data into numerical format, scaling the data, and selecting more pertinent features for machine learning models.

3.2.1 Label encoding

The datasets employed in this study both consist of categorical variables. To enhance the efficiency of machine learning algorithms in utilizing the data, categorical information is transformed into numerical data using methods like label encoding. Label encoding involves assigning a unique numeric value to each distinct categorical entry, effectively converting non-numerical values into numeric equivalents. In the scope of this study, label encoding was implemented on both datasets to improve result accuracy.

3.2.2 Normalization

The aim of normalization is to remove the influence of varying scales among features, allowing machine learning algorithms to treat all features equally during model training. With this purpose, min-max scaling is applied to both datasets. Normalization, particularly min-max scaling, is a data preprocessing technique applied to ensure that numerical features within a dataset are on a consistent scale. In the context of min-max scaling, the transformation involves adjusting the values of each feature proportionally

within the range of 0 to 1. This is accomplished by subtracting the minimum value of the feature from each data point and subsequently dividing this outcome by the range, which is the difference between the maximum and minimum values of the feature.

3.3 Feature Selection

Feature Selection is applied to enhance the effectiveness and value of features for classification models while mitigating the risk of overfitting. Researcher examined a number of feature selection methods for the diagnosis of CAD in [12],[13],[14], [15]. In this study, the chi-squared feature selection method is chosen, utilizing the SelectKBest function. First, p-values are computed for each attribute. High p-values suggest that the corresponding attributes are independent of the response variable. Consequently, attributes with high p-values may be removed, contributing to a more refined feature set for improved model performance. Figure 3.2 shows the calculated p-values for both datasets. The highlighted values in the table signify the features that are removed from the datasets. Consequently, for the Statlog dataset, nine features are selected, while for the Cleveland dataset, eleven features are chosen using the chi-squared feature selection method. These selected features are detailed in Table 3.2.

	Cleveland	Statlog
age	2.034358e-06	1.712223e-05
sex	6.404399e-03	5.487089e-03
cp	9.981719e-05	2.611301e-04
trestbps	5.823823e-05	7.131512e-05
chol	1.067196e-06	2.421437e-10
fbs	6.848935e-01	8.045311e-01
restecg	3.279787e-03	3.202088e-03
thalach	7.097172e-42	1.267379e-38
exang	6.886243e-10	1.689269e-08
oldpeak	4.436474e-17	1.838950e-14
slope	4.088886e-03	6.890083e-03
ca	1.689233e-19	7.341109e-18
thal	5.971029e-16	1.246094e-14

Figure 3.2 P-values for Cleveland and Statlog datasets

Table 3.2 The selected best features using Chi-Squared feature selection method

No	Cleveland	Statlog
1	Age	Age
2	Cp	Cp
3	Trestbps	Trestbps
4	Chol	Chol
5	Restecg	Thalach
6	Thalach	Exang
7	Exang	Oldpeak
8	Oldpeak	Ca
9	Slope	Thal
10	Ca	-
11	Thal	-

3.4 Parallel Computing

In a machine learning model, the computational time during the training phase can often be time-consuming. An effective solution to mitigate this constraint is the application of parallel computing, a method where various segments of substantial tasks are executed simultaneously. The parallelization process involves the vectorization of

codes, a technique eliminating loops to enhance computational efficiency. Subsequently, parallel computation is automatically implemented using either the Numpy or CuPy libraries.

For CPU parallelization in this study, the Numpy library is employed. Additionally, GPU parallelization is executed using the open-source library CuPy. This adaptable approach allows for a choice between CPU and GPU parallelization based on the specific requirements of the task.

3.5 Evaluation Method

In the literature, it is observed that machine learning approaches commonly employ either the hold-out method or the cross-validation method in data training and testing processes. In the hold-out method, a random subset of the data is used to train the model, while the remaining portion is reserved for testing. Many studies have adopted this approach (such as [9], [10], [16], [18], [19]). However, due to the random selection of data in this method, the model's performance on unseen data may not be accurately measured. On the other hand, the cross-validation method divides the data into k different folds, using each fold sequentially for testing while the remaining folds are utilized in training the model. This process continues for the specified number of folds, and the averages of the results are computed. Consequently, the data is effectively utilized, enabling a more accurate prediction of the model's performance on unseen data.

In this study, the 10-fold cross-validation method was preferred to assess the performance of the proposed model. The data was divided into ten equal parts, and for each iteration, training was conducted with nine folds, followed by testing with the remaining one fold. This allowed the model to process all the data in the dataset, and the average of the obtained results was calculated to evaluate the model's overall performance.

3.6 Classification Methods

The concept of classification involves constructing models or classifiers to predict categorical labels. These categories are represented by discrete values without implied ordering. For example, in the prediction of the appropriate treatment for breast cancer

patients, the values 1, 2, and 3 might be assigned to treatments A, B, and C [23]. The process of classifying data involves two fundamental stages: a learning phase, wherein a classification model is formulated, and a classification phase, wherein the model is employed to forecast class labels for provided data. During the learning phase, a classifier is established by examining a training set containing dataset tuples and their corresponding class labels. Subsequently, in the second phase, the model is applied for classification, and its accuracy is gauged using an independent test set.

In this study, several widely used classification methods for diagnosing CAD are employed to assess the performance of the proposed method in comparison with these well-known approaches.

3.6.1 Logistic regression

Logistic Regression is a machine learning algorithm frequently employed in binary classification problems. Essentially, it is a regression technique where the dependent variable is modelled through a logistic function based on a linear combination. The outcome of logistic regression provides a probability value between 0 and 1, typically utilized for classification by applying a threshold value.

3.6.2 Decision tree

A decision tree is a hierarchical structure where each node indicates a test on an attribute, branches depict test outcomes and leaf nodes hold class labels. This tree hierarchically makes decisions by partitioning the dataset, and each leaf node assigns the data to a specific class. Decision tree classifier based on this structure and it is popular for their simplicity, interpretability, and ability to handle multidimensional data. They don't require domain knowledge for construction, making them suitable for exploratory knowledge discovery. Also, the learning and classification steps are fast.

3.6.3 Random forest

Random Forest is a approach that makes a stronger classifier by putting together several decision trees. Instead of training each tree on all features, it trains them on randomly chosen subsets of features. Then, the predictions from all trees are combined to get more dependable results. This helps prevent individual trees from fitting too closely to the data. By training each tree on different feature subsets, Random Forest adds variety

to the model, making it better at generalizing. This way, the model is less affected by noise and can give more accurate predictions compared to just one decision tree.

3.6.4 Support vector machine

Support Vector Machines (SVMs) represent a method of classification applicable to both linear and nonlinear data. Essentially, an SVM is an algorithm that operates by employing a nonlinear mapping to transform the initial training data into a higher-dimensional space. In this expanded space, the approach aims to find the optimal linear separating hyperplane, essentially a "decision boundary" that distinguishes tuples of one class from another. The SVM identifies this hyperplane by leveraging support vectors, which are essential training tuples, along with the margins determined by these support vectors. In summary, SVM constructs a hyperplane to classify data points, optimizing it to maximize the margin between classes.

3.6.5 K- nearest neighbor

This method is frequently utilized in classification tasks and involves comparing each data point with all others in the dataset. It identifies the closest neighbours of each data point using a specified distance metric. The classification of a data point is subsequently based on the majority class label among its nearest neighbours. Throughout the classification process, the algorithm takes into account the labels of the K nearest neighbours, with K being a user-defined parameter indicating the number of neighbours to consider. The final classification label for the given data point is determined by the predominant class label among these K neighbours.

3.6.6 Multi layer perceptron

Multi-layer perceptrons represent a model in the realm of deep learning that utilizes artificial neural networks. These networks, consisting of multiple layers such as an input layer, one or more hidden layers, and an output layer, link neurons with weights that are defined by the outputs of neurons in the previous layers. The commonly employed backpropagation algorithm serves the purpose of training the network. Multi-layer perceptrons may offer considerable perfomance for dealing with intricate problems.

3.7 Hyperparameter Optimization

Considering the number and diversity of hyperparameters (e.g., discrete, continuous), manually tuning them is a challenging process. Furthermore, hyperparameter optimization algorithms can yield reproducible results under the same conditions. This facilitates a systematic evaluation of changes made to improve the model's performance. In this study, the Bayesian hyperparameter optimization algorithm using the "hyperopt" library [24] is preferred to tuning hyperparameters of the proposed model. Bayesian optimization is based on a probability model for P(score|configuration), which is derived by updating a prior using a history H of (configuration, score) pairs [25]. Therefore, obtaining the optimal hyperparameter usually takes less time. By the way, the Hyperopt library gives the opportunity to save more statistical and diagnostic information. Thanks to this, it is possible to store optimization results along with additional information in a text file throughout the conducted experiment.

3.8 Performance Metrics

To assess the performance of classification models, the confusion matrix is utilized. Many metrics and ratios are computed using the confusion matrix. It is created by comparing the values predicted by a model with the actual values. Fundamentally, it consists of 4 components: True Positive, True Negative, False Positive, and False Negative.

True Positive: The model accurately identifies positive values.

True Negative: The model accurately identifies negative values.

False Positive: The model erroneously classifies a negative value as positive.

False Negative: The model erroneously classifies a positive value as negative.

Among these components, True Positive and True Negative signify correct classifications made by model, whereas False Positive and False Negative represent incorrect classifications.

3.8.1 Accuracy

Accuracy, a measure of performance, reflects the proportion of accurate predictions in the confusion matrix compared to all predictions. It's computed by dividing the total of

True Positive and True Negative by the total predictions. In the literature, it's regarded as a key metric for assessing the effectiveness of a classification model.

3.8.2 F1 Score

F1 score is a leading metric preferred to assess classification performance. This measurement is obtained using the results of two different performance metrics called precision and recall. Recall measures the proportion of positive values correctly predicted by the model against the actual positive values. Precision, on the other hand, expresses the ratio of correctly predicted values by the model to all the values it predicted positively (True Positive and False Positive).

The F1 score is calculated by taking the harmonic mean of recall and precision values. This ensures a balanced contribution of both metrics, reflecting a model with good performance in both precision and recall when achieving a high F1 score.

Chapter 4

Proposed Method and Its Component

4.1 Clonal Selection Algorithm

The Clonal Selection Algorithm, belonging to the category of Artificial Immune Optimization (AIO) techniques [26], functions on the fundamental premise that organisms demonstrate an adaptive immune response when confronted with a harmful stimulus [27]. Within this algorithmic framework, antibodies undergo a replication process based on their affinity subsequent to undergoing an affinity maturation phase. The objective is to obtain antibodies characterized by the highest affinity levels. The COLONALG algorithm, introduced by Castro and Von Zuben [26], is rooted in the foundational principles of clonal selection. The version of the CLONALG algorithm employed in this particular study for optimization purposes is delineated below, encompassing a sequence of operations including selection, cloning, hyper-mutation, and receptor-editing, each performed sequentially.

The population of antibodies, denoted as $Ab=Ab_1, Ab_2, Ab_3, \ldots Ab_P$, representing the population, is generated using Eq. 4.1.1. Each antibody is represented by a vector of size D, where $Ab_i=[Ab_{i,1}, Ab_{i,2}, Ab_{i,3}, \ldots Ab_{i,D}]$, signifying a potential solution.

$$Ab_i = lb_i + rand(0,1) \times \left(ub_j - lb_j\right) \quad 4.1.1$$

The generation of each antibody is determined by Eq. 4.1.1, where lb_j and ub_j signify the lower and upper bounds for each antibody, and rand(0,1) generates random numbers between 0 and 1.

$$f\left(Ab_{ij}\right) = \frac{1}{1 + J\left(Ab_{ij}\right)} \quad 4.1.2$$

After establishing the antibody population, the fitness value for each antibody is computed using Eq. 4.1.2, where $f(Ab_{i,j})$ represents the fitness function, and $J(Ab_{i,j})$ indicates the cost function. In accordance with Eq. (4.1.2), the optimization aims to achieve higher fitness values by minimizing the cost function.

To facilitate the exploration of multiple optima [26], the clone size (C) is determined by multiplying the population size (P) by the clone factor (alpha). The hyper-mutation process is then applied to produce a matured antibody population from the generated clones, indicated as $C_i=[C_1, C_2, C_3, ... C_4]$. This process involves inverse mutation and pair-wise mutation, where σ represents the mutated clones. The algorithm checks whether the affinity value for C_i is less than the affinity value for σ_i, and if so, C_i is replaced by σ_i.

The final step of the algorithm is the receptor-editing phase, which maintains a constant population size. Antibodies exhibiting lower affinity are eliminated from the population, and the algorithm proceeds with newly generated ones until the stop criteria are met.

This optimization algorithm conducts a global search during the receptor-editing phase and a local search during the hyper-mutation phase.

4.2 Particle Swarm Optimization

Introduced in 1995 by Kenney and Eberhart, Particle Swarm Optimization (PSO) draws inspiration from the coordinated movements observed in fish and bird swarms [28]. In this approach, individual entities, referred to as particles, traverse the problem space in an effort to identify the optimal solution. Each particle possesses its own position and initial velocity, with subsequent positions determined by both individual and neighbouring experiences as they traverse the search space [29]. The position of each particle corresponds to a potential solution, and the algorithm involves several key equations.

The next position of a particle, denoted as $x_i(t+1)$, is calculated using Eq. 4.2.1.

$$x_i(t+1) = x_i(t) + v_i(t+1) \quad 4.2.1$$

Here, $x_i(t)$ and $v_i(t+1)$ represent the current position and velocity for the next iteration at time step t.

Velocity, a crucial factor determining the optimization direction, is composed of cognitive and social parameters. Eq. 4.2.2 is employed to calculate the velocity of particle i.

$$v_{i,j}(t+1) = wv_{i,j}(t) + c_1 r_1(t)\left[y_{i,j} - x_{i,j}\right] + c_2 r_2\left[y_j(t) - x_{i,j}(t)\right] \qquad 4.2.2$$

Here, j represents the iteration number, w is the inertia weight, r_1 and r_2 are uniformly distributed values between 0 and 1, and c_2 and c_2 indicate the cognitive and social components, respectively. The $y_{i,j}$ denotes the best position from the beginning, while y_i represents the best position in the neighbourhood of particle i.

The PSO algorithm develops through several steps. Initially, the swarm size (S) and control parameters are set, and the best position for each particle is determined. Both the best individual and global positions are identified. Subsequently, the velocity for each particle is updated using Eq. 4.2.2, and the position is updated using Eq. 4.1.1. These processes iteratively continue until the specified stop criterion is met.

4.3 Artificial Neural Network

An artificial neural network (ANN) serves as a computational model, comprising interconnected computational units, or neurons, organized into layers. The fundamental process involves the transmission of signals from one neuron to another for information processing.

Within an ANN, each neuron positioned between layers receives input from other neurons, undertakes a basic mathematical operation on the received input, and subsequently transmits the result to other connected neurons. The mathematical equation governing this process is expressed in Eq. 4.3.1.

$$y = f\left(\sum_{j=1}^{n} w_j x_j + b\right) \qquad 4.3.1$$

The ANN employed in this thesis comprises a configuration that includes one input layer, one hidden layer, and one output layer, as depicted in Figure 4.1. This ANN functions as a classification model aimed at detecting coronary artery disease. During the

learning phase of the model, the proposed hybrid optimization algorithm (CSA-PSO) is implemented as an alternative to the conventional backpropagation algorithm.

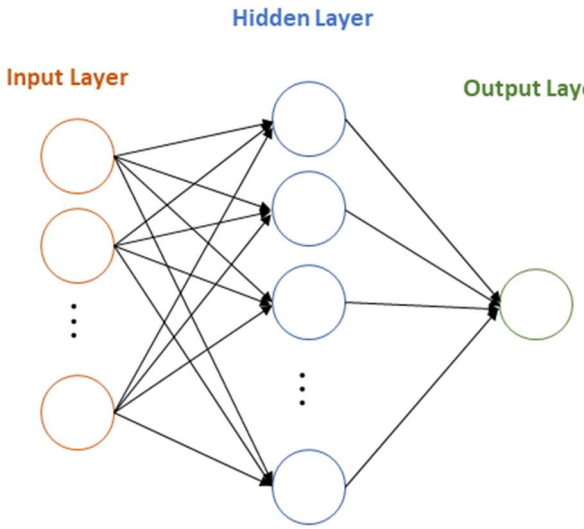

Figure 4.1 The architecture of the ANN

4.4 Proposed Hybrid Method (CSA-PSO-ANN)

This research introduces a CSA-PSO hybrid optimization algorithm designed for the training of an artificial neural network classifier. The algorithm integrates the Clonal Selection Optimization Algorithm (CSA) with the local search capabilities of the Particle Swarm Optimization Algorithm (PSO). The objective behind this integration is to identify optimal antibodies within the search space, indicative of potential solutions, by enhancing CSA's local search capability. Consequently, the overarching aim is to elevate CSA's proficiency in local search and, in turn, achieve more efficient training for neural networks.

After the cloning phase, the antibodies undergo hyper-mutation within the framework of CSA. In this intricate stage of hyper-mutation, precise operations such as inverse mutation and pair-wise mutation are executed. These operations result in the creation of a mutated clone denoted as σ_i , utilizing the parameters derived from the original clones.

The subsequent phase, which is called the affinity maturation process, is characterized by a comparison of the affinity values between the clones and σ_i. However,

this hypermutation step might impose limitations on the algorithm's search capacity for complex optimization problems [6].

In order to enhance the solution's robustness, the hyper-mutation component of CSA is excluded, and instead, the velocity component of PSO is incorporated. The velocity component, calculated according to Eq. 4.2.2, is employed in the affinity maturation process, replacing the conventional inverse and pair-wise mutation methods. This utilization of the velocity component serves to improve the identification of superior antibodies within the CSA framework. The sequential steps of the proposed algorithm are delineated as follows.

Initially, control parameters such as w, c_1, c_2, lb, ub, P, the evaluation number, hidden layer size, B, and α are set. Following this, P antibodies are generates by using the Eq. 4.1.1. The affinity value for each antibody is calculated using another equation(Eq. 4.1.2). The main iterative loop of the algorithm involves generating clones of the antibodies and subsequently updating their positions based on computed velocity components(Eq. 4.2.2). Each clone is mutated and the fitness of the mutated clone(C_i^*) is compared the original clone(C_i). If the mutated clone is fitter, the original clone is updated; otherwise, it remains unchanged. Furthermore, for each antibody in the population, the algorithm selects the clone with the highest affinity among its clones, effectively updating the antibody with a more proimisng variant. Additionaly, a certain percentage(B%) of the least fit antibodies is replaced with newly created antibodies. This dynamic replacement strategy aims to continuously improve the overall population fitness.

In the process of training Artificial Neural Networks (ANNs), the CSA-PSO approach proposed in this study aims to minimize the cost function. Specifically, Mean Absolute Error (MAE), as defined in Eq. 4.4.1, is employed as the selected cost function. MAE measures the cost by calculating the absolute difference between actual and predicted values, thereby preventing negative values. Additionally, MAE exhibits reduced sensitivity to noise or outliers in datasets, mitigating the impact of such anomalies.

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|y_i - y_i'| \qquad 4.4.1$$

Here, y_i represents actual values, and y_i' represents predicted values. The identified cost function plays a pivotal role in computing fitness values for antibodies as outlined in Eq. 4.1.2. Consequently, the training of ANNs is executed through the combined optimization algorithm of CSA-PSO. Figure 4.2 represents the flowchart of the proposed model.

Algorithm 1 : The Proposed Hybrid CSA-PSO Model

1: *Set the control parameters (w, c_1, c_2, lb, ub, P, the evaluation number, the hidden layer size, B, and α)*
2: *Produce initial population of P antibodies utilizing Eq. 4.1.1*
3: *Find the affinity value for each antibody Ab utilizing Eq. 4.1.2*
4: ***for*** *each iteration* ***do***
5: *Generate the clones of the antibodies*
6: *Calculate the affinity values of these clones*
7: ***for*** *each clone* C_i ***do***
8: *Compute velocity component utilizing Eq. 4.2.2 for* C_i *:*
9: *Update the velocity of the clone Ci generate the mutated clone Ci* utilizing Eq. 4.2.1*
10: ***if*** $f(C_i^*) < f(C_i)$ ***then***
11: $C_i := C_i$
12: ***Else***
13: *Ci := Ci**
14: ***end if***
15: ***end for***
16: ***for*** *each antibody* Ab_i ***do***
17: *Select the clone Cj with highest affinity among the clones of Abi*
18: *Abi := Ci*
19: ***end for***
20: *Change the worst B% number of antibodies with the newly created ones.*
21: ***end for***

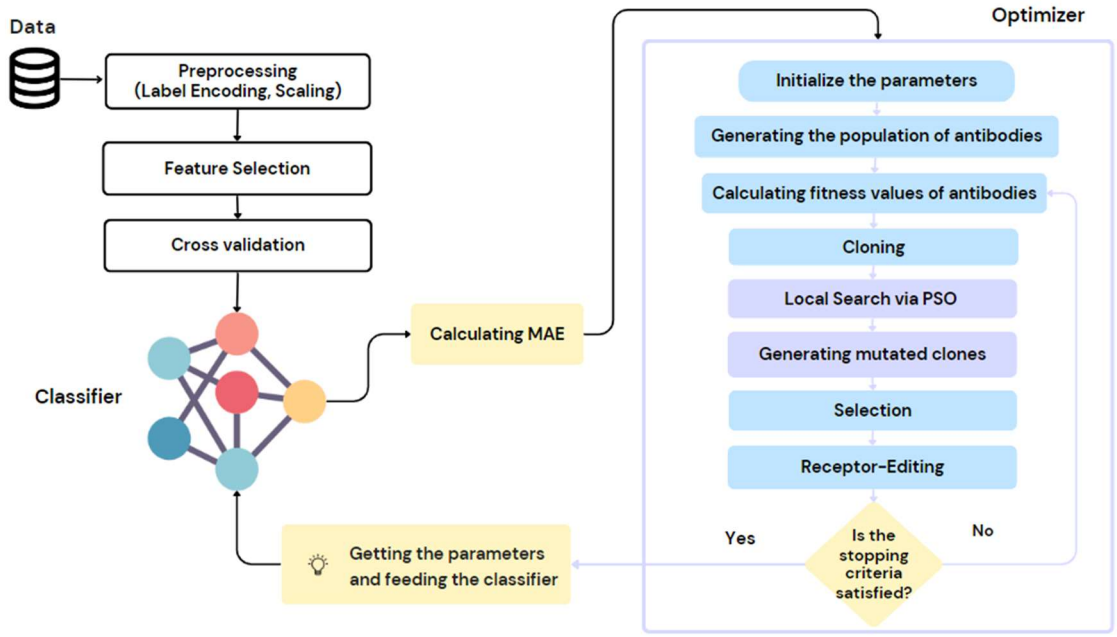

Figure 4.2 The flowchart of the proposed CSA-PSO ANN model.

Chapter 5

Experiments

This experiment aims to assess the effectiveness of the proposed method in diagnosing coronary artery disease (CAD) using two publicly accessible datasets. The experiment is conducted in several stages including data preprocessing, feature selection, model training, hyperparameter optimization, and performance evaluation.

Two datasets, namely the Statlog and the Cleveland datasets, are chosen for this investigation. These are the most commonly preferred datasets in the domain of cardiovascular health. Before experiments, a thorough examination of both datasets is carried out to understand their characteristics, such as complexity, imbalanced ratio, Fisher's discriminant ratio, and sparsity. Instances with missing values are excluded from the datasets to ensure data integrity. Subsequently, label encoding and min-max scaling techniques are applied to normalize the features and bring them within a consistent range.

Feature selection plays a crucial role in improving the performance of machine learning models by eliminating irrelevant or redundant features. In this experiment, the chi-square feature selection method, a filter-based technique, is employed to select the most informative features. For the Statlog dataset, nine features are selected, while eleven features are chosen for the Cleveland dataset based on their discriminative power.

Various machine learning algorithms are considered for CAD diagnosis, including logistic regression(LR), decision trees(DT), random forests(RF), support vector machines (SVM), k-nearest neighbours (KNN), multi-layer perceptrons (MLP). A novel approach is introduced by training an ANN using a hybrid CSA-PSO optimization algorithm. This algorithm focuses on minimizing mean absolute error, which is critical for accurate diagnosis.

The proposed method has ten different hyperparameters. To fine-tune the parameters of the models and enhance their performance, the Bayesian hyperparameter optimization technique is employed. Optimal hyperparameters for all models are obtained and listed in Table 5.1. This step ensures that the models are configured optimally for the given task.

To evaluate the performance of the proposed method, 10-fold cross-validation is applied. This technique helps in obtaining reliable performance estimates by partitioning the dataset into 10 subsets, training the model on 9 subsets, and evaluating it on the remaining subset. Accuracy and F-measure are selected as evaluation metrics to assess the classification performance comprehensively.

The entire experiment is implemented using Python programming language. Parallel computing using the NumPy library is incorporated to expedite computations, thereby reducing the overall computational time. The CSA-PSO-ANN algorithm is developed from scratch, ensuring full control over the optimization process and its integration with the neural network architecture.

In addition to the proposed method, the experimental steps outlined above are also applied to other well-known classification methods for comparative analysis. Logistic regression, decision trees, random forests, support vector machines (SVM), k-nearest neighbours (KNN), and multi-layer perceptrons (MLP) undergo the same preprocessing, feature selection, parameter optimization, and evaluation procedures.

Following the data preprocessing steps, which involve handling missing values, scaling features, and encoding categorical variables, the results are obtained utilizing all features present in the datasets. This initial phase ensures that each algorithm is provided with consistent and properly formatted data to facilitate meaningful comparisons. Subsequently, feature selection techniques are employed to identify the most relevant features for model training and prediction. By narrowing down the feature set to those deemed most informative, the subsequent execution of algorithms is streamlined, potentially improving computational efficiency and model performance. This meticulous approach aims to ensure robustness and reliability in the evaluation of each method's effectiveness in diagnosing coronary artery disease (CAD), thus contributing to a comprehensive understanding of their relative merits across various machine learning paradigms.

Table 5.1 Hyperparameters obtained with Bayesian Optimization

Method	Parameter Range
LR	penalty : [l1, l2]
	solver : [liblinear, saga]
	tol : [10^{-6}, 10^{-2}]
DT	maxdepth : [1,10]
	minsamplesplit : [0,1]
	minsamplesleaf : [0, 0.5]
	maxfeatures : [sqrt, log2, None]
RF	nestimator : [10, 1000]
	maxdepth : [1, 10]
	minsamplesplit : [0, 1]
	minsamplesleaf : [0, 0.5]
	maxfeatures: [sqrt, log2, None]
SVM	kernel : [linear, poly, rbf, sigmoid]
	gamma : [scale, auto]
KNN	neighbours : [1, 21]
	weights : [uniform, distance]
	algorithm : [auto, baltree, kdtree, brute]
	p : [1, 2]
MLP	activation : [logistic, relu]
	alpha : [10^{-6}, 1]
	learningrateinit : [10^{-6}, 1]
	maxiter : [20, 400]
	solver : [sgd, adam]
	hiddenlayersize : [5, 10, 20]
CSA-PSO-ANN	Lb: [-64, -16]
	Ub: [16, 64]
	evaluationNumber : [80000, 165000]
	B: [0.05, 0.2]
	Alpha: [1, 7]
	P : [10, 80]
	C1 : [0.5, 2.1]
	C2 : [0.5, 2.1]
	W : [0.4, 0.9]
	HLS : [2, 20]

Chapter 6

Results

The experimental results, utilizing all features in the dataset, are presented in Table 6.1. Among the various algorithms tested, CSA-PSO-ANN emerged as the top performer in terms of accuracy for both the Cleveland and Statlog datasets. The results indicate notable differences in accuracy across the algorithms tested. For the Cleveland dataset, the highest accuracy was achieved by the CSA-PSO-ANN method with a score of 86.66%, outperforming LR (85.46%), DT (82.52%), RF (84.81%), SVM (85.48%), KNN (83.84%), MLP (85.46%). Similarly, for the Statlog dataset, CSA-PSO-ANN exhibited the highest accuracy of 87.40%, surpassing LR (85.93%), DT (81.48%), RF (85.56%), SVM (85.19%), KNN (83.70%), MLP (86.30%). These results suggest that CSA-PSO-ANN is a promising algorithm for accurate classification tasks, indicating its potential for practical application in both datasets.

When incorporating selected features, CSA-PSO-ANN exhibited remarkable performance on the coronary artery disease datasets. For the Cleveland dataset, it achieved an accuracy of 87.33%, an F1 score of 85.23%, and a computational time of 2.68 seconds. Similarly, for the Statlog dataset, it attained an accuracy of 88.14%, an F1 score of 85.39%, and a training time of 2.42 seconds, as elaborated in Tables 6.2 and 6.3.

Analyzing Table 6.2, it's evident that on the Cleveland dataset, other classification algorithms with feature selection yielded varying degrees of accuracy. Notably, CSA-PSO-ANN, with carefully tuned parameters such as B: 0.15, P: 60, alpha: 5, and a hidden layer size of 5, outperformed LR, DT, RF, KNN, and MLP in terms of both F1 score and accuracy. Its runtime of 2.68 seconds also competes well with other methods, demonstrating its efficiency.

Similarly, examining Table 6.3, on the Statlog dataset, CSA-PSO-ANN, with parameter settings including B: 0.16, P: 18, alpha: 2, and a hidden layer size of 4, again showcased superior performance. Compared to LR, DT, RF, SVM, KNN, and MLP, CSA-PSO-ANN achieved the highest accuracy and F1 score. Its runtime of 2.42 seconds indicates efficient computation, further solidifying its suitability for practical applications. These results underscore the robustness and efficacy of CSA-PSO-ANN in handling feature-rich datasets for accurate classification tasks, making it a promising choice for real-world implementations in medical and statistical domains.

Table 6.4 presents a comparative analysis of our proposed model alongside state-of-the-art methods from previous studies on both datasets. To ensure fair performance evaluation, we included methods in Table 6.4 employing similar techniques found in the literature and utilizing k-fold cross-validation for model assessment, excluding [16] due to its methodological similarities. Our hybrid optimization algorithm notably outperformed methods using only CSA [16] or only PSO [8]. Furthermore, compared to methods employing traditional ANN ([15]) and other hybrid approaches ([11], [17]), our CSA-PSO-ANN method demonstrated superior classification accuracy and F1 score. Despite the limited availability of training time information in most literature studies, the computational time of our model, as depicted in Tables 6.2 and 6.3, shows improvements underscoring the role of parallel computing in reducing computational overhead.

These findings align well with the results from Table 6.4, where our CSA-PSO-ANN method showcases remarkable performance. On the Cleveland dataset, our method achieved an accuracy of 87.33% and an F1 score of 85.23%, outperforming various approaches such as NN-DEGI-BP, MLP, and ensemble methods. Similarly, on the Statlog dataset, CSA-PSO-ANN attained an accuracy of 88.14% and an F1 score of 85.39%, surpassing PSO-Em-NN and other techniques. Additionally, the computational time of our model on both datasets, as illustrated in Tables 6.2 and 6.3, further solidifies its efficiency compared to prior methods, highlighting its potential for practical applications in real-world scenarios.

The methodology employed in our approach, encompassing pre-processing, feature selection, and hyperparameter optimization, was also adopted in other established methods. Notably, within the tables, the results specific to our proposed method are highlighted in bold. CSA-PSO-ANN consistently outperformed other well-known

methods in terms of accuracy for both datasets. Particularly noteworthy is its superiority over the traditional MLP model with the stochastic gradient descent optimizer, which lagged behind in accuracy, and F1 score. This underscores the potential for enhancing the exploratory capabilities of ANN through powerful optimization algorithms.

Table 6.1 Performance measure of the other classification algorithms on both datasets without Feature Selection

Dataset	Methods	RT(sec.)	F1(%)	ACC(%)
Cleveland	LR	0.05	84.49	85.46
	DT	**0.04**	73.97	82.52
	RF	28.88	83.39	84.81
	SVM	0.11	**85.25**	85.48
	KNN	0.06	83.48	83.84
	MLP	1.6	81.78	85.46
	CSA-PSO-ANN	0.96	84.37	**86.66**
Statlog	LR	0.06	85.06	85.93
	DT	**0.05**	77.62	81.48
	RF	2.35	84.99	85.56
	SVM	0.15	84.75	85.19
	KNN	0.18	83.12	83.70
	MLP	1.16	84.70	86.30
	CSA-PSO-ANN	1.07	**85.21**	**87.40**

Table 6.2 Performance measure of other classification algorithms on the Cleveland dataset with Feature Selection

Methods	**Parameters**	**RT(sec)**	**F1(%)**	**ACC(%)**
LR	penalty: l1, solver: liblinear, tol: 0.07	0.06	84.51	85.78
DT	maxdepth: 7.0, maxfeatures: None, minsamplesleaf: 0.04, minsamplesplit:0.006	0.07	81.56	82.14
RF	maxdepth: 6.0, maxfeatures: 0, minsamplesleaf: 0.10, minsamplesplit: 0.18, nestimators: 200.0	6.62	83.58	85.44
SVM	gamma: scale, kernel: rbf	0.18	84.82	85.12
KNN	algorithm: bal tree, n_neigbours: 16, p:1, weights: uniform	0.08	84.09	84.49
MLP	activation: relu, alpha: 0.02, hiddenlayersize: 5, learningrateinit: 0.05, maxiter:78.0, solver:sgd	1.31	84.13	85.78
CSA-PSO-ANN	**B: 0.15, P:60, alpha: 5, evaluationNumber: 121304, lb: -63, ub:45, c1:1.6, c2: 1.01, w:0.85, hiddensize: 5**	**2.68**	**85.49**	**87.33**

Table 6.3 Performance measure of the other classification algorithms on the Statlog dataset with Feature Selection

Methods	Parameters	RT(sec)	F1(%)	ACC(%)
LR	penalty: l1, solver: liblinear, tol: 0.1	0.05	84	85.93
DT	maxdepth: 0.8, maxfeatures: log2, minsamplesleaf: 0.03, minsamplesplit:0.04	0.07	79.33	82.59
RF	maxdepth: 8.0, maxfeatures: 1, minsamplesleaf: 0.094, minsamplesplit: 0.32, nestimators: 70.0	3.35	83.25	85.19
SVM	gamma: auto, kernel: sigmoid	0.13	85.13	85.56
KNN	algorithm: auto, n_neigbours: 7, p:1, weights: distance	0.05	85.08	85.56
MLP	activation: logistic, alpha: 0.21, hiddenlayersize: 10, learningrateinit: 0.009, maxiter:48.0, solver:adam	0.92	84.43	86.30
CSA-PSO-ANN	**B: 0.16, P18, alpha: 2, evaluationNumber: 125756, lb: -47, ub:59, c1:1.13, c2: 1.43, w:0.74, hiddensize: 4**	**2.42**	**85.83**	**88.14**

Table 6.4 Comparison of the CSA-PSO-ANN proposed method with prior studies on both datasets

Paper	Method	K-Fold CV	F1 (%)	ACC (%)	RT (sec.)	Dataset	Year
[17]	NN-DEGI- BP	10	-	86.66	-	Cleveland	2016
[12]	MLP	20	83.80	82.50	4.01	Cleveland	2018
[14]	Ensemble	10	81.10	83.43	0.21	Cleveland	2020
[11]	PSO-EmNN	10	82.29	84	-	Cleveland	2020
[11]	PSO-Em-NN	10	84	85.20	-	Statlog	2020
[16]	CSA-KNN	-	-	78.40	-	Private	2021
[8]	MLP-PSO	5	84.40	84.60	-	Cleveland	2022
[15]	MLP	10	83.90	85.47	-	Cleveland	2023
[15]	MLP	10	85.30	85.55	-	Statlog	2023
PM	**CSA-PSO-ANN**	**10**	**85.23**	**87.33**	**2.68**	**Cleveland**	**2023**
PM	**CSA-PSO-ANN**	**10**	**85.39**	**88.14**	**2.42**	**Statlog**	**2023**

Chapter 7

Conclusions and Future Prospects

7.1 Conclusions

Cardiovascular diseases, especially coronary artery disease, continue to pose a significant global health challenge, and account for a substantial portion of total fatalities worldwide. Recently, there has been a growing interest in leveraging machine learning approaches as an alternative way of diagnosing this disease. Among these approaches, artificial neural networks (ANN) are commonly utilized classifiers, yet there is need for improvement in their classification capabilities.

This study introduces a novel method, CSA-PSO-ANN, which integrates the clonal selection algorithm (CSA) with particle swarm optimization (PSO) to enhance local exploration within the solution space. The ANN is trained using this combined approach. Additionally, a CPU-parallelized version of the method has been developed to significantly reduce training time, and optimal hyperparameters are automatically determined using Bayesian optimization. The model's performance is assessed through 10-fold cross-validation, resulting in accuracy values of 88.14% and 87.33% for the Statlog and Cleveland datasets, respectively. The outcomes indicate that this model improves the exploratory capabilities of the ANN, leading to superior classification performance compared to well-established algorithms and approaches found in previous studies.

7.2 Societal Impact and Contribution to Global Sustainability

In the contemporary landscape, the intersection of technological advancements and societal well-being is pivotal in addressing global challenges. The ongoing quest for sustainable solutions has brought forth innovative approaches that extend beyond immediate problem-solving, encompassing the broader scope of societal impact and global sustainability. This discussion investigates the ways in which emerging

technologies contribute significantly to shaping a sustainable future and fostering positive societal change.

One noteworthy aspect lies in the application of artificial intelligence (AI) and machine learning (ML) to pressing societal issues. These technologies have the potential to revolutionize sectors such as healthcare, education, and environmental conservation. For instance, AI-driven healthcare systems can enhance diagnostic accuracy, streamline treatment processes, and ultimately improve patient outcomes.

With this perspective, the proposed approach centres on diagnosing a widespread global cause of death, specifically, coronary artery disease. Detecting this ailment often involves complex and costly tests. By utilizing Artificial Intelligence technologies, especially Machine Learning algorithms, early identification of the disease becomes achievable. This, in turn, can result in a decrease in the number of sudden deaths linked to this condition. Moreover, given the typically expensive and inaccessible nature of diagnostic tests for this disease, Machine Learning-based tests offer a readily available diagnostic alternative. This thesis study focuses on enhancing the diagnostic capability of Artificial Neural Networks, the most commonly utilized machine learning techniques in coronary artery disease diagnosis. The method proposed in the study focuses on improving the detection performance of the Artificial Neural Network. Additionally, the study seeks to minimize the common waste of resources and time in training Machine Learning models by adopting two distinct parallel processing methods.

While navigating the complexities of the modern era, it is evident that technological innovation plays a pivotal role in shaping a sustainable and equitable future. By harnessing the potential of Artificial Intelligence (AI), Machine Learning (ML), and interconnected technologies, we can address societal challenges, promote inclusivity, and make a meaningful contribution to global sustainability.

7.3 Future Prospects

Further research can expand on the currently proposed method by gathering more extensive data and applying the suggested approach to diverse datasets. This will help assess its adaptability and effectiveness across various types of information. Moreover, improvements to this method can be achieved through alternative automatic hyperparameter tuning and feature selection techniques, opening the door for further refinement and optimization.

BIBLIOGRAPHY

[1] "Cardiovascular diseases-WHO." Accessed: Dec. 14, 2023. [Online]. Available: https://www.who.int/news-room/fact-sheets/detail/cardiovascular-diseases-(cvds)

[2] "Cardiovascular diseases." 2021. Accessed: Dec. 14, 2023. [Online]. Available: https://www.cdc.gov/heartdisease/coronary_ad.htm

[3] R. Alizadehsani *et al.*, "Machine learning-based coronary artery disease diagnosis: A comprehensive review," *Comput Biol Med*, vol. 111, p. 103346, 2019, doi: https://doi.org/10.1016/j.compbiomed.2019.103346.

[4] A. K. Jain, J. Mao, and K. M. Mohiuddin, "Artificial neural networks: a tutorial," *Computer (Long Beach Calif)*, vol. 29, no. 3, pp. 31–44, 1996, doi: 10.1109/2.485891.

[5] B. A. S. Emambocus, M. B. Jasser, and A. Amphawan, "A Survey on the Optimization of Artificial Neural Networks Using Swarm Intelligence Algorithms," *IEEE Access*, vol. 11, pp. 1280–1294, 2023, doi: 10.1109/ACCESS.2022.3233596.

[6] Y. Peng and B.-L. Lu, "Hybrid learning clonal selection algorithm," *Inf Sci (N Y)*, vol. 296, pp. 128–146, 2015, doi: https://doi.org/10.1016/j.ins.2014.10.056.

[7] X. Wang, X. Z. Gao, and S. J. Ovaska, "A Hybrid Particle Swarm Optimization Method," in *2006 IEEE International Conference on Systems, Man and Cybernetics*, 2006, pp. 4151–4157. doi: 10.1109/ICSMC.2006.384785.

[8] A. Al Bataineh and S. Manacek, "MLP-PSO Hybrid Algorithm for Heart Disease Prediction," *J Pers Med*, vol. 12, no. 8, 2022, doi: 10.3390/jpm12081208.

[9] U. N. Dulhare, "Prediction system for heart disease using Naive Bayes and particle swarm optimization," *Biomedical Research*, vol. 29, no. 12, pp. 2646–2649, 2018.

[10] R. P. Cherian, N. Thomas, and S. Venkitachalam, "Weight optimized neural network for heart disease prediction using hybrid lion plus particle swarm algorithm," *J Biomed Inform*, vol. 110, p. 103543, 2020, doi: https://doi.org/10.1016/j.jbi.2020.103543.

[11] A. H. Shahid and M. P. Singh, "A Novel Approach for Coronary Artery Disease Diagnosis using Hybrid Particle Swarm Optimization based Emotional Neural Network," *Biocybern Biomed Eng*, vol. 40, no. 4, pp. 1568–1585, 2020, doi: https://doi.org/10.1016/j.bbe.2020.09.005.

[12] B. Kolukisa *et al.*, "Evaluation of classification algorithms, linear discriminant analysis and a new hybrid feature selection methodology for the diagnosis of coronary artery disease," in *2018 ieee international conference on big data (big data)*, 2018, pp. 2232–2238.

[13] B. Koluksa, H. Haclar, M. Kuş, B. Bakr-Güngör, A. Aral, and V. Ç. Güngör, "Diagnosis of coronary heart disease via classification algorithms and a new feature selection methodology," *International Journal of Data Mining Science*, vol. 1, no. 1, pp. 8–15, 2019.

[14] B. Kolukisa *et al.*, "Coronary artery disease diagnosis using optimized adaptive ensemble machine learning algorithm," *Int J Biosci Biochem Bioinforma*, vol. 10, no. 1, pp. 58–65, 2020.

[15] B. Kolukisa and B. Bakir-Gungor, "Ensemble feature selection and classification methods for machine learning-based coronary artery disease diagnosis," *Comput*

Stand Interfaces, vol. 84, p. 103706, 2023, doi: https://doi.org/10.1016/j.csi.2022.103706.
[16] I. Gupta, R. Shangle, V. Latiyan, and U. Soni, "Cardiovascular Disease Detection using Artificial Immune System and other Machine Learning Models," in *Journal of Physics: Conference Series*, 2021, p. 12032.
[17] N. Leema, H. K. Nehemiah, and A. Kannan, "Neural network classifier optimization using Differential Evolution with Global Information and Back Propagation algorithm for clinical datasets," *Appl Soft Comput*, vol. 49, pp. 834–844, 2016, doi: https://doi.org/10.1016/j.asoc.2016.08.001.
[18] M. S. Gangadhar, K. V. S. Sai, S. H. S. Kumar, K. A. Kumar, M. Kavitha, and S. S. Aravinth, "Machine Learning and Deep Learning Techniques on Accurate Risk Prediction of Coronary Heart Disease," in *2023 7th International Conference on Computing Methodologies and Communication (ICCMC)*, 2023, pp. 227–232.
[19] R. Aggarwal and S. Kumar, "Classification model for meticulous presaging of heart disease through NCA using machine learning," *Evol Intell*, pp. 1–10, 2023.
[20] S. P. Barfungpa, H. K. D. Sarma, and L. Samantaray, "An intelligent heart disease prediction system using hybrid deep dense Aquila network," *Biomed Signal Process Control*, vol. 84, p. 104742, 2023.
[21] "UCI Cleveland Heart Disease Dataset." Accessed: Dec. 14, 2023. [Online]. Available: https://archive.ics.uci.edu/dataset/45/heart+disease
[22] "Statlog Heart Disease Dataset", Accessed: Dec. 14, 2023. [Online]. Available: https://archive.ics.uci.edu/dataset/145/statlog+heart
[23] J. Han, J. Pei, and H. Tong, "Data mining: concepts and techniques," in *Data mining: concepts and techniques*, Morgan kaufmann, 2011, pp. 327–439.
[24] "Hyperopt-Bayesian hyperparameter optimization library." Accessed: Dec. 14, 2023. [Online]. Available: http://hyperopt.github.com/hyperopt
[25] J. Bergstra, D. Yamins, and D. Cox, "Making a science of model search: Hyperparameter optimization in hundreds of dimensions for vision architectures," in *International conference on machine learning*, 2013, pp. 115–123.
[26] X. Wang, X. Z. Gao, and S. J. Ovaska, "Artificial immune optimization methods and applications - a survey," in *2004 IEEE International Conference on Systems, Man and Cybernetics (IEEE Cat. No.04CH37583)*, 2004, pp. 3415–3420 vol.4. doi: 10.1109/ICSMC.2004.1400870.
[27] L. N. de Castro and F. J. Von Zuben, "Learning and optimization using the clonal selection principle," *IEEE Transactions on Evolutionary Computation*, vol. 6, no. 3, pp. 239–251, 2002, doi: 10.1109/TEVC.2002.1011539.
[28] J. Kennedy and R. Eberhart, "Particle swarm optimization," in *Proceedings of ICNN'95 - International Conference on Neural Networks*, 1995, pp. 1942–1948 vol.4. doi: 10.1109/ICNN.1995.488968.

CURRICULUM VITAE

2014 – 2019	B.Sc., Computer Engineering, Erciyes University, Kayseri, TURKEY
2022 – Present	M.Sc., Electrical and Computer Engineering, Abdullah Gul University, Kayseri, TURKEY