Mustafa ETCIL

M.Sc. Thesis

AGU 2024

ENHANCING BREAST CANCER DETECTION WITH A HYBRID MACHINE LEARNING APPROACH

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
MUSTAFA ETCIL
May 2024

ENHANCING BREAST CANCER DETECTION WITH A HYBRID MACHINE LEARNING APPROACH

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
Mustafa ETCIL
May 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: Mustafa Etcil

Signature :

REGULATORY COMPLIANCE

M.Sc. thesis titled **"ENHANCING BREAST CANCER DETECTION WITH A HYBRID MACHINE LEARNING APPROACH"** has been prepared in accordance with the Thesis Writing Guidelines of the Abdullah Gül University, Graduate School of Engineering & Science.

Prepared By	Advisor	Co Advisor
Mustafa Etcil	Assoc.Dr. Burcu Bakir Güngör	Prof. Dr. V. Cagri GÜNGÖR
Signature	Signature	Signature

Head of the Electrical and Computer Engineering Graduate Program

Asst. Prof. Samet GÜLER

ACCEPTANCE AND APPROVAL

M.Sc. thesis titled **"ENHANCING BREAST CANCER DETECTION WITH A HYBRID MACHINE LEARNING APPROACH"** and prepared by Mustafa Etcil has been accepted by the jury in the Electrical and Computer Engineering Graduate Program at Abdullah Gül University, Graduate School of Engineering & Science.

30/05/2024

(Thesis Defense Exam Date)

JURY:

Advisor : Assoc. Prof. Burcu GÜNGÖR

Member : Assoc. Prof. Rıfat KURBAN

Member : Asst. Prof. Tayyip ÖZCAN

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

ENHANCING BREAST CANCER DETECTION WITH A HYBRID MACHINE LEARNING APPROACH

Mustafa ETCIL
MSc. in Electrical and Computer Engineering
Advisor: Assoc. Prof. Burcu GÜNGÖR
Co-advisor : Prof. Vehbi Çağrı GÜNGÖR
May 2024

According to the World Health Organization (WHO), breast cancer is one of the most prevalent illnesses, with 7.8 million instances recorded in the previous five years. As such, it poses a serious threat to world health. This alarming statistic underscores the urgent necessity for enhanced diagnostic methods. Against this backdrop, the current study proposes a novel diagnostic model, the CSA-PSO-LR classifier, which innovatively combines the clonal selection algorithm (CSA) with particle swarm optimization (PSO) to refine the logistic regression model training process for breast cancer detection. This research employs two extensively recognized datasets: the Wisconsin Diagnostic Breast Cancer (WDBC) and the Wisconsin Breast Cancer Database (WBCD), putting into practice a strict evaluation procedure that assesses performance using Bayesian hyperparameter optimization and 10-fold cross-validation. Furthermore, the study introduces CPU parallelization strategies to significantly curtail the model training time. Comparative analyses against machine learning algorithms, encompassing decision trees, extreme gradient boosting, k-nearest neighbors, logistic regression, random forests, and support vector machines, demonstrate the CSA-PSO-LR classifier's superior performance in detection accuracy and F1-measure. This investigation contributes a groundbreaking approach to the early detection of breast cancer, potentially facilitating more effective treatment plans and enhancing patient survival prospects.

Keywords: Clonal selection algorithm, Particle swarm optimization, Logistic regression, Bayesian optimization, Breast cancer diagnosis

ÖZET

HİBRİT MAKİNE ÖĞRENME YAKLAŞIMI İLE GÖĞÜS KANSERİ TESPİTİNİN GELİŞTİRİLMESİ

Mustafa ETCİL
Elektrik ve Bilgisayar Mühendisliği Anabilim Dalı Yüksek Lisans
Tez Danışmanı: Doç.Dr. Burcu GÜNGÖR
İkinci Tez Danışmanı: Prof.Dr. Vehbi Çağrı GÜNGÖR
Mayıs 2024

Dünya Sağlık Örgütü (WHO) tarafından belirlendiği üzere, göğüs kanseri, son beş yılda 7.8 milyon yeni vakayla en yaygın kanser türlerinden biri olarak ön plana çıkmaktadır. Bu çarpıcı istatistik, gelişmiş tanı yöntemlerine olan acil ihtiyacı vurgulamaktadır. Bu bağlamda, mevcut çalışma, göğüs kanseri tespiti için lojistik regresyon modeli eğitim sürecini iyileştirmek amacıyla klonal seçim algoritması (CSA) ile parçacık sürü optimizasyonunu (PSO) yenilikçi bir şekilde birleştiren CSA-PSO-LR sınıflandırıcısını önermektedir. Bu araştırma, geniş çapta tanınan iki veri seti olan Wisconsin Diagnostik Göğüs Kanseri (WDBC) ve Wisconsin Göğüs Kanseri Veritabanı (WBCD) kullanılarak, performans değerlendirmesi için 10 kat çapraz doğrulama ve Bayes hiperparametre optimizasyonunu içeren katı bir değerlendirme protokolü uygulamaktadır. Ayrıca, çalışma, model eğitim süresini önemli ölçüde kısaltmayı amaçlayan CPU paralelleştirme stratejilerini tanıtmaktadır. Karar ağaçları, aşırı gradyan artırma, en yakın komşular, lojistik regresyon, rastgele ormanlar ve destek vektör makineleri gibi makine öğrenimi algoritmalarına karşı yapılan karşılaştırmalı analizler, CSA-PSO-LR sınıflandırıcısının tespit doğruluğu ve F1-ölçütü açısından üstün performans sergilediğini göstermektedir. Bu araştırma, göğüs kanserinin erken tespitine yönelik yenilikçi bir yaklaşım sunarak, daha etkili tedavi planlarının kolaylaştırılmasına ve hastaların hayatta kalma beklentilerinin artırılmasına katkıda bulunmaktadır.

Anahtar kelimeler: Klonal seçim algoritması, Parçacık sürü optimizasyonu, Lojistik regresyon, Bayes optimizasyonu, Göğüs kanseri teşhisi

Acknowledgements

i. I am incredibly grateful to my supervisor, Assoc. Dr. Burcu GÜNGÖR, for their constant guidance and support during my research. Their expertise and encouragement were instrumental in helping me complete this thesis. I would also like to express my deepest thanks to my co-advisor, Prof. Dr. V. Cagri GÜNGÖR, for sharing their invaluable knowledge and for their patient and unwavering support. Their insightful feedback on my thesis significantly improved my approach and strengthened my arguments.

ii. The support, insights, and encouragement provided by Dr. Bilge Kağan DEDETÜRK and Dr. Burak KOLUKISA were instrumental in the development of this thesis. I am sincerely grateful for their contributions.

iii. I also want to express my sincere gratitude to Asst. Prof. Tayyip ÖZCAN and Assoc. Prof. Rıfat KURBAN for serving on the thesis defense committee.

iv. I wish to acknowledge the unwavering support, sacrifices, and enduring belief in my abilities bestowed upon me by my family.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

LIST OF ABBREVIATIONS

ACC	Accuracy
CSA	Clonal Selection Algorithm
CV	Cross Validation
DT	Decision Tree
F1	F1-Score
HT	Hyperparameter Tuning
KF	K-Fold
KNN	k-Nearest Neighbors
LR	Logistic Regression
ML	Machine Learning
NA	Not available
PSO	Particle Swarm Optimization
Ref	Reference
RF	Random Forest
SS	Sampling Strategy
SVM	Support Vector Machine
TT	Training Time
WBCD	Wisconsin Breast Cancer Database
WDBC	Wisconsin Diagnostic Breast Cancer
WHO	World Health Organization
XGB	XGBoosting

To my family

Chapter 1

Introduction

According to the World Health Organization (WHO), there has been a concerning rise in the number of new cases of breast cancer reported over the past five years, with an estimated 7.8 million cases worldwide. Notably prevalent among women, the year 2020 saw breast cancer cases ascend to 2.3 million globally, accompanied by a tragic toll of 685,000 fatalities attributed to this disease [1]. Projections suggest a concerning trend with an estimated 2.5 million women expected to be diagnosed by 2025 [2]. Despite extensive research, there is currently no conclusive evidence to suggest a direct link between breast cancer and specific viruses or bacterial infections [3]. Interestingly, a significant portion of breast cancer diagnoses—approximately half—cannot be attributed to known risk factors, with age and gender being the primary exceptions. This statistic highlights the urgent requirement for early detection methods to effectively fight the disease [4].

Detection of breast cancer has advanced greatly, utilizing various methods beyond machine learning. Mammography, a specialized form of X-ray imaging for breast tissue, is a widely recognized and effective detection method. By detecting cancers in their early stages, it is regarded as the gold-standard method for early identification and improves the prognosis of breast cancer [5].

Ultrasound plays a critical role in assessing breast masses and distinguishing between solid tumors and fluid-filled cysts, often used alongside mammography [6]. Magnetic resonance imaging (MRI) of the breast is also beneficial, especially for high-risk individuals, providing detailed images of the breast tissue without radiation exposure [7].

The rise of computer-assisted technologies, specifically machine learning (ML), has transformed the way breast cancer detection is approached, providing significant improvements compared to traditional diagnostic methods. ML models have played a

crucial role in identifying intricate patterns in breast imaging that may be difficult for humans to detect, thanks to their capability to learn from and predict data. Utilizing various machine learning (ML) algorithms, such as convolutional neural networks (CNNs) and support vector machines (SVM), to increase the accuracy of breast cancer detection has been the focus of recent research [8]. Incorporating machine learning (ML) approaches in breast cancer testing enhances the accuracy of detection and lowers the number of incorrect results. This leads to better patient outcomes and more effective treatments. The growing application of machine learning (ML) in medical diagnostics shows how artificial intelligence can revolutionize healthcare, particularly in the accurate and early detection of breast cancer.

The advancement of metaheuristic optimization algorithms has played a crucial role in the progression of computer techniques for addressing challenging problems, such as breast cancer diagnosis. Algorithms inspired by natural processes, such as artificial immune systems and swarm intelligence, have shown promising results. One algorithm in particular, the Clonal Selection Algorithm (CSA) [9], imitates the immune system's response to pathogens by improving the selection process to reduce the affinity between antibodies and antigens, mimicking the body's affinity maturation. This method is inspired by artificial immune system concepts. Particle Swarm Optimization (PSO) [10] has become a prominent swarm intelligence technique at the same time, influenced by the group behavior of decentralized, self-organizing systems like fish schools or bird flocks. PSO efficiently converges on optimal solutions through collaborative agent movement in the search space, with strong local search abilities [11]. By offering adaptable, trustworthy solutions for complex optimization problems, these metaheuristic techniques enhance not only breast cancer detection accuracy but also significantly impact computational biology.

In this thesis, a new classifier called CSA-PSO-LR is presented. It combines the CSA and PSO algorithms to improve the Logistic Regression (LR) framework. While the CSA is great for optimization, it lacks in exploration [12]. The CSA-PSO-LR model aims to enhance the weight optimization process of logistic regression by efficiently minimizing the cost function and leveraging the benefits of both CSA and PSO. By using the PSO technique in place of CSA's hypermutation phase, the model enhances CSA's local search capabilities.

During this thesis' experimentation phase, we used a number of strategies to raise the CSA-PSO-LR model's efficacy and efficiency. To tackle the issue of long computational times associated with the model, we utilized CPU parallelization to significantly reduce the training time. To ensure optimal performance, we also automated hyperparameter fine-tuning for our CSA-PSO-LR model and other classifiers using Bayesian hyperparameter optimization. We evaluated the model's performance using the widely-known WDBC and WBCD datasets, using a 10-fold cross-validation strategy to guarantee the accuracy and validity of the findings.

Chapter 2

Literature Review

Numerous studies have examined the use of different kinds of machine learning classifiers, either independently or in conjunction, to reliably identify breast cancer tumors as benign or malignant.

In a recent study [13], researchers examined how different algorithms, including C4.5, SVM (Support Vector Machine), NB (Naive Bayes), and KNN (k-Nearest Neighbor), can be employed to diagnose breast cancer. They found that the SVM algorithm was particularly effective, with an accuracy of 97.13% on the WBCD dataset. Another study [14] used the Recursive Feature Elimination (RFE) method to reduce the number of features in the WDBC dataset to 15, yielding a remarkable 98.06% accuracy rate with the LR classifier.

Many studies have focused on classifying breast cancer tumors using artificial immune-based optimization methods [15,16].

Instead of the common backpropagation method, researchers [17] used CSA to determine the best weights and biases for the Multi-layer Perceptron (MLP) algorithm in a study. The proposed CSA-MLP model achieved a detection accuracy of 98.24% on the WDBC dataset with a 5-fold cross-validation technique.

The researchers in study [18] utilized the Artificial Immune System for Associative Classification (AISAC) algorithm, an approach based on artificial immune system principles, to tackle challenges with incomplete and diverse data. The enhanced AISAC version, known as Associative Classification of Mixed and Missing Data (AISAC-MMD), showed impressive accuracy rates of 96.50% on the WDBC dataset and 96.90% on the WBCD dataset.

Researchers introduced CLONALG-LFS, a novel method for choosing local features based on the artificial immune system, in a paper [19]. This method uses the clonal selection algorithm to identify the best feature subsets, utilizing local clustering for evaluation. Consequently, the WBCD dataset has been reduced down to just two features, achieving a remarkable 95.48% performance rate.

Recently, many scientists have been working on improving the accuracy of classification by combining swarm intelligence-based optimization techniques with different classifiers [20,21,22].

A previous study [23] utilized Improved Particle Swarm Optimization (IPSO) and Improved Firefly Algorithm (IFA) to identify the best feature subset from the WDBC dataset, choosing 10 features in the end. For classification, the dataset was split into training and test sets at an 80-20% ratio. Different algorithms such as DT, KNN, SVM, and RF were used, with RF showing the best results in terms of performance.

A study conducted by researchers [24] used the PSO algorithm for selecting features in the WBCD dataset and optimizing the C4.5 algorithm. The method's effectiveness was evaluated through 10-fold cross-validation, showing that PSO improved C4.5's performance by 0.88%.

Another study [25] introduced a hybrid model that combined the non-parametric Kernel Density Estimation (KDE) algorithm with PSO. PSO was used to adjust the kernel bandwidth parameters for KDE and identify the best subsets. With accuracy rates of 98.45% and 98.53% on the WDBC and WBCD datasets, they employed a 10-fold cross-validation sampling approach.

Researchers in [26] conducted a study to investigate alternative optimization methods for detecting breast cancer, moving beyond PSO to explore swarm intelligence. They introduced an enhanced Gray Wolf algorithm, which was used to identify the most effective features from the WDBC dataset. After selecting six attributes, these features were entered into an SVM model. By assessing the effectiveness of their proposed EGWO-SVM approach through 10-fold cross-validation, outstanding 98.24% accuracy rates and 97.40% F1-measure rates were attained by the researchers.

The study [27] utilized the Dragonfly and Whale optimization algorithms to enhance the SVM classifier's performance. Two distinct models, DA-SVM and WOA-SVM, were introduced and evaluated using train-test splitting and 10-fold cross-validation methods. Significant improvements were observed on both the WDBC and WBCD datasets.

Study [28] presented a novel feature selection approach called BOAALO, which combines the Butterfly Optimization Algorithm (BOA) and Ant Lion Optimizer (ALO) algorithms. This method was used to choose the most effective features from the WDBC dataset, which were then fed into an Artificial Neural Network (ANN). The BOAALO-

ANN model gained an outstanding accuracy rate of 98.16% through 10-fold cross-validation.

In a recent study, experts explored a new approach called the weighted KNN (wKNN) method, which differs from the typical majority voting technique used in regular KNN algorithms. Additionally, they integrated the Crowsearch Optimization Algorithm (CSOA) to find the best parameter values for wKNN, such as the number of neighbors and distance measurement. Their innovative CSOA-wKNN model successfully detected abnormalities with a high accuracy rate of 97.36%.

In recent studies, researchers have demonstrated the promise of using metaheuristic optimization algorithms inspired by swarm intelligence or artificial immune systems to aid in diagnosing breast cancer. Despite their potential, these algorithms have been found to have drawbacks such as low accuracy rates and long training times. Additionally, the importance of hyperparameter tuning in improving classification accuracy and avoiding overfitting has not been fully recognized in these investigations.

In this thesis, a new approach is suggested for enhancing the weights of the logistic regression classifier. This is done by merging CSA and PSO algorithms to overcome limitations. The hyperparameters of this new model, CSA-PSO-LR, are adjusted using Bayesian optimization [30], and its effectiveness is tested using 10-fold cross-validation. Compared to earlier research, this method introduces a robust model that can perform both local and global searches by utilizing the capabilities of two different optimization algorithms in logistic regression training. Furthermore, using a CPU parallelization strategy can greatly decrease the time required for training. The method suggested has proven to provide reliable and satisfactory results for both datasets in terms of accuracy in detection, F1-measure, and training duration, as shown in Table 2.1.

Table 2.1 A summary of research in existing literature regarding the identification of breast cancer.

Ref.	Year	Data Set	Method	KF	HT	ACC (%)	F1(%)	TT (sec.)
[15]	2013	WDBC	I-CLONALG	-	-	98.58	-	-
[20]	2016	WDBC	PSO-SVM	10	-	93.53	-	42.3
[25]	2016	WDBC	PSO-KDE	10	-	98.45	-	-
[55]	2018	WDBC	GRU-SVM	-	-	93.75	-	-
[16]	2020	WDBC	AISAC	5	-	93.67	-	35
[21]	2020	WDBC	PSO-NDS	-	✓	98.6	-	-
[23]	2021	WDBC	IPSO-IFA-RF	-	-	97	-	1.12
[26]	2021	WDBC	EGWO-SVM	10	-	98.24	97.40	-
[28]	2021	WDBC	BOAALO-ANN	10	-	98.16	-	-
[27]	2022	WDBC	WOA-SVM	10	-	97.54	-	-
[14]	2022	WDBC	LR-RFE	-	✓	98.06	97.36	0.42
[17]	2022	WDBC	CSA-MLP	5	-	98.24	-	-
[18]	2023	WDBC	AISAC-MMD	-	-	96.50	-	-
[29]	2023	WDBC	CSOA-wKNN	5	✓	97.36	97.77	-
PM	**2024**	**WDBC**	**CSA-PSO-LR**	**10**	✓	**98.75**	**98.27**	**2.46**
[22]	2013	WBCD	PSO-ANN	-	-	97.36	-	-
[13]	2016	WBCD	SVM	10	-	97.13	96	-
[24]	2018	WBCD	PSO-C4.5	10	-	96.49	-	-
[16]	2020	WBCD	AISAC	5	-	97.42	-	4.7
[19]	2020	WBCD	CLONALG-LFS	10	-	95.48	-	-
[18]	2023	WBCD	AISAC-MMD	-	-	96.90	-	-
PM	**2024**	**WBCD**	**CSA-PSO-LR**	**10**	✓	**97.94**	**97.53**	**3.03**

Chapter 3

Material and Methods

3.1 Datasets

In this thesis, we employ two prominent and accessible datasets from the UCI Machine Learning Repository: the Wisconsin Diagnostic Breast Cancer (WDBC) [31] and the Wisconsin Breast Cancer Database (WBCD) [32]. These datasets are derived from Fine Needle Aspirate (FNA) images of breast masses, a minimally invasive technique for collecting breast tissue samples for examination. The WDBC dataset is detailed, containing 569 instances with 30 features for each instance. These features include the mean, standard error, and the "worst" or largest value (calculated as the mean of the three largest values) for 10 primary measurements, totaling 30 features. Out of the total instances, 212 (37%) are malignant, and 357 (63%) are benign. On the other hand, the WBCD dataset comprises 699 instances with 11 attributes. This includes 241 cases (34%) identified as malignant and 458 cases (66%) identified as benign. Table 3.1 and Table 3.2 give the feature definitions of the WDBC and WBCD datasets respectively.

Previous research [33,34] emphasizes the importance of evaluating databases on the grounds of their complexity, linearity, and the balance between different classifications. In line with this approach, our analysis, depicted in Figure 3.1, undertakes a side-by-side evaluation of the WDBC and WBCD databases, focusing on these very aspects.

To facilitate this comparative analysis, the study leverages Fisher's discriminant ratio (F1) [35] and the N1 metrics, innovations introduced in a preceding study, to gauge the linearity and complexity of these databases. The F1 metric is particularly insightful, measuring the extent of feature overlap across different categories. A low F1 score suggests a more complex issue where no single characteristic is able to clearly distinguish the groups. Conversely, an elevated F1 score suggests at least one attribute significantly overlaps between categories, simplifying the classification task. The N1 metric assesses the separability of class distributions, with higher values signaling a reduced distinction

between classes and, by extension, a more arduous classification challenge. Our analysis, as illustrated in Figure 3.1, reveals that the WDBC database embodies a more complex challenge compared to the WBCD database, when assessed through the lens of F1 and N1 scores.

Moreover, the concept of an imbalance ratio, which denotes the proportion of malignant to benign tumors within a dataset, serves as another critical evaluation criterion. The WDBC database exhibits an imbalance ratio of 0.59, whereas the WBCD database presents a slightly lower ratio of 0.53. This indicates that the WDBC database is relatively more imbalanced compared to the WBCD database, adding an additional layer of complexity to the classification tasks associated with the WDBC dataset.

Table 3.1 The WDBC dataset's feature explanation

Feature name	**Feature type**	**Description**
Diagnosis	Categorical	Classifies the tumor as either Malignant (dangerous) or Benign (harmless).
Radius	Numerical	Average distance from the tumor's center to its outer boundary.
Texture	Numerical	Variation in the grayscale intensity across the tumor, showing texture irregularity.
Perimeter	Numerical	The length of the tumor's outer edge.
Area	Numerical	The overall size measured within the tumor's boundaries.
Smoothness	Numerical	A quantification of how even or uneven the surface variations are on the tumor's boundary.
Compactness	Numerical	Reflects the density or tightness of the tumor, based on area and perimeter measurements.
Concavity	Numerical	The extent of indentations or hollow areas on the tumor's surface.
Concave points	Numerical	Counts the number of indentations or hollow points on the tumor's surface.
Symmetry	Numerical	The balance in shape and structure of the tumor when divided into halves.
Fractal dimension	Numerical	A measure of the complexity of the tumor's boundary, indicating how much the shape deviates from smooth.

Table 3.2 The WBCD dataset's feature explanation

Feature name	Feature type	Description
Clump Thickness	Numerical	Thickness of the cell clump.
Uniformity of Cell Size	Numerical	Uniformity in cancer cell sizes.
Uniformity of Cell Shape	Numerical	Similarity in cancer cell shapes.
Marginal Adhesion	Numerical	Cell adhesion strength to neighbors.
Single Epithelial Cell Size	Numerical	Size of epithelial cells.
Bare Nuclei	Numerical	Presence of bare nuclei.
Bland Chromatin	Numerical	'Blandness' of chromatin.
Normal Nucleoli	Numerical	Normalcy of nucleoli.
Mitoses	Numerical	Number of cell divisions.
Class	Categorical	Diagnosis (2: benign, 4: malignant).

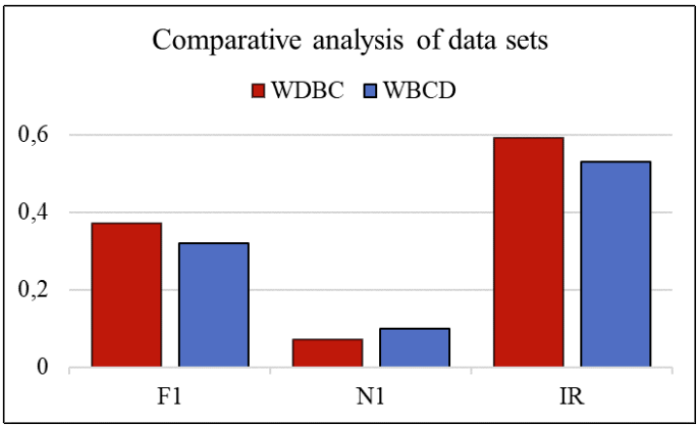

Figure 3.1 Evaluating the WDBC and WBCD datasets in terms of their complexity, linearity, and the balance of classes

3.2 Data Preprocessing

Data preprocessing is a crucial initial stage in the workflow for machine learning and data analysis, in order to prepare raw data for subsequent processing and analysis. Its

primary objective is to raise the caliber of data and prepare it for the development of precise and effective models. In the context of this thesis, the data preprocessing phase is meticulously designed to encompass merely two critical steps.

3.2.1 Handling missing values

The initial step in the preprocessing sequence involves a thorough examination of the WBCD to identify and eliminate samples that are incomplete, particularly those lacking data in the "Bare Nuclei" attribute. An audit of the dataset revealed that 16 samples were afflicted by this issue of missing values. The exclusion of these records is a necessary measure to maintain the dataset's consistency and reliability. Following this purging process, the dataset's composition was reevaluated, revealing a revised distribution wherein 239 samples, equivalent to 35% of the dataset, were classified as malignant tumors, while the remaining 444 samples, constituting 65% of the dataset, were categorized as benign tumors.

3.2.2 Scaling

The second step in the preprocessing involves the application of a normalization procedure to both the WBCD and WDBC datasets. This is accomplished utilizing the Standard-Scaler technique, a method facilitated by the Scikit-Learn [36] library, renowned for its efficacy in data preprocessing for machine learning tasks. By giving each feature a zero mean and a one standard deviation, the Standard-Scaler algorithm standardizes the feature collection.

3.3 Machine Learning

Machine learning (ML) is a subfield of artificial intelligence that is defined by statistical models and algorithms that allow computers to carry out tasks by deriving conclusions and patterns from data instead than directly following instructions that have been programmed. ML models become adept at tasks like predicting future outcomes, classifying data points, and identifying anomalies in unseen data.

Supervised and unsupervised learning are the two primary subcategories of machine learning. In the domain of supervised learning, models are trained on a dataset containing both inputs and their corresponding correct outputs, with the aim of learning

a mapping function with the ability to forecast results for novel, unforeseen inputs. This methodology encompasses techniques like classification, where the goal is to assign discrete labels to data points, and regression, where the prediction of continuous quantities is required [37]. Conversely, unsupervised learning involves algorithms that are used to analyze and cluster input data without any labeled responses, thereby identifying patterns or structures within the data autonomously. Common techniques in unsupervised learning include cluster analysis, which groups data points based on similarity, and principal component analysis, which simplifies the complexity of high-dimensional data.

3.4 Optimization Algorithms

3.4.1 Optimization Definition and Purpose

The process of optimizing involves determining which solution is best for a particular problem, which typically involves reducing expenses, augmenting efficiency, enhancing profits, or achieving various other objectives. Central to this endeavor are two critical elements: the objective function and the decision variables. The objective function—sometimes termed as the cost or loss function—is the quantitative expression of the aspect one seeks to minimize or maximize, such as operational costs, the error rates of a predictive model, or other measurable criteria. Decision variables, in contrast, are the modifiable parameters that are adjusted to achieve the optimal value of the objective function. The ultimate aim of optimization is to find the values of these decision variables that result in the most advantageous outcome for the objective function.

3.4.2 Concepts of Local and Global Minima

In the field of optimization, the concepts of global and local minima are pivotal. A global minimum at a point x^* for a function f is defined when $f(x^*)$ is less than or equal to $f(x)$ for all x within the defined domain; meaning that at x^*, f attains the lowest possible value across the entire domain. Conversely, a local minimum at x^* is characterized by $f(x^*)$ being less than or equal to $f(x)$ for all x in the vicinity of x^*, implying that $f(x^*)$ is the lowest value in that local neighborhood, though not necessarily the lowest in the entire domain. Figure 3.2 shows the local and global minimum concepts.

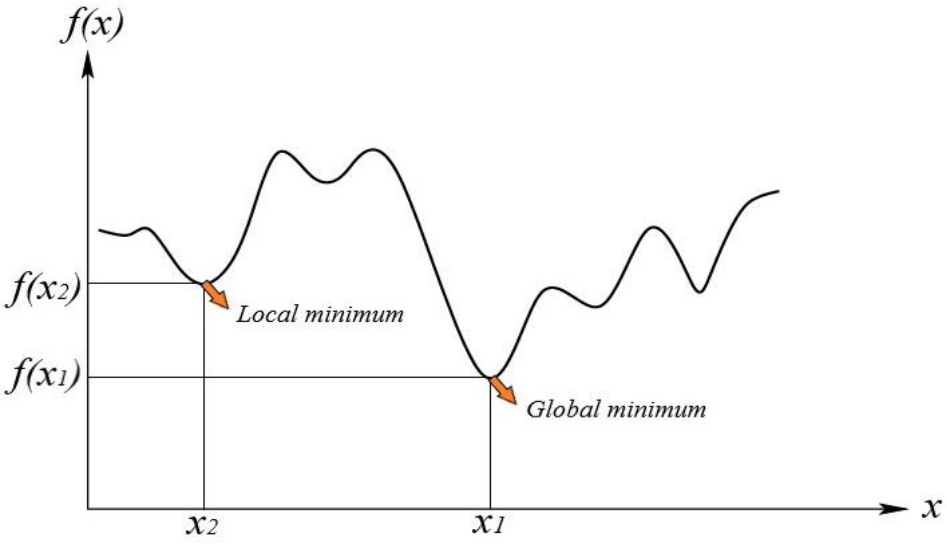

Figure 3.2 A function graph displaying the local minimum and global minimum concepts

3.4.3 Heuristics and Metaheuristics

Optimization techniques are broadly categorized into heuristics and metaheuristics, each suited to different types of problems and solution strategies. Heuristics are pragmatic, problem-specific approaches that forgo complex mathematical procedures, providing expeditious yet potentially suboptimal solutions, exemplified by the Greedy Algorithm, which pursues stepwise local optima in the hopes of achieving a globally optimal solution [38].

Metaheuristics, on the other hand, operate at a higher abstraction level, applying more universally to a range of problems [39]. Within metaheuristics, there are two subcategories: Single Solution Metaheuristics and Population-based Metaheuristics. The former focuses on iteratively improving a standalone solution, as seen in Simulated Annealing [40] and Tabu Search [41], while the latter encompasses evolutionary or biologically inspired algorithms that concurrently refine a group of solutions, such as Genetic Algorithms [42], Ant Colony Optimization [43], the Artificial Bee Colony Algorithm [44], and the Differential Evolution Algorithm [45].

3.4.4 Gradient Descent Algorithm

The Gradient Descent algorithm is a method for minimizing an objective function with the goal of identifying the combination of parameters that results in the lowest value.

It utilizes the direction of the steepest descent, which is opposite to the gradient of the function, to progress towards the global minimum [46]. For a given objective function $J(\theta)$, and the parameters of the model, θ, the algorithm iteratively adjusts θ, following a specific rule, to reduce $J(\theta)$ effectively. The update rule as follows:

$$\theta := \theta - \mu . \nabla_{\theta} J(\theta) \qquad 3.4.4.1$$

In this expression, the term μ represents the learning rate, which dictates the magnitude of the steps taken towards the minimum, while $\nabla_{\theta} J(\theta)$ symbolizes the gradient of the cost function $J(\theta)$ with regard to the parameters θ. Figure 3.3 shows an illustration of the gradient descent algorithm.

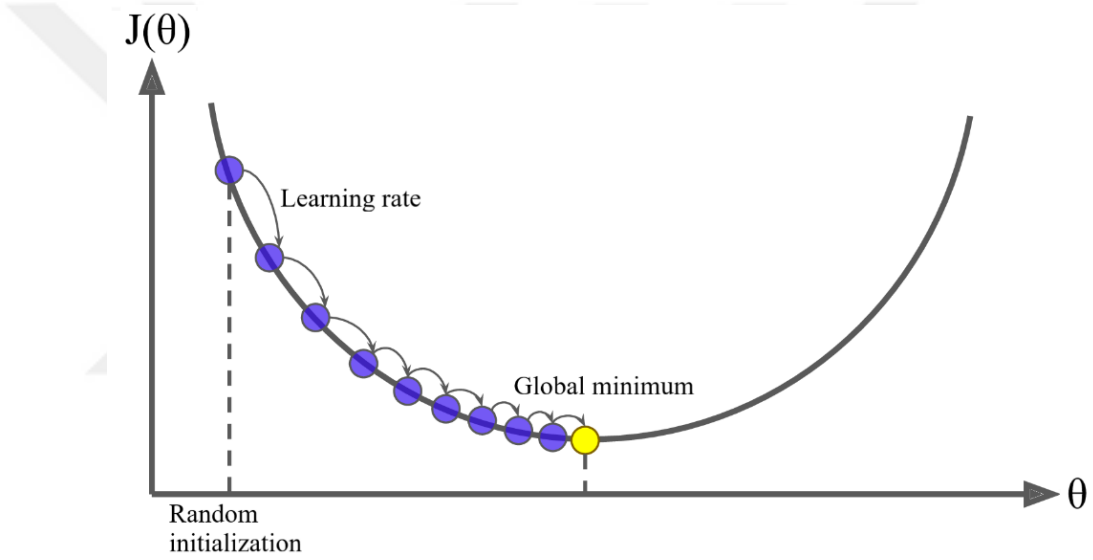

Figure 3.3 Visualization of the gradient descent optimizer

The optimal selection of the learning rate, denoted as μ, is critical for the efficacy of the Gradient Descent algorithm. An excessively large μ may result in the algorithm overshooting the minimum, leading to divergence, whereas an excessively small μ may lead to protracted convergence times or entrapment in local optima or saddle points. The phenomenon of entrapment in local optima or saddle points, as delineated in Figure 3.4, represents a significant impediment to the algorithm, occurring when it converges to a point that is not the global minimum. At such points, due to the gradient being null, the update formula maintains the parameters θ static, thereby prematurely halting the algorithm's progress. This highlights the imperative of judiciously setting the learning

rate and parameter initialization to navigate the challenges inherent in Gradient Descent optimization effectively.

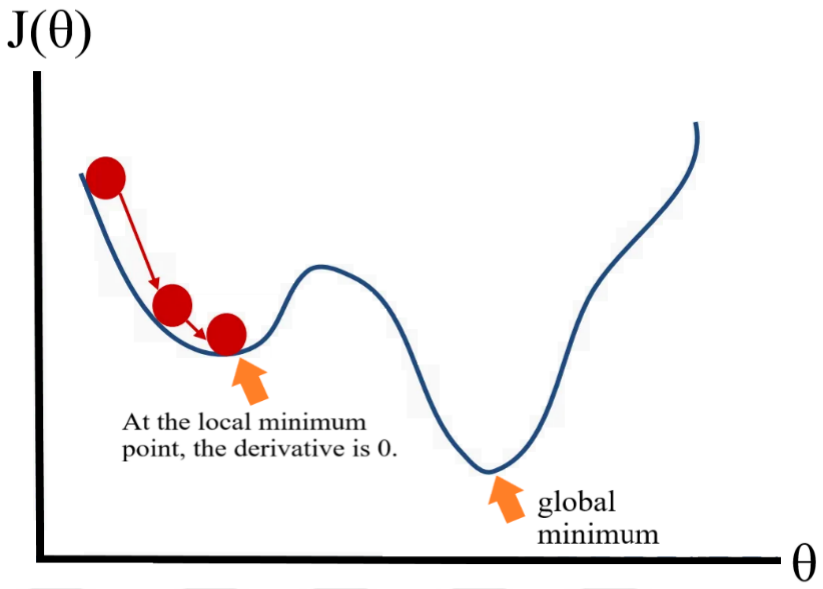

Figure 3.4 Local minimum trapping problem in gradient descent algorithm

3.5 Machine Learning Classifiers

Classification constitutes a fundamental component in machine learning, targeting the assignment of novel instances into predefined classes. Algorithms are thus engineered to categorize emails as spam or legitimate, or to infer medical diagnoses from patient records. This procedure predominantly leverages supervised learning, wherein a model, trained on a dataset comprising known outcomes, is tasked with the precise classification of novel, unobserved data.

In this thesis, a detailed comparison of several classifiers, including logistic regression (LR), support vector machines (SVM), k-nearest neighbors (KNN), decision trees (DT), random forest (RF), and extreme gradient boosting (XGB), is applied to assess the efficacy of the method being proposed.

3.5.1 K- nearest neighbors

The K-Nearest Neighbors algorithm is a non-parametric method for regression and classification, based on the idea that comparable occurrences produce similar results. Using a predetermined distance metric (such as Manhattan or Euclidean), it classifies a sample according to the prevalent class among its k-nearest neighbors. The distance metric and k selection have a significant impact on how effective the algorithm is.

3.5.2 Support vector machine

A popular supervised learning technique for classification and regression applications is the support vector machine. In order to maximize the margin between the nearest points in each class—known as support vectors—it works on the basis of identifying the hyperplane in the feature space that best divides the various classes. SVMs are especially well-known for working well in high-dimensional environments and situations where there are more dimensions than samples [47].

3.5.3 Decision tree

In order to simulate decision paths based on input features, the Decision Tree method builds a tree structure. The tree's internal nodes stand in for attribute testing, branches for the test results, and leaf nodes for the conclusion or forecast. The objective of this technique is to create homogenous subsets about the target variable in classification tasks by recursively partitioning the data space into subsets based on feature values.

3.5.4 Random forest

By constructing several decision trees throughout training, the Random Forest algorithm is an ensemble learning method that determines the mode of the classes of the individual trees. Random forests aggregate the output of numerous trees built from different data subsets and use a random feature selection at each split in the tree-building process to lessen the variation of individual decision trees. This method reduces overfitting and increases tree variability. Every tree in the forest is trained using a random sample extracted with replacement from the training set by a procedure called bootstrap aggregating, also referred to as bagging.

3.5.5 XGBoost

A complex gradient boosting technique that puts speed and efficiency first is called Extreme Gradient Boosting, or XGBoost. It functions by adding predictors to an ensemble one after another, each of which corrects the one before it, progressively raising the model's accuracy. XGBoost specifically enhances the gradient boosting architecture and lessens the likelihood of overfitting by incorporating regularization terms into its objective function.

3.6 Cpu Parallelization

In this thesis, the concept of CPU parallelization is explored as a means to enhance the utilization of existing hardware capabilities through the concurrent execution of computational processes across several CPU cores. For the purposes of implementing a CPU-parallelized variant of CSA-PSO, this research utilizes the NumPy library [48]. NumPy is instrumental in facilitating efficient operations on arrays and executing complex mathematical computations, which are foundational to the methodologies employed in machine learning. This is accomplished by taking advantage of the CPUs' built-in parallel processing capabilities, which speeds up operations like sorting arrays, manipulating matrices, and doing statistical analysis [49]. By dividing up work among several cores in a single CPU, this method of parallelization aims to maximize computational efficiency. This strategy seeks to enhance performance without the need for execution on multiple machines or in clustered computing environments. Detailed documentation of the CPU version's implementation is made available through the cited references [50,51].

3.7 Model Evaluation and Validation

3.7.1 Cross validation

In the context of this thesis, evaluation is conducted using a 10-fold cross-validation technique on the presented model as well as several additional categorization frameworks during the classification process. A statistical model validation method called 10-fold cross-validation is used to assess how well machine learning models predict the future. The initial sample is divided into ten equal parts, or folds, in this process. In this procedure, the model is trained using nine folds, and tested using the one fold that remains. Ten iterations of this process are carried out, using one testing set for every ten fold. To provide a thorough assessment of the model's predicted accuracy, the final model performance is then averaged over the course of 10 trials.

3.7.2 Evaluation metrics

A table that describes how well a categorization algorithm performs is called a confusion matrix. It consists of four elements: True Positive (TP), which is the count of cases correctly identified as positive; False Positive (FP), which represents the cases incorrectly labeled as positive; True Negative (TN), indicating the cases correctly identified as negative; and False Negative (FN), representing the cases incorrectly labeled as negative. The confusion matrix adopted in this thesis is given in Figure 3.5.

		Predicted class	
		Negative	Positive
Actual class	Negative	TN	FP
	Positive	FN	TP

Figure 3.5 Confusion matrix

The confusion matrix is used to determine accuracy, which is then computed as the sum of true positives and true negatives over all cases, as stated in Eq. 3.7.2.1. This metric gives a clear assessment of the model's overall performance by calculating the percentage of all predictions that the model made correctly.

$$Accuracy = \frac{TP + TN}{TP + FP + TN + FN} \qquad 3.7.2.1$$

On the other hand, The F1-measure as given in Eq. 3.7.2.2 is another important metric calculated from the confusion matrix, designed to assess the model's performance especially on unbalanced datasets.

$$F1 - measure = \frac{2TP}{2TP + FP + FN} \qquad 3.7.2.2$$

3.8 Hyperparamater Optimization

In this thesis, the proposed CSA-PSO-LR and other classifiers utilize Bayesian hyperparameter optimization to automatically adjust their hyperparameters. Bayesian hyperparameter optimization [52] is a sophisticated methodology in machine learning that optimizes model hyperparameters through probabilistic modeling of the objective function. This technique employs Bayesian inference to forecast model performance on

unseen datasets, guiding the selection of hyperparameters towards those with the highest expected efficacy. Central to Bayesian optimization is the use of a surrogate probabilistic model, typically a Gaussian Process, to approximate the objective function based on prior observations of hyperparameter configurations and their resultant performance metrics. Iterative updates to this surrogate model incorporate new data, refining the prediction of the objective function's landscape.

Chapter 4

Proposed Algorithm

4.1 Clonal Selection Algorithm

The clonal selection algorithm is inspired by how the immune system responds to antigens. Essentially, when the immune system encounters an antigen, it activates the antibodies most effective against it [9]. These selected antibodies are then replicated through cloning. During cloning, they undergo a process known as somatic hypermutation, improving their ability to target the antigen. Over time, less effective antibodies are phased out and replaced by more adept ones. This natural selection process leads to the production of highly efficient antibodies tailored to combat specific antigens [53]. This mechanism serves as the foundation for the clonal selection algorithm, encapsulating its core principle. The steps of the algorithm are outlined as follows:

1. Set up the values of the initial parameters of the algorithm: the size of the population (P), the number of clones that will be produced for each antibody (α), and the percentage of antibodies to be replaced (B).
2. Randomly generate the first P antibody population $Ab = \{Ab_1, Ab_2, \dots, Ab_P\}$.
3. Determine the affinity score for each Ab_i antibody.
4. To create a group of clones, make sure each antibody generates an identical number of clones. This approach is particularly useful when the goal is to identify multiple optimal solutions. By incorporating every antibody in the cloning phase, and ensuring each one yields the same quantity of clones, a diverse set of solutions can be explored . The total count of clones generated is determined by a specific formula.

$$C = \alpha . P \qquad 4.1.1$$

where the total number of clones produced is denoted by C.

5. In the cloning group, every antibody C_i goes through a hypermutation phase to improve its affinity. This hypermutation includes two stages: reverse mutation and pairwise mutation. With reverse mutation, if the distance between two random points k and l on the antibody exceeds two, the segment between them is flipped. This step is deemed successful if the newly mutated clone C_i^* displays a higher affinity than the original C_i. If not, the clone undergoes a pairwise mutation where the positions $C_{i,k}$ and $C_{i,l}$ are exchanged. The process stops if the mutated clone C_i^* proves to be a more effective solution than the original C_i.
6. To keep the antibody count constant in the population, a selection process is carried out where the clone with the best affinity from each antibody Ab_i 's clones is picked and retained as Ab_i.
7. Next, a certain percentage of antibodies *B*% with the weakest affinity in the population are replaced by freshly created antibodies, a step known as "receptor editing."
8. The cycle of steps from 4 to 7 is repeated until the predetermined stopping condition is achieved.

4.2 Particle Swarm Optimization

The Particle Swarm Optimization (PSO) technique, inspired by birds searching for food, operates with a group of particles where each one represents a potential answer within the search space. These particles have two main attributes: velocity and position, both of which adjust over time to converge on the most effective solution. The steps of PSO are:

1. Initialize the algorithm with certain parameters such as c_1, c_2, w and the number of particles.
2. Randomly set the starting position and velocity for each particle.
3. Evaluate the performance of each particle.
4. For every particle, update its personal-best position $p_i^{(t)}$, which shows the best position that particle has reached.
5. Identify the leading particle in the swarm and update the group's overall best-known position $g^{(t)}$.

6. Adjust the particles' velocity based on their prior velocity, personal best position, and the swarm's best position, following a specific formula as in Eq. 4.2.1.

$$V_i^{(t+1)} = w.V_i^{(t)} + c_1.r_1.\left(p_i^{(t)} - X_i^{(t)}\right) + c_2.r_2.\left(g^{(t)} - X_i^{(t)}\right) \quad 4.2.1$$

Here, $X_i^{(t)}$ represents the current location of a particle, r_1 and r_2 are random values between 0 and 1, c_1 and c_2 are learning rates associated with the particle's personal best and the global best positions, respectively, and w is the inertia that influences the update in velocity.

7. Update the position of each particle by Eq. 4.2.2.:

$$X_i^{(t+1)} = X_i^{(t)} + V_i^{(t+1)} \quad 4.2.2$$

8. Repeat the process from step 2 to step 7 until a stopping condition is met.

4.3 Logistic Regression

In the field of machine learning, logistic regression is a prominently used approach, especially for solving binary classification issues [54]. This algorithm's basic idea is to calculate the likelihood that a certain data point belongs to one of two potential groups.

For a collection of data points from the training dataset, which can be denoted as $\{(\overrightarrow{x_1}, y_1), (\overrightarrow{x_2}, y_2), \ldots, (\overrightarrow{x_n}, y_n)$ where each x_i is a feature vector that exists within the real number space, and each corresponding y_i is a binary label that can either be 0 or 1 for $1 \leq i \leq n$, LR applies a computation to gauge the likelihood of a data point belonging to a certain class. This is accomplished by employing the logit function, also known as the sigmoid function, as given in Eq.4.3.1., which serves as the activation function.

$$\hat{y}_i = \begin{cases} 1, & p_i \geq 0.5 \\ 0, & p_i < 0.5 \end{cases} \quad 4.3.1$$

where p_i denotes the probability value calculated using the sigmoid activation function given in Eq.4.3.2 for a training set sample. On the other hand, σ is the sigmoid activation function as in Eq.4.3.3.

$$p_i = \sigma(\vec{w} \cdot \overrightarrow{x_1}) \quad 4.3.2$$

$$\sigma(z) = \frac{1}{1 + e^{-z}} \quad 4.3.3$$

In the training process of logistic regression, a cost metric, specifically the mean absolute error as delineated in Eq 4.3.4, quantifies the disparity between the predicted class outputs and the actual class labels across individual instances. The objective of the training algorithm is to determine the optimum set of parameters (weights) that result in the minimization of this cost metric.

$$J(\vec{w}) = \frac{1}{n} \sum_{i=1}^{n} |y_i - p_i| \quad 4.3.4$$

Usually, the loss function is minimized by iteratively adjusting the weights and bias in the direction that minimizes the loss, such as with gradient descent.

4.4 Proposed Hybrid Algorithm

Logistic regression is typically trained using gradient descent optimization, which determines the best weights through a derivative-based method. This process, though, can be resource-heavy and sometimes gets stuck in suboptimal solutions, hindering its effectiveness.

This thesis presents an improved logistic regression training technique named CSA-PSO-LR, which combines CSA with PSO. This combined strategy takes advantage of both algorithms' strengths to more effectively explore the solution space.

In this enhanced method, logistic regression weights are treated as the CSA's antibodies, and the cost function is analogous to the antigens. Finding the optimal antibody, or the weight set with the highest affinity or lowest cost function value, is the aim. This is done through a dynamic process that involves cloning the antibodies, inducing high rates of mutation, selecting the most effective variants, and fine-tuning them, as depicted in Figure 4.1 which illustrates this CSA-PSO-LR method.

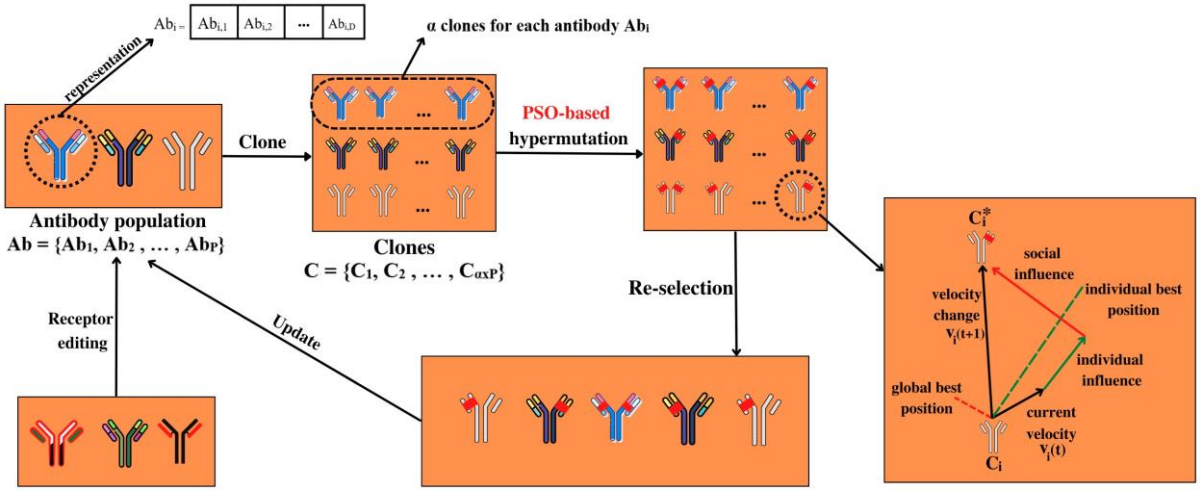

Figure 4.1 An illustration of the proposed method

The proposed model's process is outlined as follows:

1. Begin by setting the initial parameters of the model, including population size (P), memory size (B), the influence of previous velocity (α), and other factors like lower and upper bounds (lb, ub), and maximum number of evaluations (max_evals).
2. Create an initial group of P antibodies, represented by $\{Ab_1, Ab_2, \dots, Ab_P\}$, by assigning them random positions within specified limits as shown in Figure 4.2, following Equation 4.4.1. The position of each parameter for an antibody is randomly determined between its upper and lower bounds.

$$Ab_{i,j} = rand(0,1)\ x\ (ub_j - lb_j) + lb_j \qquad 4.4.1$$

where $rand(0,1)$ is a technique that produces a random number between 0 and 1, and lb_j is the antibody Abi's lower bound for the j^{th} parameter and ub_j is the antibody Abi's upper bound for the j^{th} parameter.

3. Determine the fitness, or affinity value, for each antibody using Equation 4.4.2. The fitness is inversely related to the cost function: the lower the cost, the higher the fitness. The aim is to identify the antibody with the highest fitness, thus the lowest cost.

$$f(Ab_i) = \frac{1}{1 + J(Ab_i)} \qquad 4.4.2$$

where $f(Ab_i)$ represents the affinity value of the antibody and $J(Ab_i)$ is the value of the cost function, as provided in Equation 4.3.4.

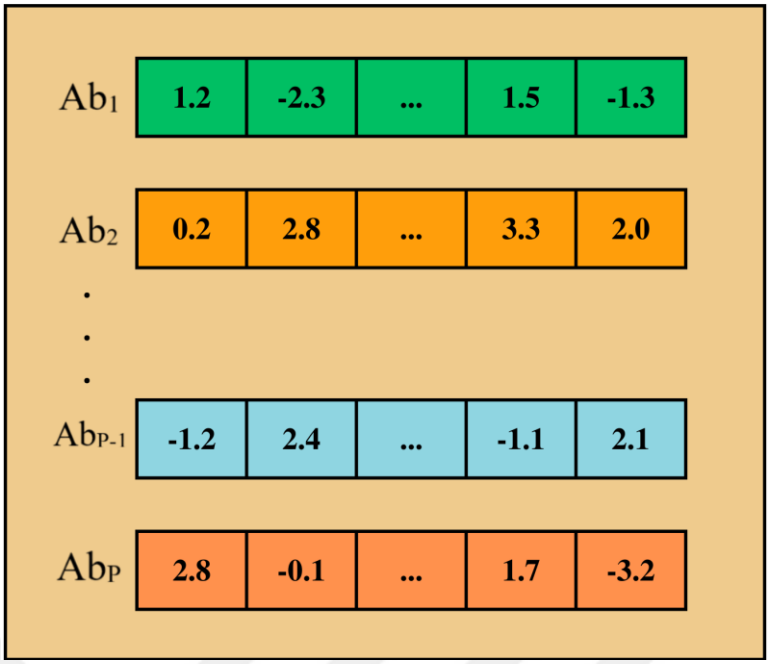

Figure 4.2 An initial population of antibodies chosen at random to correspond to various weight vectors

4. Clone each antibody Ab_i in the population to a specified degree, α, resulting in a new group of clones, $C = \{C_1, C_2, C_3, \dots, C_{\alpha.P}\}$
5. Apply hypermutation to each clone C_i using the PSO algorithm instead of the usual mutation methods. The current state of each clone C_i is analogous to a particle's position $X_i^{(t)}$, and its mutated state C_i^* is equivalent to the updated position $X_i^{(t+1)}$. The velocity and position updates for each clone C_i are calculated with Equations 4.4.3 and 4.4.4.

$$V_i^{(t+1)} = w.V_i^{(t)} + c_1.r_1.(p_{best} - C_i) + c_2.r_2.(g_{best} - C_i) \qquad 4.4.3$$

where , the current position of a clone is denoted by C_i, with p_{best} being the most optimal position that this particular clone has discovered up to the current moment. The term g_{best} represents the most optimal position any clone in the group has identified. Using this information, the next position of the clone is calculated with Equation 4.4.4. Through this process, a new version of the clone, indicated as Ci*, is generated, which is considered to be the mutated clone.

$$C_i^* = C_i + V_i^{(t+1)} \quad 4.4.4$$

6. To keep the antibody population stable, re-select the antibodies by retaining the clone with the highest affinity from each set of clones. Replace the original antibody with this selected clone.
7. Replace the worst performing %B antibodies with the best ones from the new clones, a process known as receptor editing, effectively refreshing the population with better candidates.

Algorithm 1 : The Proposed CSA-PSO Approach

1: *Identify the model's initial values for the parameters (max_evals, P, B, P, α).*
2: *Generate the initial population of P antibodies using Eq. 4.4.1*
3: *Determine the affinity value for each antibody Abi, using Eq. 4.4.2.*
4 : *Set the evaluation count to 0 at the beginning.*
5: ***while*** *evaluation count < max_evals* ***do***
6: ***for*** *each iteration* ***do***
7: *Create a population of clones C, using Eq. 4.1.1.*
8: *Apply Eq. 4.4.2. to determine each clone Ci's affinity value.*
9: ***for*** *each clone* C_i ***do***
10: *Determine the velocity change for clone Ci using Eq.4.4.3.*
11: *Utilizing Eq.4.4.4, update the location of the clone Ci to obtain the mutated clone Ci*.*
12: ***if*** $f(C_i^*) > f(C_i)$ ***then***
13: $C_i := Ci^*$
14: ***else***
15: *Ci := Ci*
16: ***end if***
17: ***end for***
18: *Increment evaluation count*
19: ***for*** *each antibody* Ab_i ***do***
20: *Pick the clone Cj that corresponds to Abi that has the highest affinity.*
21: *Abi := Ci*
22: ***end for***
23: *Change the worst B% number of antibodies with the newly created ones.*
24: ***end for***

Algorithm 1 outlines the algorithm of the model, with $'max\_evals'$ serving as an essential hyperparameter. This parameter sets the limit for the number of times the objective function will be assessed. It serves as a cutoff point for the optimization procedure, ensuring a compromise between the computational effort involved and the precision of the optimization outcomes. The evaluation count, which is started at zero and increased after each function evaluation, is iterated through the algorithm, changing the population of clones and antibodies until it reaches $'max\_evals'$.

Chapter 5

Experiments

The study introduces the development of a logistic regression approach, meticulously crafted for the detection of breast cancer, which stands out due to its integration of the CSA and PSO techniques for the optimization of the logistic regression model's training process. This hybrid CSA-PSO-LR model showcases a pioneering method in leveraging evolutionary computation algorithms to refine the logistic regression technique, hence improving its prognostic potential with regard to the diagnosis of breast cancer.

Using two well-known publicly available breast cancer datasets, extensive experiments were carried out to assess the effectiveness of the CSA-PSO-LR model, specifically the WDBC and WBCD. To guarantee an accurate evaluation of the model's performance, the datasets were meticulously divided into subsets for the training and testing stages of the model. Additionally, a rigorous 10-fold cross-validation methodology was employed.

In terms of performance comparison, the CSA-PSO-LR model was benchmarked against an array of leading machine learning algorithms, namely decision trees, extreme gradient boosting, k-nearest neighbors, logistic regression, random forests, and support vector machines.

To ensure the highest quality data is used by the model, a comprehensive data preparation strategy was implemented across both datasets. This included the scaling of all features to a consistent level to prevent any single feature from skewing the model’s predictions, and the management of missing values to maintain the datasets' completeness and accuracy.

Moreover, to optimize the classifier parameters and thereby augment the detection accuracy of the model, a sophisticated Bayesian optimization algorithm was employed, executing 200 iterations on each dataset. To ensure clarity and provide a comprehensive overview of the optimization process, Table 5.1. meticulously outlines the search ranges determined for the hyperparameters of each classifier included in the study.

Table 5.1 Defining the range of values for classifier hyperparameter tuning

Method	**Parameter Range**
LR	C : [0.0001, 1000]
DT	min_samples_split : [2,50]
	max_samples_split: [2,50]
RF	n_estimators : [5,100]
SVM	C : [0.001, 100]
KNN	n_neighbors : [2, 30]
	weights : [uniform, distance]
	metric : [minkowski, euclidean, manhattan]
XGB	learning_rate : [0.01, 0.3]
	n_estimators : [5, 200]
	subsample : [0.1, 1]
CSA-PSO-LR	lb: [-65, 20]
	ub: [20, 65]
	max_evals: [60000, 160000]
	B: [0.05, 0.2]
	Alpha: [1, 6]
	P : [20, 100]
	c1 : [0.5, 2.1]
	c2 : [0.5, 2.1]
	w : [0.4, 0.9]

Chapter 6

Results

This thesis presents detailed analysis where the Clonal Selection Algorithm and Particle Swarm Optimization-Logistic Regression (CSA-PSO-LR) is benchmarked against other established classifiers, with a comprehensive set of results outlined in **Table 6.1.** The CSA-PSO-LR exhibits a remarkable proficiency, outclassing competing models by achieving higher F1-scores and detection accuracies on the Wisconsin Diagnostic Breast Cancer (WDBC) and Wisconsin Breast Cancer Database (WBCD) datasets. The peak performance on the WDBC dataset is evidenced by an accuracy of 98.75% and an F1-score of 98.27%. This optimal performance is derived from a finely-tuned set of hyperparameters: a lower boundary (lb) of -16, an upper boundary (ub) of 46, maximum number on evaluations (max_evals) at 95,992, a B parameter of 0.18, a swarm size (P) of 51, an inertia weight (α) of 6, a personal learning coefficient (c1) of 0.708, a social learning coefficient (c2) of 1.45, and a velocity factor (w) of 0.43.

For the WBCD dataset, the CSA-PSO-LR sustains its robust performance, evidenced by achieving an accuracy rate of 97.94% and an F1-score of 97.53%, with an optimally configured hyperparameter set: lb of -33, ub of 19, max_evals set to 113,506, B at 0.073, P increased to 76, α adjusted to 5, c1 at a higher 1.736, c2 at 1.203, and w at a slightly increased value of 0.74. The strength and consistency of the CSA-PSO-LR classifier are further verified in Table 6.2, which provides insights into the model's classification accuracy and F1-measure performance metrics across a rigorous 10-fold cross-validation conducted on both WDBC and WBCD datasets. This validation process highlights the CSA-PSO-LR's dependable performance over varied segments of the datasets, reinforcing its credibility as a diagnostic predictive tool in the biomedical field.

A detailed empirical analysis is presented through the data synthesized in Table 6.3, which delineates the optimal hyperparameters for a suite of computational models. These hyperparameters were meticulously derived via a Bayesian optimization process, a

statistical strategy that iteratively improves the search for parameter sets that optimize model performance.

Table 6.1 Performance results of the CSA-PSO-LR and other classifiers using 10 fold cross-validation and Bayesian hyperparameter optimization techniques

Dataset	Methods	ACC(%)	F1(%)
WDBC	LR	98.24	98.13
	DT	93.68	93.48
	RF	96.13	95.91
	SVM	98.24	98.15
	KNN	97.18	97.05
	XGB	97.19	97.01
	CSA-PSO-LR	**98.75**	**98.27**
WBCD	LR	96.79	96.51
	DT	95.31	94.95
	RF	96.92	96.70
	SVM	97.22	97.02
	KNN	97.22	97.02
	XGB	96.93	96.69
	CSA-PSO-LR	**97.94**	**97.53**

Additionally, Table 6.4 broadens the comparative analysis by providing a thorough evaluation of the CSA-PSO-LR model's classifier performance metrics in comparison to a variety of other classification algorithms as documented in recent research. The WDBC and WBCD are the two separate datasets used in these comparisons. By contrasting the CSA-PSO-LR model's sampling strategies, accuracy rates, F1-scores, and computational training duration with those of other models proposed by researchers in the field, this comparative exposition carefully illustrates the model's effectiveness.

In examining the computational performance of various classification models as detailed in the provided Table 6.4, it becomes evident that the implementation of cross-validation as a sampling strategy, particularly when utilized in the CSA-PSO-LR method, has implications for the computational training time. Cross-validation, involves partitioning the data set into ten subsets, systematically using one subset for validation

and the remaining subsets for training, and iteratively rotating the validation subset through all partitions. This technique ensures a comprehensive evaluation of the model's performance but at the cost of increased computational time due to its repetitive nature.

Table 6.2. The results of classification accuracy and F1-measure metrics obtained by the proposed model for each fold

Dataset	Fold	ACC(%)	F1(%)
WDBC	1	98.21	98.87
	2	98.21	97.77
	3	98.21	97.43
	4	98.21	98.24
	5	96.42	95.83
	6	98.21	96
	7	100	100
	8	100	100
	9	100	100
	10	100	100
	Average	**98.75**	**98.27**
WBCD	1	97.05	97.14
	2	100	100
	3	98.52	98.11
	4	94.11	94.87
	5	95.58	95.65
	6	100	100
	7	97.05	94.73
	8	100	100
	9	98.52	97.77
	10	98.52	95.23
	Average	**97.94**	**97.53**

Conversely, studies employing a single split strategy, such as the 70-30 or 80-20 partitioning, inherently necessitate a reduced computational duration. This is attributed to the fact that the model is trained and validated on fixed non-overlapping segments of the data set, thus requiring less processing time. Table 6.4 confirms this, where models employing the 80-20 splitting strategy demonstrate markedly shorter training times in comparison to those utilizing cross-validation.

Table 6.3 Optimum hyperparameters found by Bayesian hyperparameter Optimizer

Method	**Parameter Range**	**WDBC**	**WBCD**
LR	C : [0.0001, 1000]	6.48	3.24
DT	min_samples_split : [2,50]	2	2
	max_samples_split: [2,50]	10	6
RF	n_estimators : [5,100]	72	39
SVM	C : [0.001, 100]	2.77	0.71
KNN	n_neighbors : [2, 30]	4	8
	weights : [uniform, distance]	distance	distance
	metric : [minkowski, euclidean, manhattan]	manhattan	euclidian
XGB	learning_rate : [0.01, 0.3]	0.16	0.19
	n_estimators : [5, 200]	193	138
	subsample : [0.1, 1]	0.38	0.48
CSA-PSO-LR	lb: [-65, 20]	-16	-33
	ub: [20, 65]	46	19
	max_evals: [60000, 160000]	95992	113506
	B: [0.05, 0.2]	0.18	0.073
	alpha: [1, 6]	6	5
	P : [20, 100]	51	76
	c1 : [0.5, 2.1]	0.708	1.736
	c2 : [0.5, 2.1]	1.45	1.203
	w : [0.4, 0.9]	0.43	0.74

It is thus observable that there exists a trade-off between the thoroughness of the model evaluation and the temporal efficiency of the training process. Models employing single-split strategies benefit from a more streamlined computational demand, leading to faster execution times, which can be a critical factor in time-sensitive applications. However, this expedience may come at the expense of the model's generalizability and the robustness of performance metrics. Therefore, while the CSA-PSO-LR model's application of 10x CV endows it with a potentially more reliable performance assessment, it does so by incurring greater computational time, which is consistently demonstrated across the referenced literature.

Table 6.4 Analysis of the CSA-PSO-LR's performance using the WBCD and WDBC data set in contrast to a few other models from the literature

Ref.	Year	Data Set	Method	SS	ACC	F1	TT (sec.)
[15]	2013	WDBC	I-CLONALG	80-20 %	98.58	NA	NA
[20]	2016	WDBC	PSO-SVM	10 CV	93.53	NA	42.3
[25]	2016	WDBC	PSO-KDE	10 CV	98.45	NA	NA
[55]	2018	WDBC	GRU-SVM	70-30 %	93.75	NA	NA
[16]	2020	WDBC	AISAC	5 CV	93.67	NA	35
[21]	2020	WDBC	PSO-NDS	NA	98.6	NA	NA
[23]	2021	WDBC	IPSO-IFA-RF	80-20 %	97	NA	1.12
[26]	2021	WDBC	EGWO-SVM	10 CV	98.24	97.40	NA
[28]	2021	WDBC	BOAALO-ANN	10 CV	98.16	NA	NA
[27]	2022	WDBC	WOA-SVM	10 CV	97.54	NA	NA
[14]	2022	WDBC	LR-RFE	80-20 %	98.06	97.36	0.42
[17]	2022	WDBC	CSA-MLP	5 CV	98.24	NA	NA
[18]	2023	WDBC	AISAC-MMD	NA	96.50	NA	NA
[29]	2023	WDBC	CSOA-wKNN	5 CV	97.36	97.77	NA
PM	**2024**	**WDBC**	**CSA-PSO-LR**	**10 CV**	**98.75**	**98.27**	**2.46**
[22]	2013	WBCD	PSO-ANN	50-50 %	97.36	NA	NA
[13]	2016	WBCD	SVM	10 CV	97.13	96	NA
[24]	2018	WBCD	PSO-C4.5	10 CV	96.49	NA	NA
[16]	2020	WBCD	AISAC	5 CV	97.42	NA	4.7
[19]	2020	WBCD	CLONALG-LFS	10 CV	95.48	NA	NA
[18]	2023	WBCD	AISAC-MMD	NA	96.90	NA	NA
PM	**2024**	**WBCD**	**CSA-PSO-LR**	**10 CV**	**97.94**	**97.53**	**3.03**

Chapter 7

Conclusions and Future Prospects

7.1 Conclusions

In the current research, a novel integrative computational model designated as CSA-PSO-LR has been developed for the precise classification of breast cancer tumors into benign and malignant categories. This innovative model harnesses the robust optimization capabilities of the Clonal Selection Algorithm (CSA) in tandem with the Particle Swarm Optimization (PSO) technique to effectively calibrate the weights in a logistic regression framework, thereby optimizing the cost function.

The strategic incorporation of the PSO algorithm within the hypermutation phase of the CSA algorithm significantly augments the model's convergence speed and bolsters its skill in local search, improving its forecasting abilities. To evaluate the robustness of the CSA-PSO-LR model, advanced Bayesian methods for hyperparameter tuning have been employed alongside the widely recognized 10-fold cross-validation approach on the WDBC and WBCD datasets, which are open to public access.

To further enhance the model's operational efficiency, a parallel processing approach has been implemented on the computational processing unit (CPU), effectively diminishing the duration required for the CSA-PSO-LR model's training phase. This method guarantees that the model runs in a reasonable amount of time and attains excellent precision.

To directly compare the performance metrics of the CSA-PSO-LR model with a number of sophisticated classifiers such as decision trees, extreme gradient boosting, k-nearest neighbors, logistic regression, random forests, and support vector machines, a comparison analysis has been carried out. The analysis's findings imply that the CSA-PSO-LR model excels in terms of accuracy and the F1-measure when compared to these other methods.

In light of the model's satisfactory and consistent performance across both WDBC and WBCD datasets, it is evident that the CSA-PSO-LR model establishes a new benchmark when evaluated against recent scholarly contributions. The empirical evidence derived from the dataset analyses conclusively demonstrates the model's exceptional capability to navigate the intricacies of complex and imbalanced datasets.

The utilization of metaheuristic optimization algorithms in this study, such as CSA and PSO, provides a substantial advantage due to their capability to escape local optima and discover more optimal solutions in complex search spaces. These algorithms are particularly adept at dealing with non-linear, high-dimensional, and multimodal function optimization, which are common challenges in medical data classification tasks. Their application in the CSA-PSO-LR model underscores the potential of hybrid optimization techniques to significantly elevate the performance of predictive modeling in the realm of medical diagnostics.

7.2 Societal Impact and Contribution to Global Sustainability

Incorporating Artificial Intelligence (AI) and Machine Learning (ML) within healthcare diagnosis represents a transformative shift with profound societal impacts and significant contributions to global sustainability. These technologies are not merely augmenting the capabilities of medical professionals but are reshaping the accessibility, efficiency, and accuracy of healthcare services on a global scale. The integration of AI and ML in diagnostic processes heralds a new era of precision medicine, where personalized treatment plans can be developed with a higher degree of specificity and effectiveness, thereby directly contributing to the enhancement of patient outcomes and the optimization of healthcare resources.

One of the most notable societal impacts of AI and ML in healthcare diagnosis is their potential to democratize access to high-quality medical services. In regions with limited access to medical experts, AI-powered diagnostic tools can serve as a critical resource, providing reliable diagnoses that would otherwise be inaccessible. This capability is especially crucial in managing diseases with high morbidity and mortality rates, where early detection can significantly alter treatment success rates. Moreover, the

efficiency introduced by AI and ML in processing and analyzing vast amounts of medical data not only accelerates the diagnostic process but also reduces the likelihood of human error, contributing to more sustainable healthcare systems by minimizing unnecessary treatments and optimizing the use of medical resources.

Furthermore, the integration of AI and ML into healthcare diagnostics aligns with the global sustainability goals by promoting health and well-being across diverse populations. By enabling early detection and more accurate diagnoses, these technologies play a pivotal role in reducing disease burden and improving life expectancy. Additionally, the data-driven insights generated by AI and ML can inform public health strategies and policies, leading to more effective disease prevention measures and health promotion initiatives. In essence, the adoption of AI and ML in healthcare diagnostics is not just an advancement in medical technology; it is a step toward realizing a more equitable, efficient, and sustainable global healthcare system.

7.3 Future Prospects

In the future, the potential exists to extend the capabilities of this proposed model to detect various other cancer types. This could lead to faster and more accurate diagnoses across the medical field. Additionally, the model's performance in breast cancer detection could be further enhanced by exploring alternative algorithms designed for more efficient and precise navigation and optimization of the search space.

BIBLIOGRAPHY

[1] "Breast cancer, 2021" Accessed: 2022-03-20. [Online]. Available: https://www.who.int/news-room/fact-sheets/detail/breast-cancer

[2] Cezary Wojtyla, Paola Bertuccio, Michal Ciebiera, and Carlo La Vecchia. Breast cancer mortality in the americas and australasia over the period 1980–2017 with predictions for 2025. Biology, 10(8), 2021. ISSN 2079-7737. doi: 10.3390/biology10080814.

[3] James S Lawson and Benjamin Heng. Viruses and breast cancer. Cancers, 2(2):752–772, 2010. doi: https://doi.org/10.3390/ cancers2020752.

[4] Ziyu He, Zhu Chen, Miduo Tan, Sauli Elingarami, Yuan Liu, Taotao Li, Yan Deng, Nongyue He, Song Li, Juan Fu, et al. A review on methods for diagnosis of breast cancer cells and tissues. Cell proliferation, 53(7):e12822, 2020.

[5] "Mammograms for breast cancer detection." Accessed: 2023-11-20. [Online]. Available: https://www.cancer.org/cancer/breast-cancer/screening-tests-and-early-detection/mammograms.html

[6] "Breast Ultrasound." Accessed: 2023-11-20. [Online]. Available: https://www.radiologyinfo.org/en/info/breastus

[7] "Breast MRI Scans" Accessed: 2023-11-23. [Online]. Available: https://www.breastcancer.org/symptoms/testing/types/mri

[8] Yue, W., Wang, Z., Chen, H., Payne, A., & Liu, X. (2018). Machine learning with applications in breast cancer diagnosis and prognosis. *Designs*, *2*(2), 13.

[9] L.N. de Castro and F.J. Von Zuben. Learning and optimization using the clonal selection principle. IEEE Transactions on Evolutionary Computation, 6(3):239–251, 2002. doi: 10.1109/TEVC.2002. 1011539.

[10] J. Kennedy and R. Eberhart. Particle swarm optimization. In Proceedings of ICNN'95 - International Conference on Neural Networks, volume 4, pages 1942–1948 vol.4, 1995. doi: 10.1109/ICNN.1995. 488968.

[11] Y. Shi and R.C. Eberhart. Empirical study of particle swarm optimization. In Proceedings of the 1999 Congress on Evolutionary Computation-CEC99 (Cat. No. 99TH8406), volume 3, pages 1945– 1950 Vol. 3, 1999. doi: 10.1109/CEC.1999.785511.

[12] Lining Zhang, Maoguo Gong, Licheng Jiao, and Jie Yang. Optimal approximation of linear systems by an improved clonal selection algorithm. In 2008 IEEE Congress on Evolutionary Computation (IEEE World Congress on Computational Intelligence), pages 527– 534, 2008. doi: 10.1109/CEC.2008.4630847.

[13] Hiba Asri, Hajar Mousannif, Hassan Al Moatassime, and Thomas Noel. Using machine learning algorithms for breast cancer risk prediction and diagnosis. Procedia Computer Science, 83:1064– 1069, 2016. ISSN 1877-0509. doi: https://doi.org/10.1016/j.procs. 2016.04.224. The 7th International Conference on Ambient Systems, Networks and Technologies (ANT 2016) / The 6th International Conference on Sustainable Energy Information Technology (SEIT2016) / Affiliated Workshops.

[14] Abdur Rasool, Chayut Bunterngchit, Luo Tiejian, Md Ruhul Islam, Qiang Qu, and Qingshan Jiang. Improved machine learning-based predictive models for breast cancer diagnosis. International journal of environmental research and public health, 19(6):3211, 2022. doi: 10.3390/ijerph19063211.

[15] Ryma Daoudi, Khalifa Djemal, and Abdelkader Benyettou. Cells clonal selection for breast cancer classification. In 10th International Multi-Conferences on Systems, Signals & Devices 2013 (SSD13), pages 1–4. IEEE, 2013. doi: 10.1109/SSD.2013.6564016.

[16] David González-Patiño, Yenny Villuendas-Rey, Amadeo José Argüelles-Cruz, Oscar Camacho-Nieto, and Cornelio YáñezMárquez. Aisac: An artificial immune system for associative classification applied to breast cancer detection. Applied Sciences, 10(2):515, 2020. doi: 10.3390/app10020515.

[17] Ali Al Bataineh, Devinder Kaur, and Seyed Mohammad J. Jalali. Multi-layer perceptron training optimization using nature inspired computing. IEEE Access, 10:36963–36977, 2022. doi: 10.1109/ ACCESS.2022.3164669.

[18] David González-Patiño, Yenny Villuendas-Rey, Magdalena SaldañaPérez, and Amadeo-José Argüelles-Cruz. A novel bioinspired algorithm for mixed and incomplete breast cancer data classification. International Journal of Environmental Research and Public Health, 20(4), 2023. ISSN 1660-4601. doi: 10.3390/ijerph20043240.

[19] Yi Wang and Tao Li. Local feature selection based on artificial immune system for classification. Applied Soft Computing, 87: 105989, 2020. ISSN 1568-4946. doi: https://doi.org/10.1016/j.asoc. 2019.105989.

[20] Abbas Ahmadi and Parnian Afshar. Intelligent breast cancer recognition using particle swarm optimization and support vector machines. Journal of Experimental & Theoretical Artificial Intelligence, 28 (6):1021–1034, 2016. doi: https://doi.org/10.1080/0952813X.2015. 1055828.

[21] Senthilkumar Mohan, Sweta Bhattacharya, Rajesh Kaluri, Guang Feng, Usman Tariq, et al. Multi-modal prediction of breast cancer using particle swarm optimization with non-dominating sorting. International Journal of Distributed Sensor Networks, 16(11), 2020. doi: https://doi.org/10.1177/1550147720971505

[22] Hasan Koyuncu and Rahime Ceylan. Artificial neural network based on rotation forest for biomedical pattern classification. In 2013 36th International conference on telecommunications and signal processing (TSP), pages 581–585. IEEE, 2013 doi: 10.1109/TSP.2013. 6614001.

[23] Moolchand Sharma, Shubbham Gupta, and Suman Deswal. Modified bio-inspired algorithms for diagnosis of breast cancer using aggregation. International Journal of Innovative Computing and Applications, 12(1):37–47, 2021. doi: https://dx.doi.org/10.1504/ IJICA.2021.10036070.

[24] Much Muslim, Siti Hardiyanti, E Sugiharti, Budi Prasetiyo, and Siti Alimah. Optimization of c4.5 algorithm-based particle swarm optimization for breast cancer diagnosis. Journal of Physics: Conference Series, 983:012063, 03 2018. doi: 10.1088/1742-6596/983/1/012063.

[25] Razieh Sheikhpour, Mehdi Agha Sarram, and Robab Sheikhpour. Particle swarm optimization for bandwidth determination and feature selection of kernel density

estimation based classifiers in diagnosis of breast cancer. Applied Soft Computing, 40:113–131, 2016. doi: https://doi.org/10.1016/j.asoc.2015.10.005.

[26] Sunil Kumar and Maninder Singh. Breast cancer detection based on feature selection using enhanced grey wolf optimizer and support vector machine algorithms. Vietnam Journal of Computer Science, 8(02): 177–197, 2021. doi: https://doi.org/10.1142/S219688882150007X.

[27] Ahmed S. Elkorany, Mohamed Marey, Khaled M. Almustafa, and Zeinab F. Elsharkawy. Breast cancer diagnosis using support vector machines optimized by whale optimization and dragonfly algorithms. IEEE Access, 10:69688–69699, 2022. doi: 10.1109/ACCESS.2022. 3186021.

[28] Shankar Thawkar, Satish Sharma, Munish Khanna, and Law kumar Singh. Breast cancer prediction using a hybrid method based on butterfly optimization algorithm and ant lion optimizer. Computers in Biology and Medicine, 139:104968, 2021. ISSN 0010-4825. doi: https://doi.org/10.1016/j.compbiomed.2021.104968.

[29] S.R. Sannasi Chakravarthy, N. Bharanidharan, and H. Rajaguru. Deep learning-based metaheuristic weighted k-nearest neighbor algorithm for the severity classification of breast cancer. IRBM, 44(3):100749, 2023. ISSN 1959-0318. doi: https://doi.org/10.1016/j.irbm.2022. 100749.

[30] Jasper Snoek, Hugo Larochelle, and Ryan P Adams. Practical bayesian optimization of machine learning algorithms. Advances in neural information processing systems, 25, 2012.

[31] "Breast cancer wisconsin (diagnostic) data set, 1995." Accessed: 2023-10-05. [Online]. Available: https://archive.ics.uci.edu/dataset/17/breast+cancer+wisconsin+diagnostic

[32] "Breast cancer wisconsin (original) data set, 1995." Accessed: 2023-10-05. [Online]. Available: https://archive.ics.uci.edu/dataset/15/breast+cancer+wisconsin+original

[33] Bilge Kagan Dedeturk and Bahriye Akay. Spam filtering using a logistic regression model trained by an artificial bee colony algorithm. Applied Soft Computing, 91:106229, 2020. ISSN 1568-4946. doi: https://doi.org/10.1016/j.asoc.2020.106229.

[34] Burak Kolukisa and Burcu Bakir-Gungor. Ensemble feature selection and classification methods for machine learning-based coronary artery disease diagnosis. Computer Standards Interfaces, 84:103706, 2023. ISSN 0920-5489. doi: https://doi.org/10.1016/j.csi.2022. 103706.

[35] Tin Kam Ho and M. Basu. Complexity measures of supervised classification problems. IEEE Transactions on Pattern Analysis and Machine Intelligence, 24(3):289–300, 2002. doi: 10.1109/34.990132.

[36] Pedregosa, F., Varoquaux, Ga"el, Gramfort, A., Michel, V., Thirion, B., Grisel, O., … others. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research, 12*(Oct), 2825–2830.

[37] Hastie, T., Tibshirani, R., & Friedman, J. (2009). The Elements of Statistical Learning: Data Mining, Inference, and Prediction (2nd ed.). Springer Series in Statistics.

[38] Vince, A. (2002). A framework for the greedy algorithm. *Discrete Applied Mathematics*, *121*(1-3), 247-260.

[39] Ólafsson, S. (2006). Metaheuristics. *Handbooks in operations research and management science*, *13*, 633-654.

[40] Rutenbar, R. A. (1989). Simulated annealing algorithms: An overview. *IEEE Circuits and Devices magazine*, *5*(1), 19-26.

[41] Brandão, J. (2004). A tabu search algorithm for the open vehicle routing problem. *European Journal of Operational Research*, *157*(3), 552-564.

[42] Tang, K. S., Man, K. F., Kwong, S., & He, Q. (1996). Genetic algorithms and their applications. *IEEE signal processing magazine*, *13*(6), 22-37.

[43] Blum, C. (2005). Ant colony optimization: Introduction and recent trends. *Physics of Life reviews*, *2*(4), 353-373.

[44] Karaboga, D. (2005). *An idea based on honey bee swarm for numerical optimization* (Vol. 200, pp. 1-10). Technical report-tr06, Erciyes university, engineering faculty, computer engineering department.

[45] Qin, A. K., Huang, V. L., & Suganthan, P. N. (2008). Differential evolution algorithm with strategy adaptation for global numerical optimization. *IEEE transactions on Evolutionary Computation*, *13*(2), 398-417.

[46] Ruder, S. (2016). An overview of gradient descent optimization algorithms. *arXiv preprint arXiv:1609.04747*.

[47] Cortes, C., & Vapnik, V. (1995). Support-vector networks. Machine Learning, 20(3), 273–297.

[48] Charles R Harris, K Jarrod Millman, Stéfan J Van Der Walt, Ralf Gommers, Pauli Virtanen, David Cournapeau, Eric Wieser, Julian Taylor, Sebastian Berg, Nathaniel J Smith, et al. Array programming with numpy. Nature, 585(7825):357–362, 2020. doi: https://doi.org/ 10.1038/s41586-020-2649-2.

[49] Burak Kolukisa, Bilge Kagan Dedeturk, Hilal Hacilar, and Vehbi Cagri Gungor. An efficient network intrusion detection approach based on logistic regression model and parallel artificial bee colony algorithm. Computer Standards Interfaces, 89:103808, 2024. ISSN 0920- 5489. doi: https://doi.org/10.1016/j.csi.2023.103808. URL https: //www.sciencedirect.com/science/article/pii/S0920548923000892.

[50] "Pypi : Abc-lr, 2022." Accessed: 2023-08-20. [Online]. Available: https://pypi.org/project/abcLR/.

[51] "Github : Abc-lr, 2022." Accessed: 2023-08-21. [Online]. Available: https://github.com/kagandedeturk/ ABC-LR.

[52] Jasper Snoek, Hugo Larochelle, and Ryan P Adams. Practical bayesian optimization of machine learning algorithms. Advances in neural information processing systems, 25, 2012.

[53] Janeway Jr, C. A. (2001). How the immune system works to protect the host from infection: a personal view. Proceedings of the National Academy of Sciences, 98(13), 7461-7468.

[54] David R Cox. The regression analysis of binary sequences. *Journal of the Royal Statistical Society: Series B (Methodological)*, 20(2):215–232, 1958.

[55] Abien Fred M Agarap. On breast cancer detection: an application of machine learning algorithms on the wisconsin diagnostic dataset. In Proceedings of the 2nd international conference on machine learning and soft computing, pages 5–9, 2018. doi: 10.1145/3184066.3184080.

CURRICULUM VITAE

2016 – 2021	B.Sc., Computer Engineering, Erciyes University, Kayseri, TURKEY
2022 – Present	M.Sc., Electrical and Computer Engineering, Abdullah Gul University, Kayseri, TURKEY
2022 – Present	Research Assistant, Computer Engineering Abdullah Gul University, Kayseri, TURKEY