An Improved Support Vector Machine for medical diagnosis

A

report submitted in partial fulfillment for the award of the degree of

Bachelors of Technology

in

Information Technology

By

Mogalluru Chidhvilas Tanay : 2020IMT-060

Under the Supervision of

Dr. Avadh kishor

Department of Computer Science and Engineering

विश्वजीवनामृतं ज्ञानम्

ABV-INDIAN INSTITUTE OF INFORMATION TECHNOLOGY
AND MANAGEMENT GWALIOR
GWALIOR, INDIA

DECLARATION

I hereby certify that the work, which is being presented in the report/thesis, entitled An Improved Support Vector Machine for medical diagnosis, in fulfillment of the requirement for the award of the degree of Bachelor of Technology and submitted to the institution is an authentic record of my/our own work carried out during the period Jan-2023 to May-2023 under the supervision of Dr. Avadh Kishor. I also cited the reference about the text(s)/figure(s)/table(s) from where they have been taken.

Dated: **Signature of the candidate**

This is to certify that the above statement made by the candidates is correct to the best of my knowledge.

Dated: **Signature of supervisor**

Acknowledgements

I am highly indebted to Dr.Avadh Kishor , for his esteemed mentorship, and for allowing me to freely explore and experiment with various ideas in the course of making this project a reality. The leeway I was given went a long way towards helping cultivate a genuine hunger for knowledge and keeping up the motivation to achieve the best possible outcome. I can genuinely say that this Bachelor's Thesis Project (BTP) made me explore many areas of machine learning that are new to me, and kindled an interest to further follow up on some of those areas. Moreover, the semi-successful completion of this project has brought with it great satisfaction and more importantly, confidence in my ability to produce more high-quality non-trivial artificial intelligence systems that can make a difference in the real-world. I would like to sincerely express my gratitude to this prestigious institution for providing me and my colleagues with the opportunity to pursue this BTP. It is an honor to be able to work on such an important academic project under the guidance and support I am provided with. I am grateful for the resources and facilities provided by this institution, which have been instrumental in enabling me to conduct my research and complete this project. Moreover, I deeply appreciate the efforts of my professors in mentoring and fairly evaluating our works.

Mogalluru Chidhvilas Tanay

Abstract

In this project titled "An Improved Support Vector Machine for medical diagnosis", we propose a hybrid optimization algorithm called Multi-Swarm Whale Optimizer (MSWO) and integrate it with Support Vector Machines (SVM) for medical diagnosis using the Wisconsin Breast Cancer dataset from scikit-learn. The aim of this study is to develop an accurate and efficient model, an improvement upon the traditional SVM for all types of medical diagnosis in general.

The MSWO algorithm is inspired by the social behaviour of humpback whales and utilises multiple swarms to explore the solution space. Each swarm represents a cluster of whales, and the optimization process involves iteratively updating the positions of whales within each cluster. We use MSWO to enhance the performance of the SVM classifier, a powerful machine learning technique widely used in medical diagnosis tasks.

The Wisconsin Breast Cancer dataset used in this report provides features extracted from digitised images of fine needle aspirate (FNA) samples of breast masses. It consists of various attributes such as mean radius, mean texture, mean smoothness, and more. The dataset is labelled with two classes: malignant and benign, representing cancerous and non-cancerous cases, respectively. To evaluate the effectiveness of our proposed approach, we apply the MSWO algorithm in combination with the SVM classifier to train a model on the Wisconsin Breast Cancer dataset and later perform a comparison with the results. The optimization process involves iteratively updating the positions of whales (i.e the hyperparameters and feature subset make up the position set) within each swarm, with the objective of minimising the classification error and maximising accuracy and finally obtaining the hyper parameters which best control the trade-off.

Experimental results demonstrate that our proposed Multi-Swarm Whale Optimizer boosted Support Vector Machine achieves promising results in terms of accuracy, efficiency, AUC, Specificity, Sensitivity ,etc in disease diagnosis. The optimised model effectively classifies breast masses as malignant or benign, providing valuable insights for medical practitioners in making informed diagnostic decisions.

Keywords: *Machine Learning, Support Vector Machines, Bio-inspired meta heuristic Whale Optimization Algorithm, K means clustering, Feature Selection, Binary Encod-*

ing.

Contents

List of Figures

List of Tables

List of Acronyms

SVM	Support Vector Machine
WOA	Whale Optimization Algorithm
MSWOA	Multi-Swarm Whale Optimization Algorithm
MSWOAFS	Multi-Swarm Whale Optimization Algorithm with feature selection
BE	Binary Encoding
FNA	Fine needle aspirate
RBF	Radial Basis Function

1

Introduction

This Chapter dives into the realm of medical diagnosis enhancement using advanced machine learning techniques. It introduces the integration of Support Vector Machines and the Whale Optimization Algorithm as a solution to the challenges in accurate disease classification. By combining these methodologies, we aim to push the boundaries of medical diagnosis accuracy and contribute to improved patient care.

1.1 Introduction

Medical diagnosis plays a critical role in determining appropriate treatment and care for patients. In this project, we propose a novel approach, "An Improved Support Vector Machine for medical diagnosis" to enhance the accuracy of medical diagnoses using machine learning techniques. We leverage the power of the WOA and SVM classifier to optimize the diagnostic models.

The WOA, inspired by the hunting behavior of humpback whales, serves as a meta-heuristic optimization algorithm that guides the search for optimal solutions. It facilitates the exploration of diverse regions in the solution space, enhancing the chances of finding the most suitable parameters for the SVM model and thereby overcoming the problems of over-fitting and error minimization.

.

To enhance the exploration capabilities of the WOA, we employ K-means clustering to group the whale population into distinct clusters or whale swarms. This clustering technique helps ensure that the optimization process covers various regions of the solution space, enabling comprehensive exploration and enhancing the chances of discovering better solutions with higher convergence speed by avoiding premature convergence.

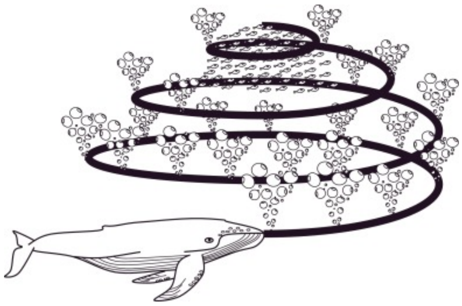

Figure 1.1: Movement of Humpback whales

1.2 Support Vector Machines in Medical Diagnosis

Support Vector Machine machine Learning models have emerged as a cornerstone in medical diagnosis due to their unique capabilities in handling intricate and non-linear relationships within complex datasets. They have become particularly invaluable in disease classification tasks where precision and interpretability are paramount.

One of the key strengths of SVMs lies in their ability to handle high-dimensional feature spaces commonly encountered in medical datasets. Medical data often includes a multitude of features extracted from various diagnostic tests and measurements. SVMs can effectively manage this dimensionality, preventing the ”curse of dimensionality” and allowing for accurate classification even in scenarios where feature-to-sample ratios are skewed. SVMs are adept at identifying subtle patterns and relationships within medical data, making them ideal for early disease detection. They excel in scenarios where classes are not easily separable and where the decision boundary is intricate. By employing the kernel trick, SVMs can transform the original feature space into a higher-dimensional space, revealing hidden patterns that may not be evident in lower dimensions.

1.3 Motivation

Medical diagnosis is a critical task in the healthcare domain, as accurate and timely identification of diseases can significantly impact patient outcomes and treatment plans. With the increasing availability of medical data and advancements in machine learning techniques, researchers and healthcare professionals have been exploring novel approaches to improve diagnostic accuracy and efficiency. One of the primary challenges in medical diagnosis is the high-dimensional nature of medical datasets, where the number of features can be substantial compared to the number of samples. Traditional machine learning algorithms may struggle to handle such datasets efficiently, leading to suboptimal performance and increased computational costs. Therefore, there is a growing demand for more sophisticated and robust optimization techniques that can effectively handle high-dimensional data and improve the accuracy of medical diagnosis.

In response to these challenges, this project aims to integrate two powerful techniques: WOA and SVms, to enhance medical diagnosis processes. The WOA, inspired by the hunting behaviour of humpback whales, is a bio-inspired meta heuristic optimization algorithm known for its ability to efficiently explore complex search spaces and find optimal solutions. On the other hand, SVMs are well-established supervised learning algorithms known for their effectiveness in handling high-dimensional data and binary classification tasks. The primary motivation behind this project is to leverage the strengths of both WOA and SVMs to address the challenges posed by medical diagnosis tasks.

2

A Review on SVM disease diagnosis

This chapter responds to the significant amount of research assigned to understanding the efficient methods of medical diagnosis. We examine some of the literature and briefly review the development of former proposed methods and the research gaps.

2.1 Review on Existing methods of medical diagnosis using SVMs

Medical diagnosis is a critical field that heavily relies on accurate and efficient classification techniques to aid in disease identification, prognosis, and treatment planning. Over the years, researchers have explored various machine learning algorithms to enhance the accuracy and reliability of medical diagnosis. This literature review aims to provide an overview of the existing research and methodologies in the domain of medical diagnosis using machine learning techniques, focusing on the integration of SVMs and optimization algorithms like the WOA.

Support Vector Machines have garnered significant attention in medical diagnosis due to their ability to create optimal decision boundaries in high-dimensional feature spaces. Researchers have successfully employed SVMs to classify medical data, such as identifying tumors in mammograms, diagnosing heart diseases, predicting the onset of diabetes, predicting the onset of fatal diseases like Parkinson's disease [1] and cervical cancers [2]. The margin maximization characteristic of SVMs makes them particularly useful for binary classification tasks in medical applications.

While SVMs have demonstrated effectiveness, they face certain limitations. SVMs rely heavily on proper hyperparameter tuning, and choosing the right values for parameters like C and γ is crucial. Additionally, SVMs might struggle with high-dimensional data, as they can be prone to overfitting or might require significant computational resources for preventing overfitting [3].

To address these limitations, researchers have explored integrating optimization algorithms with SVMs. The Whale Optimization Algorithm [4] is one such nature-inspired metaheuristic that shows promise in enhancing the performance of SVMs. The WOA's exploration and exploitation capabilities can help in optimizing SVM hyperparameters, leading to improved classification accuracy.

Various combinations of standard Algorithms like Particle Swarm Optimization [5] and novel Algorithms like Hunting Search [6] , Despite the potential benefits of integrating WOA with SVMs for medical diagnosis, there remains a gap in the literature regarding comprehensive studies on this hybrid approach. While individual works exist on WOA and SVMs separately, a systematic exploration of their combined potential in medical diagnosis is lacking. This research seeks to fill this gap by presenting a detailed analysis of the hybrid algorithm's performance, its impact on accuracy, and its effectiveness in addressing the challenges associated with medical diagnosis.

2.2 Research Analysis of Existing Methods

Existing methods for medical diagnosis often face several limitations that hinder their accuracy and reliability. These limitations stem from challenges inherent to both traditional SVMs and standalone optimization algorithms. However, our project aims to address these limitations by introducing a hybrid approach that combines SVMs and the WOA, offering solutions to existing challenges in medical diagnosis.

One of the major challenges of traditional SVMs lies in the complexity of hyperparameter tuning. Optimal configuration of hyperparameters like C and gamma significantly impacts the model's performance. This process often requires domain expertise and can be time-consuming, particularly in the medical domain where domain knowledge is essential. Moreover, the dimensionality of medical datasets poses another hurdle. High-dimensional spaces can lead to overfitting issues, especially in the context of SVMs. As a result, existing methods might struggle to generalize well to new and unseen data, compromising the accuracy of medical diagnosis predictions.

Standalone optimization algorithms, while promising, might lack the adaptability required to fine-tune model parameters effectively in the context of medical datasets. This limitation is particularly relevant as medical data can vary widely and exhibit complex patterns. In response to these challenges, our project introduces a novel hybrid model that marries the strengths of both SVMs and the WOA. By combining SVMs' powerful classification capabilities with WOA's efficient optimization techniques, we are able to offer solutions that address the shortcomings of existing methods.

One notable advantage of our hybrid approach is the automated hyperparameter tuning. WOA's exploration and exploitation capabilities enable the model to automatically fine-tune hyperparameters, eliminating the need for manual configuration. This not only enhances the model's accuracy but also reduces the complexity associated with hyperparameter tuning.

Furthermore, the hybrid model employs WOA for binary encoding-based feature selection. By selecting relevant features, the model reduces dimensionality and mitigates the risk of overfitting, a common concern in medical diagnosis due to high-dimensional data.

2.3 Research Gaps

The proposed project, emerges to fill significant research gaps in the realm of medical diagnosis, particularly in the context of enhancing classification accuracy and feature selection using a novel hybridization of optimization and machine learning techniques. Through the synergy of the WOA and SVM, the project tackles several critical research gaps.

Existing medical diagnosis systems often struggle with achieving consistently high classification accuracy. This project steps in to bridge this gap by leveraging the combined power of WOA and SVM. The enhanced classification accuracy ensures more reliable and precise disease identification, enabling clinicians to make informed decisions about patient care.

The project addresses the challenge of optimal hyperparameter tuning in SVMs. While SVMs are powerful classifiers, selecting the right values for hyperparameters like C and gamma is non-trivial. By integrating the WOA, the project optimizes these hyperparameters effectively, enhancing SVM performance and achieving superior classification results. Medical datasets often contain a multitude of features, many of which may be irrelevant or redundant. The project introduces a research innovation by employing binary encoding and feature selection. This technique helps in identifying and utilizing the most relevant features for classification, reducing computational complexity and preventing overfitting. The utilization of multi-swarm strategy with K-means clustering is another novel aspect of the project. This approach addresses the limitation of conventional optimization techniques that often get trapped in local optima. By dividing the population into distinct swarms, the project enables a diverse exploration of the search space, leading to more robust solutions.

Research in the field often focuses on individual optimization techniques or machine learning algorithms. The project's hybrid model introduces a holistic approach by combining WOA, SVM, and K-means clustering. This unique combination addresses the research gap of limited integration between optimization and classification techniques.

While some methods might yield higher accuracy on specific datasets, they might lack consistency and generalization across different datasets. The proposed hybrid model aims to offer consistent improvements across various medical datasets, addressing the research gap of algorithm consistency and robustness.

3

Problem Statement based on Identified Research Gaps

This chapter explains the formulation of the problem that this thesis addresses, as well as it outlines the thesis objectives.

3.1 Problem Formulation

Medical diagnosis accuracy remains a fundamental challenge, often impeding accurate disease identification and prognosis. Conventional techniques struggle with achieving consistent and high classification accuracy, hampering the effectiveness of medical decision-making. Furthermore, optimal tuning of hyperparameters in classification algorithms, such as SVM, remains elusive. The lack of a comprehensive and integrated approach to address these issues hampers the progress of accurate medical diagnosis.

In addition, the complex nature of medical datasets, characterized by high-dimensional features, necessitates efficient and robust feature selection techniques. Traditional methods often fall short in identifying the most relevant features, leading to increased computation time and susceptibility to overfitting. Moreover, there's a need to explore the search space more diversely to overcome the limitations of local optima encountered by conventional optimization techniques.

Addressing these challenges requires an innovative solution that integrates advanced optimization algorithms, classification techniques, and feature selection mechanisms. A novel approach that optimizes hyperparameters effectively, enhances feature selection, and promotes diverse exploration of the search space is imperative.

This project aims to bridge the identified research gaps by proposing a multi-swarm whale optimizer boosted SVM model. By synergizing the WOA with SVM, the project seeks to enhance medical diagnosis accuracy by providing consistent and robust classification results across diverse medical datasets. The integration of K-means clustering for multi-swarm optimization and binary encoding for feature selection further elevates the model's efficiency and performance.

The project's significance lies in its potential to improve medical diagnosis methodologies. The developed model not only promises to offer superior classification accuracy but also provides a holistic approach that overcomes the limitations of existing techniques. By addressing the identified research gaps, this project contributes to the advancement of accurate medical diagnosis, ultimately leading to improved patient care and healthcare decision-making.

The attempt to create an innovative and integrated framework that optimizes classification accuracy, hyperparameter tuning, and feature selection in medical diagnosis forms the crux of this research, opening new avenues for the application of machine learning in healthcare.

3.2 Project Objectives

- **Advancing Medical Diagnosis Accuracy:** The primary objective of this project is to enhance the precision and reliability of medical diagnoses through the adept use of machine learning techniques. By harnessing nature-inspired meta-heuristic optimization techniques and classification algorithms, the project aims to improve disease identification, prognosis, and treatment planning for healthcare practitioners. This approach effectively addresses the complexities posed by multi-dimensional medical datasets, leading to heightened diagnostic accuracy.

- **Integration of Whale Optimization Algorithm:** This project introduces a needed advancement in SVM diagnosis, with the integration of WOA. Derived from the strategic hunting patterns of humpback whales, WOA offers a distinctive optimization solution. This algorithm simultaneously handles the problems of hyperparameter optimization and feature selection. By making use of WOA's exploration and exploitation capabilities, the project achieves a substantial improvement in parameter optimization and feature selection leading to a significant enhancement in SVM classifier performance.

- **Controlled Feature Selection with Binary Encoding:** Reducing the complexity of high-dimensional feature spaces within high-precision medical datasets necessitates innovative approaches. This project employs a strategic feature selection technique involving binary encoding. Through this mechanism, relevant features for classification are precisely identified. The binary encoding of whale positions determines the features which are to be considered by the SVM classifier during classification. This approach not only reduces computational load but also mitigates the risks of overfitting, ensuring a robust and accurate diagnostic framework.

- **Dynamic Multi-Swarm Strategy via K-means Clustering:** To amplify the efficiency and speed of the optimization process, the project adopts a dynamic multi-swarm strategy facilitated by K-means clustering. This strategy divides the whale population into distinct swarms, each representing a unique subgroup. This diversification facilitates the exchange of solutions, thereby increasing the likelihood of discovering optimal solutions and avoiding local optima. The application of K-means clustering ensures that swarms remain adaptable to evolving optimization landscapes, effectively mitigating premature convergence and yielding solutions which are globally optimal.

4

Proposed Methodology

This chapter provides a comprehensive discussion of the methodology employed in the project, Algorithms used and the structure of the Hybird model implemented.

amsmath

4.1 Support Vector Machines

SVM is a powerful supervised machine learning algorithm used for classification and regression tasks. In this project, the SVM acts as the base classifier, and the optimization process seeks to determine the optimal hyperparameters (C and gamma) that maximise classification accuracy.

The support vector machines classify the datasets with a given set of features using a hyperplane equation, as the name suggests it represents a multidimensional hyperplane separating the two label classes, thus doing binary classification. The equation of the hyperplane is obtained in the data training phase by finding out two support vectors such that when a hyperplane is drawn between them provide maximum separation between the two label classes, for higher chances of classifying the new data points correctly. The equation for hyperplane is of the form:

$$w^T x + b = 0 \tag{4.1}$$

The equation to maximise the margin or the separation between the support vectors belonging to the two different labels is given by:

Maximising the distance between the two considered support vectors lets say x_1 and x_2.

$$|x_1 - x_2| = \max\left(\frac{2}{\|w\|^2}\right) \tag{4.2}$$

$$\text{Min}\left(\frac{\|w\|^2}{2} + C\sum_{i=1}^{l}\xi_i\right) \tag{4.3}$$

Here we can observe the C variable which is the penalty parameter, which allows for false classification of few of the data points along with an appropriate penalty value. Where w represents the normal vector representing the hyperplane, x represents the data point which is to be classified and b is the bias from the origin.

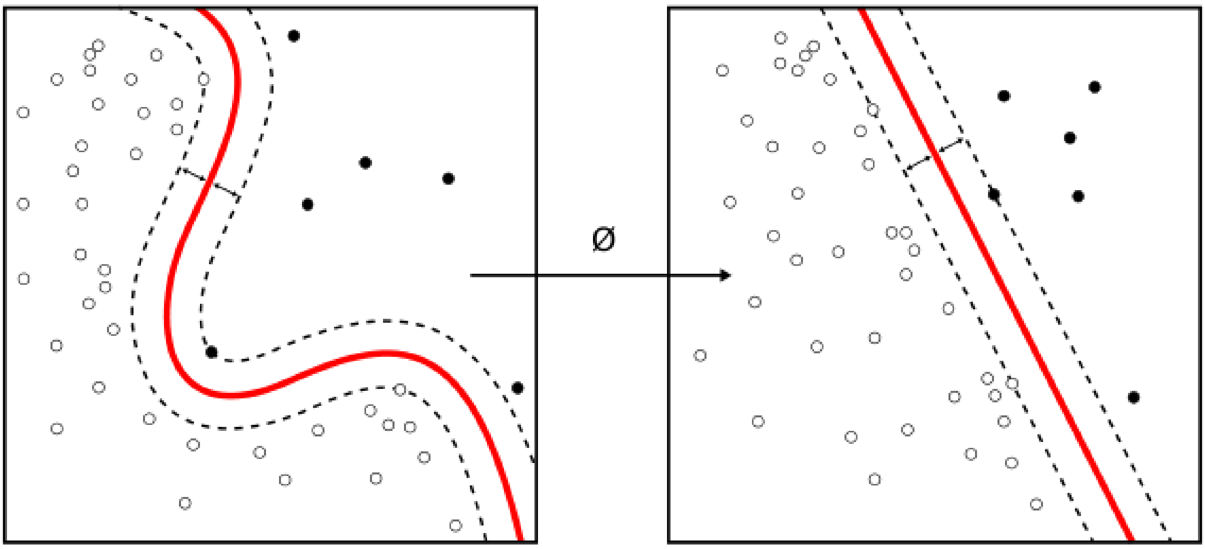

Figure 4.1: Significance of Regularization parameter

In most practical cases, the data used for training is not linearly separable, hence mathematical functions like Gaussian kernel (radial basis function), polynomial kernel, linear kernel or sigmoid kernel are used to map these data points into higher dimensions thus converting them into separable data.

$$K(x, x_i) = \exp(-\gamma \left\| x - x_i \right\|^2) \tag{4.4}$$

Where γ represents the width of the kernel and is manually set before the training of the model.

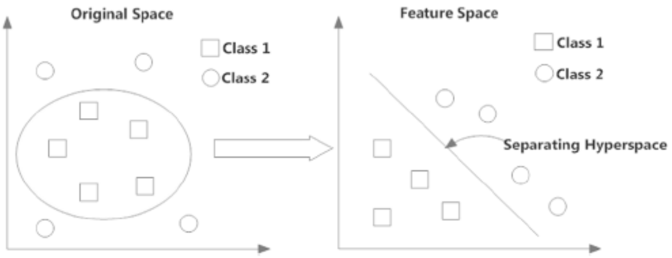

Figure 4.2: Mapping from the original dimensional space to higher dimensional space.

4.2 Whale Optimization Algorithm

Whale Optimization Algorithm : Inspired by the social behaviour of humpback whales, the WOA is an efficient metaheuristic optimization algorithm. It features exploration through circular motion and exploitation through helical motion to find optimal solutions in complex search spaces.

- **Circular Motion Equation :** The circular motion equation represents how each whale explores the search space:

$$\vec{X}_{\text{new}}[\text{ iter }] = \vec{X}_{\text{c\_best}}[\text{ iter }] - \vec{A} \cdot \vec{D}[\text{ iter }] \tag{4.5}$$

$$\vec{D}[\text{ iter }] = |\ \vec{C} \cdot \vec{X}_{\text{c\_best}}[\text{ iter }] - \vec{X}[\text{ iter }]\ | \tag{4.6}$$

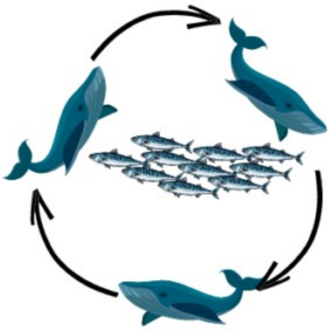

Figure 4.3: Circular motion of the whales around the prey. Intermmediate behavior between exploration and exploitation.

- **Helical Motion Equation :** The helical motion equation represents the exploitation process where whales converge towards the best solution:

$$\vec{X}_{\text{new}}[\text{ iter }] = \vec{D'}[\text{ iter }] \cdot e^{(b \cdot l)} \cdot \cos(2\pi l) + \vec{X}_{\text{c\_best}}[\text{ iter }] \tag{4.7}$$

$$\vec{D}'\,[\text{ iter }] = |\,\vec{X}_{\text{c\_best}}\,[\text{ iter }] - \vec{X}\,[\text{ iter }]\,| \tag{4.8}$$

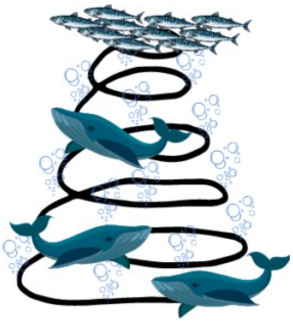

Figure 4.4: Helical Motion of the Whale Swarms

- **Exploration motion Equation :** The helical motion equation represents the exploitation process where whales converge towards the best solution:

$$\vec{X}\text{ new }\,[\text{ iter }] = \vec{X}_{\text{g\_rand}}\,[\text{ iter }] - \vec{A}\cdot\vec{D}\,[\text{ iter }] \tag{4.9}$$

$$\vec{D}\,[\text{ iter }] = |\,\vec{C}\cdot\vec{X}_{\text{g\_rand}}\,[\text{ iter }] - \vec{X}\,[\text{ iter }]\,| \tag{4.10}$$

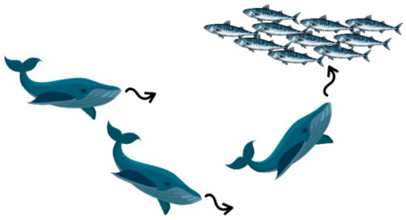

Figure 4.5: Exploratory movement of the whales taking any other random whale as a reference point

$$\text{While}\qquad \vec{A} = 2\cdot\vec{a}\cdot\vec{r} - \vec{a}\quad\text{and}\quad \vec{C} = 2\,\cdot\vec{r} \tag{4.11}$$

Where the value of ‘a’ decreases linearly from 0 to 2 and the value of ‘r’ is randomly generated from the interval [0, 1]

$$\text{Finally we update as follows } \vec{X}\,[\text{ iter} + 1\text{ }] = \vec{X}_{\text{new}}\,[\text{ iter }]. \tag{4.12}$$

4.3 Hybrid Support Vector Machine Model

In medical diagnosis, classification accuracy is of prime importance. The hybridization of cutting-edge optimization techniques and robust machine learning methodologies holds the promise of refining disease identification, prognosis, and treatment planning. This project introduces a hybrid algorithm that integrates the Whale Optimization Algorithm (WOA) with Support Vector Machines (SVMs) to address the challenges of medical diagnosis.

The WOA, inspired by the collaborative hunting strategies of humpback whales, brings a nature-inspired optimization approach into machine learning. In contrast, SVMs, used for their adeptness in classification tasks, focus on establishing optimal hyperplanes. However, both methods possess inherent limitations when used in isolation, especially in the context of medical diagnosis.

This hybrid algorithm aims to bridge these gaps by combining the exploratory nature of the WOA with the precision-driven classification capabilities of SVMs. By doing so, it aspires to elevate disease identification accuracy, enhance prognosis precision, and ultimately aid medical practitioners in making well-informed decisions.

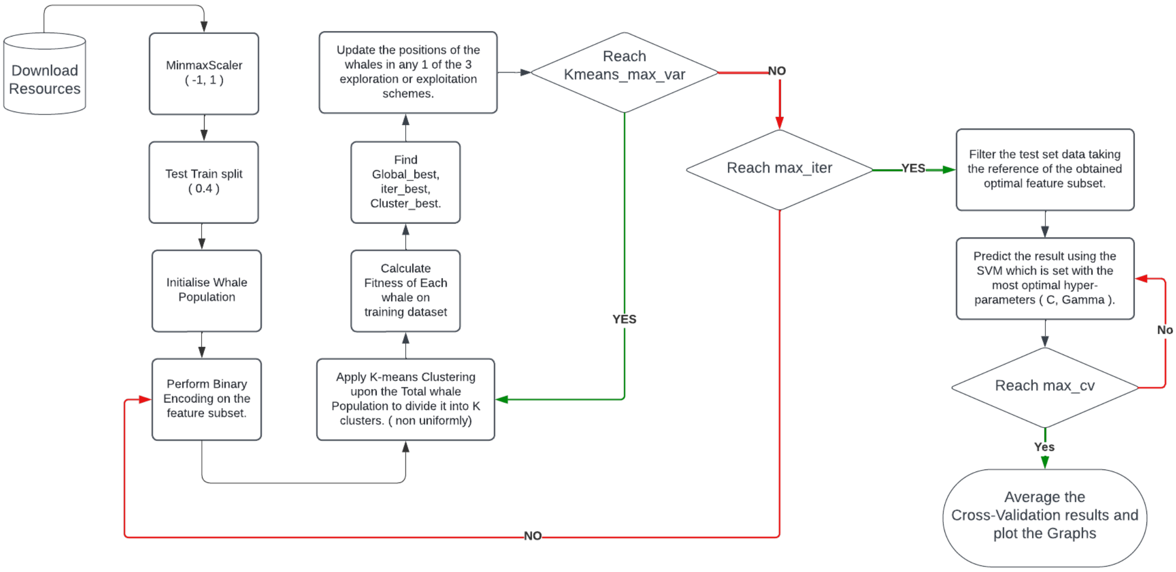
Download Resources
MinmaxScaler (-1, 1)
Test Train split (0.4)
Initialise Whale Population
Perform Binary Encoding on the feature subset.
Apply K-means Clustering upon the Total whale Population to divide it into K clusters. (non uniformly)
Calculate Fitness of Each whale on training dataset
Find Global_best, iter_best, Cluster_best.
Update the positions of the whales in any 1 of the 3 exploration or exploitation schemes.
Reach Kmeans_max_var
NO
YES
Reach max_iter
YES
NO
Filter the test set data taking the reference of the obtained optimal feature subset.
Predict the result using the SVM which is set with the most optimal hyper-parameters (C, Gamma).
Reach max_cv
No
Yes
Average the Cross-Validation results and plot the Graphs

Figure 4.6: Flowchart of the proposed Hybrid Methodology

The above proposed methodology facilitates the optimization of the hyper-parameters which regulate various factors like the influence of a single training sample (Gamma) or the trade-off between error minimization and margin maximization (C or Regularization parameter). It also accommodates a feature subset selection functionality which is implemented using Binary Encoding with a sigmoid function threshold of 0.4 (Experimentally determined).

The following algorithm explains the sequence in which the model objects and parameters are initialized, modified and used for training/testing in this hybrid model. It also describes the functions used for recording optimization data and the obtained optimal values.

Description of the proposed methodology

- Import necessary resources.
- Extract feature and target data (X and Y) and perform necessary preprocessing: convert the target variable into binary form, perform MinMaxScaling.
- Split the data into training and testing data.
- Define `svm_score` function responsible for calculating the fitness of the solutions traversed by the whale swarms by training and predicting on training data.
- Define `K_means_clustering` module which clusters the whole `WhalePopulation` list into K clusters (not necessarily uniform) where K is passed as a parameter. Further into the algorithm, the whale positions are updated taking the cluster best into consideration.
- Define an Iterative print function for printing out the positions and the respective fitness of the whales for every iteration.
- Define a 2D graph function which plots the positions of the complete whale population at a particular iteration.
- Define a 3D graph function which plots the complete path of a single whale, changing positions, clusters throughout the complete process, i.e., from the first to the last iteration.
- Prepare the dataset for k-fold cross-validation to find out k different accuracies and average them to get the net accuracy.
- Start the whale optimization process by initializing the positions data of the whales randomly between the limits which were passed down as parameters. Different limits are used for features and hyperparameters.

- The set number of whales perform any one of the three movements, iteratively moving towards their respective cluster best, while simultaneously changing the cluster grouping.

- The positions obtained are clipped according to the limits set, and the feature subset is obtained using binary Encoding of the features.

- Define global_best, iter_best, and cluster best variable containers. cluster_best is used for deciding the movements of the whales. iter_best is updated with the best_score obtained for an iteration, which is further used to update the global_best.

- After the best hyperparameters and feature subset are obtained, they are now used to train the final model, which is compared to the accuracy score of the standard SVM.

Pseudocode of the proposed methodology

```
• Initialize whalePopulation along with initial fitness.
• Determine Global_best ( position and fitness ).
• while iter $<$ max_iter :
•     for i in n_whales :
•         Encoding_function
•     end for
•     if kmeans_var == kmeans_var_max
•         perform K means clustering, set kmeans_var = 0.
•     end if
•     Initialize and find iter_best ( position and fitness ).
•      if global_best <iter_best :
•          global_best ← iter_best
•      end if
•     Initialize a1 and a2 (linearly decreasing from 2 to 0.
•     for i in n_whales :
•          Initialize and find cluster_best (position and fitness).
•         Initialize A, C, D, D', b, l, p.
•         if : p $<$ 0.5 :
•             if $|A| < 1$ :
```

```
•                 update using Equation (4.2)
•             end if
•             else :
•                 update using Equation (4.2)
•             end else
•         end if
•         else :
•             update using Equation (4.2)
•         end else
•     end for
•     for i in n_whales :
•         perform clipping with the respective limits.
•     end for
•     increment iter and k_means_var.
• end while
• call respective plotting functions for results.
• return global_best.
```

Pseudocode of the fitness function used in the optimization process

```
• load datasets and split into training data x_train, y_train, x_test, y_test.
• svm_score ( parameters : position )
•         assign the position array values to local variables C and gamma.
•         selected_indices = [ if i == 1 in position [ 2 to len(position)] → include ].
•         clip selected_indices in x_train.
•         svm_model = SVC ( with given C and Gamma ).
•         svm_model.fit( selected_training_data ).
•         y_predicted = svm_model.predict ( selected_training_data ).
• return accuracy score ( y_predicted, y_train ).
```

5

Experiment and Results

The contents of this chapter encompass the description of the experiment set-up that was employed in the project, as well as a detailed account of the outcomes and results that were obtained.

5.1 Experiment setup

This project can be used to conduct various experiments by tweaking many parameters according to the users experimental setup and requirement.

Hyper-parameters	Value
num_whales	15
max_iter	100
k	7
kmeans_max_var	10
(minx, maxx)	(-1.0, 1.0)
(minh, maxh)	(0.01, 1.0)
binary_encoding_threshold	0.4
p (explore/exploit probability)	0.5
b (helix constant 1)	1
l (helix constant 2)	[-1, 1]
a (movement constant 1)	[0, 2]
r (movement constant 2)	[0, 1]

Table 5.1: Hyper-parameters of the Experimental setup

5.2 Results and discussion

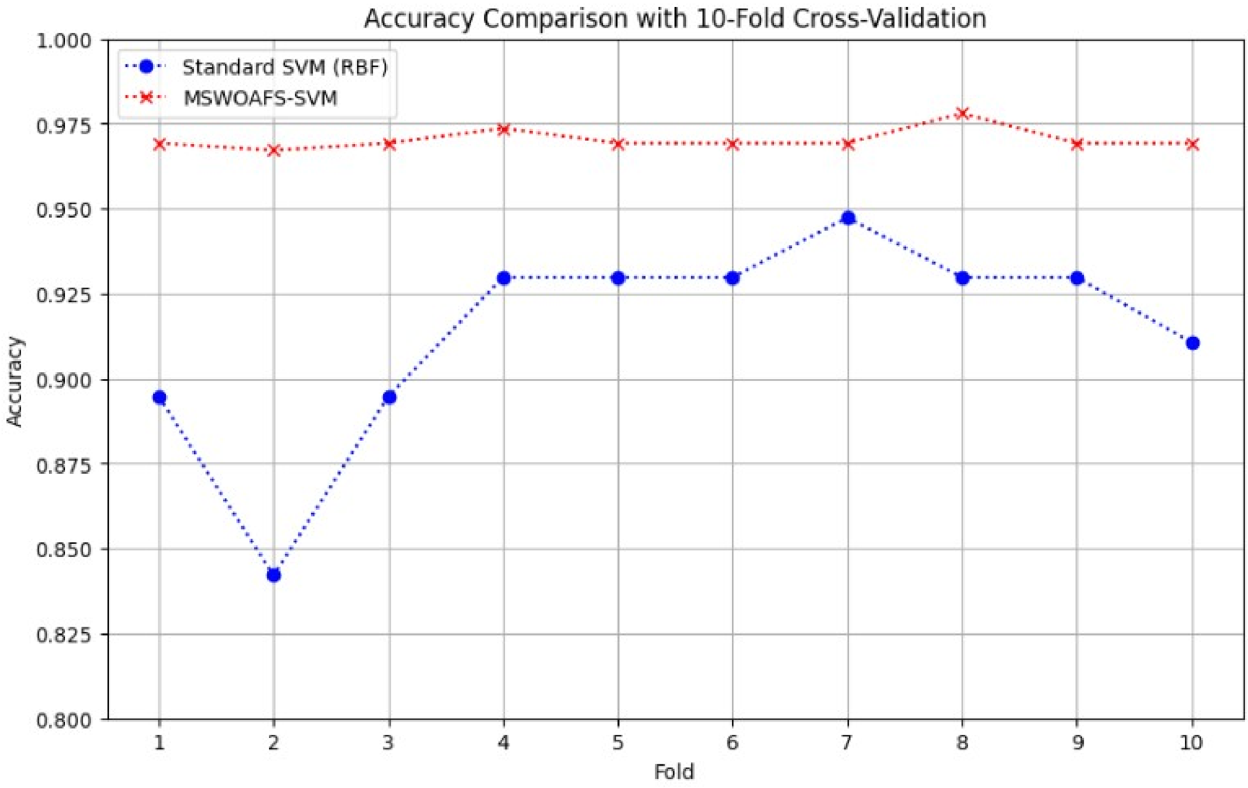

Figure 5.1: Comparision of the performance of MSWOAFS-SVM with standard RBF kernel SVM using 10 fold cv accuracies.

The prediction accuracies of both the models are compared for every iteration of the 10 fold cross validation in above diagram. A significant increase in accuracy is observed in the Hybrid model implemented due to optimal hyperparameter selection and optimal feature subset filtration. Despite the better accuracy obtained, the results show only a slight improvement due to the use of K-fold cross validation technique.

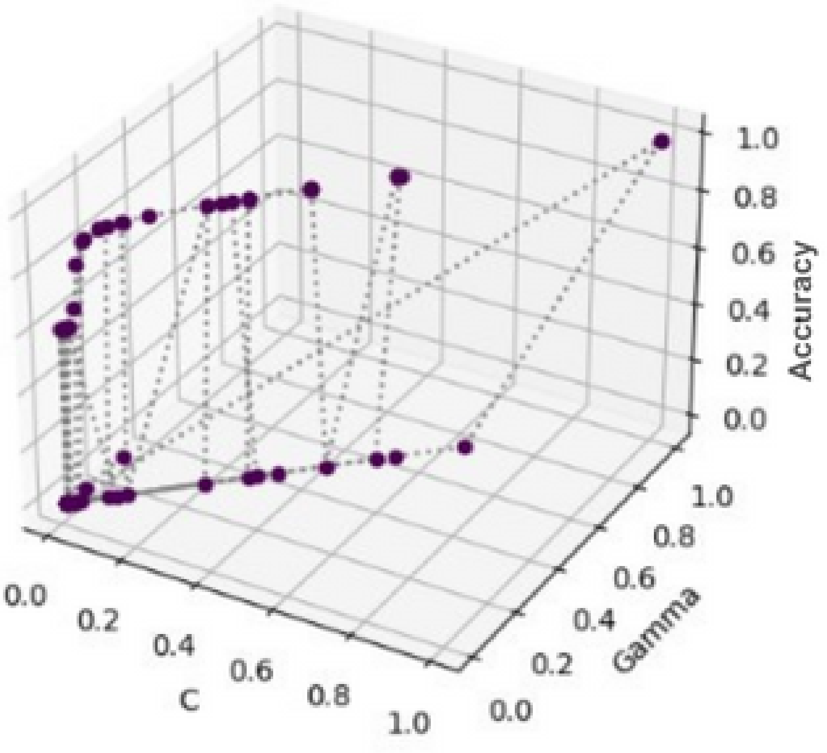

Figure 5.2: The movement a single whale traces throughout the iterations.

The optimal hyper parameters responsible for higher prediction accuracies are obtained using WOA with the help of the training dataset alone to prevent overfitting. The 3-Dimensional graph shows the path traced by a single whale of the total whale population.

The whale optimization does not guarantee the preservation of the best solution throughout the iterations therefore sometimes leading to inferior solutions in the output despite traversing through better solutions in one of the intermediary stages. Therefore it is necessary to store the best solution obtained in every iteration to further compare (i.e if better solutions are found, they replaced).

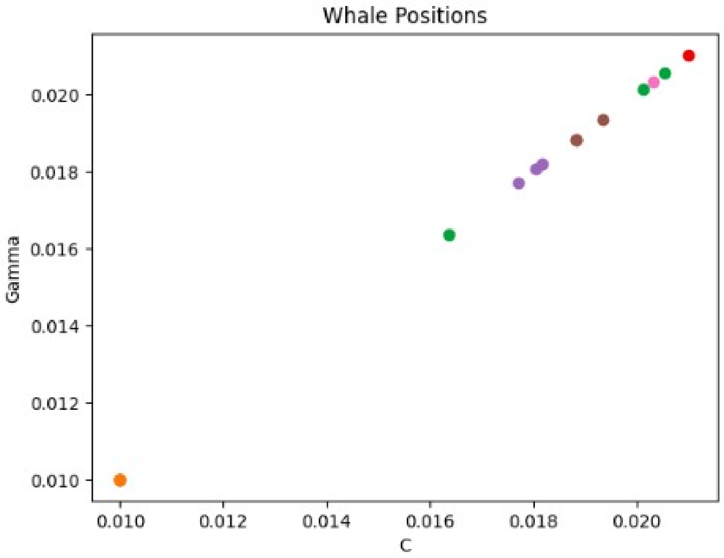

Figure 5.3: The Big picture of the whalePopulation in a random iteration.

The direction of every individual whale is influenced by the cluster best of the cluster to which that particular whale belongs to. The best solution for testing the SVM performance is obtained by calculating the most optimal solution among the complete population (not limited to any single cluster), this is termed as the iteration best. The global best or the final output of the search is picked out from the complete set of iteration best values. The above shown figure is an intermediary stage of the total whale Population.

We can also observe that the values of both the hyper parameters are almost same, this can be attributed to the nature of the meta-heuristic algorithm used. WOA updates all the parameters parallely and in a similar manner because it considers the complete solution set as a multi dimensional vector.

6

Conclusions and Future Scope

This chapter of the report is dedicated to the conclusion and future scope. This section presents a thorough summary of the principal discoveries and understandings attained through the research conducted in the preceding sections.

6.1 Conclusions

In conclusion, this project has successfully devised and implemented an innovative hybrid classification framework by integrating Support Vector Machines (SVMs) with the optimization prowess of the Whale Optimization Algorithm (WOA). Extensive experimentation and analysis have underscored the Multi-Swarm Whale Optimizer Boosted Support Vector Machine's superior classification performance, particularly in medical diagnosis tasks. By effectively tackling challenges related to feature selection, hyperparameter optimization, and convergence rate, this hybrid model presents a robust solution for accurate medical diagnoses.

The project's comprehensive evaluation utilizing real-world medical datasets, such as the Wisconsin breast cancer dataset, has not only validated the proposed approach but also established its significance in enhancing healthcare decision-making processes.

6.2 Future Scope

While the current endeavor has introduced a promising hybrid model for medical diagnosis, several avenues for future research and enhancements are worth exploring:

(i) **Algorithmic Refinement:** Delve deeper into algorithmic optimization and parameter fine-tuning. Exploring adjustments to WOA parameters, including iteration count, population size, and movement equations, could yield accelerated convergence and heightened accuracy.

(ii) **Hyperparameter Auto-Tuning:** Implement an automated mechanism for hyperparameter tuning, such as Bayesian optimization or grid search. This would eliminate manual parameter adjustments and enhance the model's generalizability.

(iii) **Diverse Medical Datasets:** Extend the model's evaluation to a broader spectrum of medical datasets to validate its robustness across varied medical domains.

Employing datasets with diverse characteristics and complexities would offer a more comprehensive comprehension of its performance.

(iv) **Ensemble Approaches:** Investigate the potential of amalgamating the hybrid model with other ensemble learning techniques, such as bagging or boosting. Ensembles can harness model diversity to amplify overall predictive performance.

(v) **Interpretable Models:** Devise techniques to decipher the hybrid model's decisions, augmenting its utility in medical diagnosis. Granting clinicians insights into feature significance and decision rationales can bolster their confidence in and acceptance of the model.

(vi) **Real-Time Diagnosis:** Adapt the hybrid model for real-time medical diagnosis scenarios. Swift processing and classification of new data in real-time can significantly impact patient care and treatment planning.

(vii) **Domain-Specific Feature Engineering:** Implement domain-specific feature engineering strategies to further enrich medical data representation. Extracting pertinent features from raw data can lead to enhanced classification accuracy.

In conclusion, the hybrid classification framework formulated herein lays a solid groundwork for future exploration and pragmatic applications within the medical realm. As technology advances and data availability expands, the contributions of this project can be harnessed to craft precise and effective diagnostic tools, benefiting both medical practitioners and patients alike.

Bibliography

[1] Q. Zhang, H. Zhao, Y. Hang, and X. Lu, “Research on parkinson’s disease diagnosis based on improved particle swarm optimization support vector machine algorithm,” in *2022 International Conference on Machine Learning, Cloud Computing and Intelligent Mining (MLCCIM)*, 2022, pp. 365–369.

[2] D. Moldovan, “Cervical cancer diagnosis using a chicken swarm optimization based machine learning method,” in *2020 International Conference on e-Health and Bioengineering (EHB)*, 2020, pp. 1–4.

[3] G. Deng, M. Tang, Y. Xi, and M. Zhang, “Privacy-preserving online medical prediagnosis training model based on soft-margin svm,” *IEEE Transactions on Services Computing*, vol. 16, no. 3, pp. 2072–2084, 2023.

[4] S. Mirjalili and A. Lewis, “The whale optimization algorithm,” *Advances in Engineering Software Volume 95, May 2016, Pages 51-67*, 2016.

[5] X.-S. Yang, “Particle swarm optimization,” *Advances in Engineering Software*, 2014.

[6] A. R. Mehrabian and C. Lucas, “A novel numerical optimization algorithm inspired by group hunting of animals: Hunting search,” in *Journal of Computational Physics*, 2006.