



# Tomas Bata University in Zlín

## Faculty of Applied Informatics

Doctoral Thesis

### **Soft Computing Techniques for Sentiment Analysis and Feature Selection**

**Soft computingové techniky pro analýzu sentimentu a výběr  
příznaků**

Author: **Ing. Raphael Kwaku Botchway**

Degree programme: Engineering Informatics

Degree course: Engineering Informatics

Supervisor: assoc. prof. Ing. Zuzana Komínková Oplatková, Ph.D.

Zlín, 2023



## **DEDICATION**

I dedicate this dissertation to my dear wife Obaapa Mabel for her wonderful support, encouragement, and patience demonstrated throughout my Ph.D. journey. It is also dedicated to my kids Kwaku Jerry, David Obrempong, and Sally Adepa.

It is further dedicated to Freda Amoah, Benjamin Caleb, Akosua Dorcas as well as my brother-in-law Emmanuel Abrokwah for the various roles played in helping my family during my absence. I am extremely grateful to you all and may the good Lord bless you greatly.

## **ACKNOWLEDGEMENTS**

I could not have completed my doctoral studies without the tremendous support received from some persons. My supervisor, assoc. prof. Ing. Zuzana Komínková Oplatková Ph.D., whose assistance and guidance I most appreciate, deserves special recognition. She helped make this dissertation a reality by reading through my countless revisions, for which I am very grateful. I appreciate her enlightening remarks and criticisms of my work.

Also worthy of note are the many good friends and colleagues I met in Zlín, particularly Ing. Abdul Bashiru Jibril Ph.D., and MSc. Vinod Yadav, who in various ways made significant contributions to my studies.

Finally, I would like to acknowledge Prof. Andries Engelbrecht, the Voight Chair in Data Science at the department of Industrial Engineering, Stellenbosch University, prof. Mgr. Roman Jašek Ph.D., head of the department of Informatics and Artificial Intelligence, and the members of the A.I lab headed by prof. Ing. Roman Šenkeřík Ph.D.

## ABSTRAKT

Pochopení významu sociálních médií v poslední době přitahuje akademickou pozornost. Jak kdysi řekl významný učenec, sociální média již nejsou pomíjivým pocitem nebo módou. Názory zákazníků vyjádřené na sociálních sítích mohou předávat důležité zprávy, které mohou podniky využít k budování pevných vztahů se zákazníky. S rostoucím využíváním sociálních médií mezi běžnou populací roste i jejich využití v obchodním světě, protože stále více firem využívá sociální média jako efektivní způsob, jak se spojit s mnoha klienty.

Navzdory rychlému přechodu od tradičních k sociálním médiím se firmy v této éře takzvaných velkých dat stále snaží plně porozumět potřebám a obavám svých zákazníků. Navíc schopnost rychle porozumět spotřebitelské komunikaci, aby management mohl reagovat včas a efektivně, zůstává klíčovou výzvou. Dále, velké množství nestrukturovaných dat a nedostatek praktických nástrojů pro analýzu nestrukturovaných dat tuto analýzu komplikuje.

Tato disertační práce představuje stručný přehled aplikací soft computing technik pro analýzu sentimentu a výběr příznaků. Zpočátku autor disertační práce využívá množství dat ze sociálních médií dostupných online k ovlivňování tím, že využívá techniky dolování textu k analýze obsahu generovaného uživateli z příspěvků na sociálních sítích (tweetů) na podporu spotřebitelského rozhodování a marketingové komunikace. Tento nestrukturovaný obsah vytvářený uživateli silně obsahuje slangy, slova s nesprávným pravopisem atd., což představuje výzvu pro výběr funkcí kvůli vágnosti, nepřesnosti a nejednoznačnosti, které jsou v něm obsaženy. V důsledku toho je implementováno řešení založené na metaheuristickém algoritmu Particle Swarm Optimization (PSO) pro optimální výběr textových prvků během analýzy sentimentu, aby se zvýšila přesnost predikce sentimentu.

Druhá část disertační práce kombinuje techniky evolučních výpočtů s úhlovou modulací pro řešení problému výběru příznaků (feature selection). Při hodnocení výkonnosti navržené techniky je použito osmnáct klasických datových sad strojového učení UCI. Zjištění potvrzují konkurenceschopnost a vynikající výkonnost navrženého přístupu při porovnání s jinými metaheuristickými metodami souvisejícími s prací, které jsou k dispozici v literatuře s tématem výběru příznaků. Další statistické testy rovněž potvrzují, že navrhovaná metoda je účinným nástrojem pro řešení binárních optimalizačních problémů v různých oblastech.

**Klíčová slova:** úhlová modulace, evoluční výpočetní techniky, výběr příznaků / atributů, optimalizace rojem částic, analýza sentimentu, sociální média

## **ABSTRACT**

Understanding the significance of social media has attracted academic attention in recent times. As a prominent scholar once put it, social media is no longer a passing sensation or fad. Customer opinions expressed on social media can convey important messages that businesses can use to build strong relationships with customers. As social media usage among the general population grows, so are its uses in the business world as more businesses turn to social media as a cost-effective and efficient way to connect with many clients.

Despite the quick transition from traditional to social media, firms still struggle to fully comprehend the needs and concerns of their customers in this era of the so-called big data. Moreover, the ability to quickly comprehend consumer communications so that management can respond in a timely and effective manner remains a key challenge. Further, the huge amount of unstructured data and a scarcity of practical tools for analysing this unstructured data makes such analysis more complicated.

This dissertation presents a brief overview of the application of soft computing techniques for sentiment analysis and feature selection. Initially, the author of the dissertation utilizes the abundance of social media data available online as leverage by employing text mining techniques to analyze user-generated content from social media posts (tweets) to support consumer decision-making and marketing communications. This unstructured user-generated content heavily includes slang, misspelt words, etc... thereby presenting a challenge to feature selection due to the vagueness, imprecision, and ambiguity contained therein. Consequently, a metaheuristic-based solution using the Particle Swarm Optimization (PSO) algorithm for optimal text feature selection during sentiment analysis is implemented to enhance sentiment prediction accuracy.

The second segment of the dissertation combines evolutionary computation techniques with angle modulation to solve feature selection problems. Eighteen classical UCI machine learning datasets are employed in evaluating the performance of the proposed technique. The findings confirm the competitive and superior performance of the proposed approach when juxtaposed with other work-related metaheuristics methods available in feature selection literature. Further statistical tests also confirm the proposed method as a potent tool for resolving binary optimization problems across different domains.

**Keywords:** angle modulation, evolutionary computation, feature selection, particle swarm optimization, sentiment analysis, social media

© Raphael Kwaku Botchway

The publication was issued in the year 2023.

## TABLE OF CONTENTS

1.	INTRODUCTION .....	10
2.	THE STATE OF THE ART .....	14
2.1	SOCIAL MEDIA .....	14
2.2	NATURAL LANGUAGE PROCESSING .....	15
2.3	SENTIMENT ANALYSIS APPROACHES .....	16
2.3.1	LEXICON-BASED SENTIMENT ANALYSIS .....	16
2.3.2	MACHINE LEARNING TECHNIQUES .....	18
2.4	DEEP LEARNING MODELS .....	19
2.5	METAHEURISTIC ALGORITHMS .....	20
2.5.1	PARTICLE SWARM OPTIMIZATION (PSO) .....	23
2.5.2	GREY WOLF OPTIMIZATION .....	24
2.5.3	WHALE OPTIMIZATION ALGORITHM .....	26
2.5.4	ANGLE MODULATED PARTICLE SWARM OPTIMIZATION (AMPSO) .....	28
2.6	FEATURE SELECTION.....	30
2.7	HYBRID METAHEURISTIC ALGORITHMS FOR FEATURE SELECTION .....	31
3.	OBJECTIVES OF THE THESIS .....	33
4.	WORKFLOW .....	34
4.1	SENTIMENT ANALYSIS WORKFLOW .....	34
4.2	FEATURE SELECTION WORKFLOW .....	35
5.	HYBRID BIO-INSPIRED FEATURE SELECTION TECHNIQUE USING ANGLE MODULATION .....	37
5.1	CONTINUOUS HYBRID PSOGWO .....	37
5.2	PROPOSED HYBRID ANGLE MODULATED GWOPSO (AMGWOPSO).....	38
5.3	MODELING THE HYBRID AMGWOPSO .....	38
5.4	EXPERIMENTAL SETUP FOR PROPOSED AMGWOPSO .....	40
5.5	DATASETS .....	41
6.	RESULTS .....	42
6.1	DEDUCTIONS FROM A SUB-SAHARAN AFRICAN BANK: A SENTIMENT ANALYSIS APPROACH.....	42
6.2	TEXT-BASED FEATURE SELECTION USING BINARY PSO FOR SENTIMENT ANALYSIS .....	45



6.2.1	EXPERIMENTAL SETUP .....	47
6.2.2	RESULTS REALIZED .....	47
6.3	HYBRID BIO-INSPIRED FEATURE SELECTION USING ANGLE MODULATION.....	48
6.3.1	PARAMETERS UTILIZED.....	48
6.3.2	METRICS FOR EVALUATION .....	49
6.3.3	EXPERIMENTAL RESULTS AND DISCUSSION .....	50
6.3.4	PROPOSED AMGWOPSO .....	51
6.3.5	COMPARING THE PROPOSED HYBRID AMGWOPSO WITH OTHER METHODS 53	
6.3.6	STATISTICAL SIGNIFICANCE ANALYSES .....	54
6.3.7	COMPUTATIONAL COMPLEXITY .....	55
6.3.8	ACHIEVED RESULTS FOR AMGWOPSO .....	56
7.	SUMMARY OF RESULTS AND DISCUSSION .....	75
8.	CONTRIBUTION OF THESIS TO SCIENCE AND PRACTICE.....	78
9.	CONCLUSION.....	80
	BIBLIOGRAPHY .....	81
	LIST OF FIGURES .....	89
	LIST OF TABLES.....	92
	LIST OF SYMBOLS, ACRONYMS, AND ABBREVIATIONS.....	93
	LIST OF PUBLICATIONS BY THE AUTHOR .....	95
	AUTHOR’S PROFESSIONAL CURRICULUM VITAE .....	97

# 1. INTRODUCTION

The functioning of our society has been significantly impacted by the internet [1], [2]. As the literature suggests, the internet and the rapid growth of its associated technologies (Web 2.0) in terms of volume, velocity, and variety of opinion-rich information available online has convinced a lot of researchers to pay much attention to social media (SM). To put it another way, social media is no longer unimportant and discretionary. It includes a huge selection of online spaces that encourage user engagement, teamwork, and content sharing. Users have the ability to impact other users individually and collectively through social media by sharing their opinions with them. Thus, social media acts as a natural laboratory to analyse new generation netizens' attitudes and access large-scale discussions in real-time [76].

Emotions play a significant role in influencing attitudes and comprehending behavioural motives. The cognitive consistency theory, created by Festinger [77], contends that people are driven to take actions that are in line with their perceptions and beliefs [78]. Furthermore, if their experiences do not match their perceptions, people are also motivated to alter their behaviour. As such, customers share their opinions about their experiences with products through ratings, reviews, and recommendations on virtual platforms like Twitter.

The phrase “Sentiment Analysis” was first coined by Dave et al., in the year 2003 [75]. Sentiment analysis (SA), originally established as a natural language processing (NLP) text classification task, facilitates knowledge extraction from unstructured data abundantly available from social media sources for efficient decision-making. Research conducted by [3], [4] investigated customer sentiment expressed as either positive or negative emotional words in user-generated content (UGC) such as tweets.

UGC simply refers to user-generated content and denotes any type of online content that is produced, initiated, shared, and consumed by users [79]. Since emotional words are ingrained in the descriptions of personal experiences by nature, examining UGC is an additional way to assess customers' sentiments. Thus, to better understand user sentiments, UGC offers extensive data. According to [80], opinions expressed in UGC have a big impact on review readers' propensity to buy. The UGC-like tweets have become vital for multinational brands because it impacts their branding strategy in numerous ways.

Multinational brands like Louis Vuitton, Visa, Amazon, and Tesco need customer feedback and engagement through social media platforms e.g., Instagram, LinkedIn, Twitter, etc... to stay up with the changing consumer behaviour made possible by technological advances. The importance of sentiment polarity, be it positive, negative, or neutral, has long been acknowledged. Sentiment polarity in UGC has also been observed to be strongly correlated with sales volume and increased revenues [81].

Earlier studies [13] on online shopping revealed that the sentiment polarity of tweets from various businesses (e.g., John Lewis, and Tesco) can reflect the differences in their marketing strategies. Extant literature reports the application of the use of some soft computing techniques namely, machine learning (ML), lexicon-based, and rule-based approaches [5], [6] in numerous disciplines.

Soft Computing (SC) techniques are generally organized into five groups (Fig. 1.1) namely Machine Learning (ML), Evolutionary Computation (EC), Fuzzy Logic (FL), Probabilistic Reasoning (PR), and Neural Networks (NN) [7]. All SC approaches have one thing in common: they are capable of self-tuning, which means that they may learn from experimental data and approximate it to gain generalization power [17]. The use of SC techniques for sentiment analysis (SA) helps to transform unstructured social media data from sources such as Twitter into a structured data format for business intelligence purposes [8].

Twitter is the most popular and preferred microblogging network that allows users to publish their ideas voluntarily, thereby enervating the impact of biases. This is in contrast to some traditional research approaches, such as empirical surveys, focus groups, qualitative case studies, interviews, and opinion polls [82].

Grover et al., [83] claim that the majority of current research procedures are vulnerable to the ingrained response biases of optimism, social desirability, and skewed impressions. Twitter has emerged as a major revolution in the social media space among Web 2.0 tools. Hence, for the exploratory experiments in this dissertation, Twitter data was chosen. Table 1.1 categorizes the advantages of Web 2.0 into three different groups in line with the research of Bughin et al., [84] and the McKinsey quarterly poll.

Despite the quick transition from traditional to social media, firms still struggle to fully comprehend the needs and concerns of their customers in this era of the so-called big data. Moreso, the ability to quickly comprehend consumer communications so that management can respond in a timely and effective manner remains a key challenge facing businesses. Again, the huge amount of unstructured data and the scarcity of practical tools for analysing this (unstructured) data makes such analysis more complicated.

Consequently, the first segment of this dissertation utilizes the abundance of social media data available online as leverage to explore the use of soft computing techniques for sentiment analysis. During this process, text mining techniques are employed to analyze UGC from social media posts (tweets) to support consumer decision-making and marketing communications.

The second part of the dissertation builds on the earlier segment by extending the use of evolutionary computation techniques to solve feature selection problems. In this phase, a metaheuristic-based solution using the Particle Swarm Optimization (PSO) algorithm for optimal subset text feature selection during sentiment analysis is implemented.

Furthermore, a new hybrid metaheuristic algorithm is developed by combining a variant of the PSO with the Grey Wolf Optimization (GWO) algorithm [28],

[40] for optimal subset feature selection. The variant of the PSO referred to earlier is the Angle Modulated Particle Swarm Optimizer (AMPSO). The AMPSO utilizes a trigonometric function derived from a technique used in the signal processing field in the telecommunication industry [27], [30]. The resulting hybrid version is the Angle Modulated GWOPSO (AMGWOPSO). A low-level coevolutionary mixed hybrid approach is used in this study when combining AMPSO and GWO.

The proposed solution is evaluated and analyzed on different publicly available datasets to understand the benefits and drawbacks of the approach to formulate recommendations for future work.

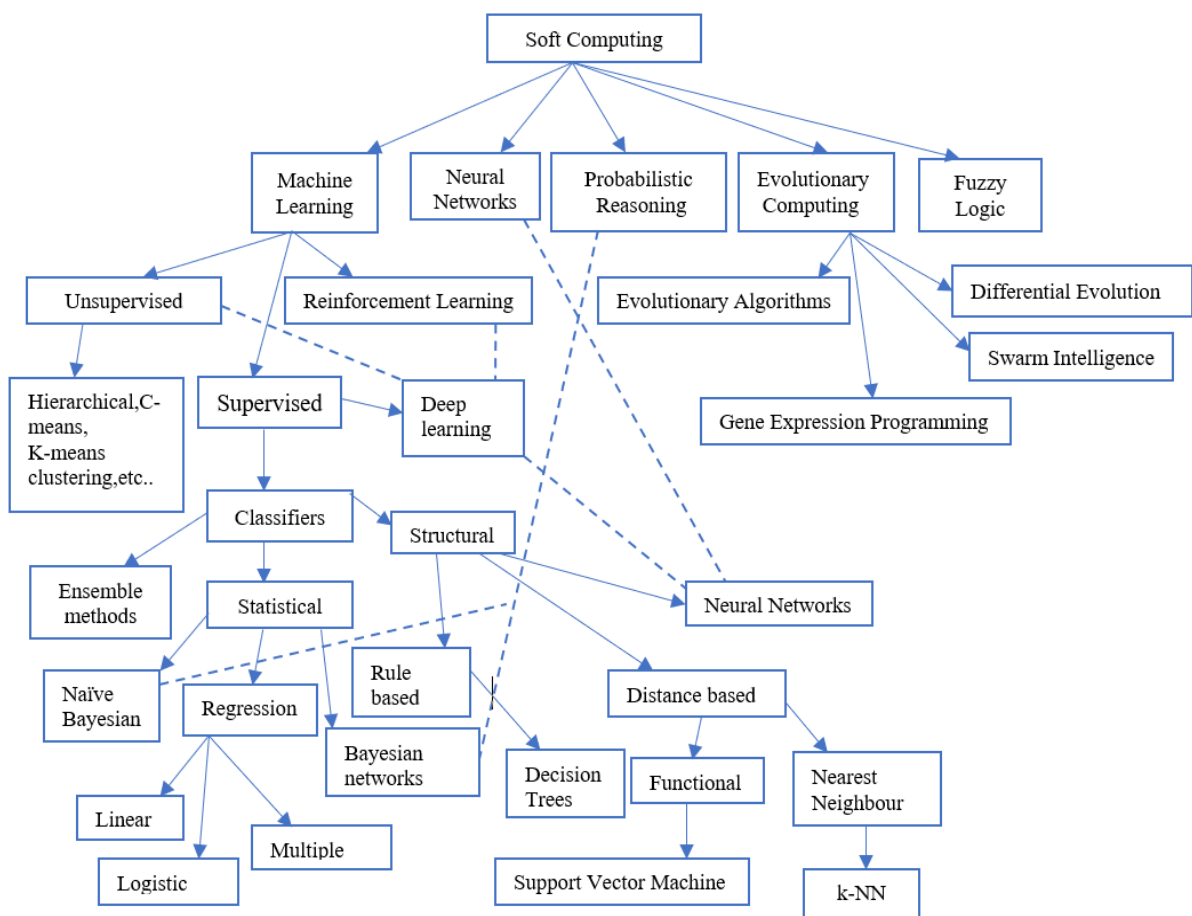


Fig. 1.1: Overview of Soft computing techniques [17]

**Table 1.1 Business benefits from Web 2.0 [84]**

Benefits (%)	Internal Purpose	Customer-related purpose	External partners/suppliers
Increasing effectiveness of marketing and speed of access to knowledge	68	52	51
Increasing customer/employee/supplier satisfaction	35	43	37
Increasing revenue	14	18	16
Reducing travel costs		32	40
Reducing marketing costs		38	
Reducing customer support costs		32	
Reducing communication costs	54		49
Reducing time to market for products/services	25	24	24
increasing the number of successful innovations for new products/services	25	22	19

## **2. THE STATE OF THE ART**

This section begins by initiating a discussion on sentiment analysis which is a subfield of natural language processing (NLP). Guided by the aims of the dissertation, special attention is paid to approaches that help to understand and gain valuable insights from the vast amount of unstructured social media data (Twitter) available, as well as the role of soft computing techniques in sentiment analysis and feature selection.

Furthermore, a brief background and state-of-the-art solutions covering core concepts related to social media, NLP, sentiment analysis, evolutionary computation techniques (metaheuristics algorithms), and feature selection are presented in the sub-sections below.

### **2.1 SOCIAL MEDIA**

As mentioned in the previous chapter, social media acts as an ideal platform to instantaneously access extensive conversations and understand the attitudes of new generation netizens [76]. According to current statistics [85], [86], the majority of people in the United States (79%), East Asia (70%), and Northern Europe (67%), all had social media profiles. Also, more than 80% of customers get useful information from social media-based channels, and about 77% of consumers read online evaluations of products written by other consumers and trust those reviews more than personal recommendations [87].

Further, businesses can use online customer reviews to determine the general level of interest, contentment, or needs for their products. Exceptionally expressive postings with copious details might offer vital pointers on how to build new products, improve functions, or maintain quality [88]. From the foregoing, new age netizens are not simply passive buyers of goods, but active users who, by their strong social media voices, influence the product's survival and future development in either a direct or indirect way.

In reality, a lot of studies have gathered a lot of customer-generated social media data and developed statistical findings, quantifiable satisfaction levels, or themes that can assist commercial organizations in making strategic decisions in their business environment. Specifically, Jeong et al., [89] gathered online reviews of Samsung Galaxy Note 5 from Reddit, another popular virtual platform, and applied topic modeling to extract customers' concerns about battery life, touch recognition, and camera.

Additionally, they evaluated each topic's relevance and satisfaction rating, computed the opportunity score, and measured their satisfaction and importance levels. By giving product attributes with high relevance ratings but low satisfaction, the knowledge they gathered can aid the related firm's decision-making during the development of a good or service. Scholarly attention has been drawn toward understanding the importance of social media, which has become one of the best virtual places for customers to express their views. A scholarly

activity known as social media analytics has evolved in response to this increase in academic interest in social media data. This endeavour entails acquiring and analysing diverse social media data to uncover important hidden information.

By gathering information from Amazon reviews of three competing smartphones, Trappey et al., [90] were able to quantify user interest in each of the key qualities they identified. These findings helped the customer-centric product enhancement by revealing customers' intentions and the relative placement of each product in the competitive market. Besides, many studies have applied sentiment analysis to customer-generated web data to quantitatively analyse customer satisfaction.

In a society where it is challenging to recognize prominent people, analysing social media data has revealed a tremendous shift in how public opinion is influenced or shaped. The term "social media influencers" has been used to describe social media opinion leaders [91]. Intriguingly, discovering influential users has attracted considerable attention towards cross-domain sustainable applications by supporting society across multiple levels, encompassing e-governance, financial risk evaluation, viral marketing, etc... [92].

Twitter is regarded as the de facto communication medium amongst socio-technical internet channels facilitating human and technological interactions. Exploring Twitter influencers is now strategically important from both a global and regional perspective due to advancements in Twitter science and enriched network technologies and methods [92].

Influencers on social media usually have a large following and a prominent position in their networks. They often promote their posts aggressively. Because of this, influencer postings receive more likes, favourites, comments, and shares from the general audience. The reactions to their posts, especially shares, are indications of influence. According to studies, influencers are particularly effective in persuasion because of the concerted and relentless efforts they make to advance a particular cause, viewpoint, or product. They utilize the internet differently than the average user, who shares and posts things in a far less direct manner [93].

## **2.2 NATURAL LANGUAGE PROCESSING**

Natural Language Processing (NLP) is a sub-field of Artificial Intelligence (AI) defined as a collection of computational techniques for automatic analysis and representation of human languages, motivated by theory [9]. However, the automatic analysis of text considered to be at par with humans requires a higher-level understanding of natural language by machines which may not happen anytime soon.

NLP examples such as question-answering, online information retrieval, and aggregation are known to be mainly based on algorithms that rely solely on the textual representation on web pages. However, these algorithms perform better when used for tasks such as text retrieval, spell checks, and word level analysis

but perform poorly when used for analysis at the sentence and paragraph level. It is clear from the above explanation that these algorithms have limited capabilities when it comes to the issue of sentence interpretation and meaningful information extraction. Among NLP tasks and challenges include some well-known applications of NLP such as text summarization, topic segmentation, question-answering, sentiment analysis, text generation/dialogues, automatic language translation, etc. Natural languages usually contain a lot of ambiguities with several words having more than one meaning. Again, they tend to be large and contain infinitely many words. Consequently, developing a program that understands natural languages is quite challenging.

## 2.3 SENTIMENT ANALYSIS APPROACHES

Sentiment Analysis (SA), also known as opinion mining (OM), is essentially an NLP activity that entails the identification of user sentiment, attitude, emotion, and opinion in natural language text.

According to [10], SA can be classified as machine learning-based or lexicon-based as shown in Fig. 2.1. A detailed description of the lexicon-based approach is detailed in the next section.

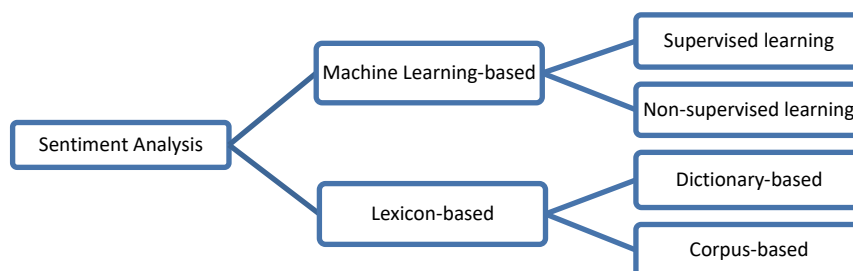


Fig. 2.1: Sentiment analysis approaches

### 2.3.1 LEXICON-BASED SENTIMENT ANALYSIS

In the unsupervised lexicon-based method, a sentiment lexicon is used to calculate the overall sentiment polarity of a text document based on the sum of the polarities of the individual words embedded in the text [94]. Lexical resources are typically employed in a variety of studies that use lexicon-based methods for unsupervised sentiment classification or analysis.

The main drawback of this approach is that some features are inaccessible to human analysis; however, supervised classification techniques can find these concealed features. On the other hand, the supervised classification-based technique entails building classifiers utilizing supervised machine learning that are fed explicitly trained data with labels for the classification problem. The main drawback to this approach is that it inherently has a hidden, black-box mechanism, is computationally expensive, and needs manually labelled training data to get a reasonably good accuracy. Sentimental words and phrases make a significant



contribution and are an essential resource for SA tasks. They are usually assembled in a sentiment lexicon, which is a linguistic resource that contains word-level a priori knowledge throughout the semantic dimension of sentiment. A standard sentiment lexicon includes attributes for each word, such as the polarity and intensity of the polarity.

The challenge is that it would require an extraordinary amount of annotator time and effort to manually tag words to build a sentiment lexicon. As a result, a significant amount of work focusing on automated sentiment lexicon development has emerged in this field. Using lexical resources and internet-based dictionaries (WordNet, Merriam Webster, etc...), the dictionary-based method automatically associates phrases with the polarity of their corresponding sentiments [45], [96].

The corpus-based technique, on the other hand, makes use of co-occurrence data or syntactic patterns in a corpus. [94], [97]. When a domain-independent sentiment lexicon needs to be transformed into a domain-specific vocabulary, text corpora are usually preferred [98].

In these modern times, sentiment analysis systems must be able to parse free-form social media text. With this, sentiment lexicon generation methods have drastically transformed from using dictionaries to using social media corpora. Bandhakavi et al. [99] modelled an emotion corpus for social media sentiment analysis using a unigram mixture model, combined with an emotion sentiment mapping for the generation of word sentiment lexicons that capture emotion-sensitive vocabulary. They evaluate the proposed mixture model in learning emotion-sensitive sentiment lexicons with those generated using supervised Latent Dirichlet Allocation (sLDA) as well as word document frequency (WDF) frequency information.

The WordNet [11] lexicon database was used by Kamp et al. [12] in their work to find the emotional content of a word along different dimensions. Using WordNet, they built a distance metric for the semantic orientation of adjectives. In another study [51], Ibrahim and Wang used SentiStrength to decode the sentiment dynamics of online retail customers [13].

Hutto and Gilbert introduced VADER[15], a simple sentiment analysis model for social media sentiment analysis. VADER is an acronym for Valence Aware Dictionary sEntiment Reasoner [15]. As a pre-trained sentiment analyzer available via the Natural language Toolkit (NLTK) library, it examines the lexical features of a document to compute a preliminary sentiment score and then applies five different rules based on general syntactic and grammatical conventions to modify the score.

The values returned by VADER [15] are *pos*, *neu*, *neg*, and *compound* (Comp). These VADER values represent the probability of a sentiment being positive, neutral, negative, and the normalized compound score respectively. The compound score computes the sum of all the lexicon ratings lies within the range

-1 (strongly negative) and 1 (strongly positive). Ho and Huang in their work [16] formulated equation (1) using the VADER to determine sentiments.

$$Sentiment = \begin{cases} \textit{strongly positive} & \textit{if } Comp > 0.5 \\ \textit{positive} & \textit{if } Comp \in [0.5, 0) \\ \textit{neutral} & \textit{if } Comp = 0 \\ \textit{negative} & \textit{if } Comp \in (0, -0.5] \\ \textit{strongly negative} & \textit{if } Comp < -0.5 \end{cases} \quad (1)$$

### 2.3.2 MACHINE LEARNING TECHNIQUES

Machine Learning (ML) algorithms have demonstrated high performance in several application domains such as user behaviour analytics, NLP, computer vision, and many others [19]. However, the type of problem to be solved determines the choice of the ML algorithm and dataset needed. Usually, supervised ML algorithms utilize both a training set and a test set for sentiment classification.

The polarity of a text can be accurately detected by classifiers created using supervised techniques. Such classifiers, nonetheless, perform admirably in the domain in which they were initially trained, but when the same classifier is applied to a different domain, their performance plummets dramatically [100].

The training set contains input feature vectors and their corresponding class labels. Using this training set, a classification model is developed that attempts to classify the input feature vectors into corresponding class labels. Thereafter, a test set is used to validate the model by predicting the class labels of unseen feature vectors.

Machine learning techniques such as Naive Bayes (NB) [22], Maximum Entropy (ME), K-Nearest Neighbour (k-NN), and Support Vector Machines (SVM) [20],[21] among several others are used for sentiment analysis tasks. However, pertinent literature in sentiment analysis suggests that SVM is the widely utilized learning algorithm by the research community [18] followed closely by k-NN and NB. As indicated in the literature [20], the support vector machine (SVM) is a supervised learning algorithm that can handle both classification and regression tasks. They are based on the concept of mapping data points from low-dimensional into high-dimensional space to make them linearly separable. Assuming there are  $n$  data points, the objective function of SVM is denoted as follows [21]:

$$\arg \min_w \left\{ \frac{1}{n} \sum_{i=1}^n \max \{0, 1 - y_i f(x_i)\} + Cw^T w \right\} \quad (2)$$

where  $w$  is a normalization vector;  $C$  is the penalty parameter of the error term, which is an important hyper-parameter of all SVM models.

The kernel function  $f(x_i)$  which is used to measure the similarity between two data points  $x_i$  and  $x_j$ , can be chosen from multiple types of kernels in SVM models. Therefore, the kernel type would be a vital hyper-parameter to be tuned. Common SVM kernel types include linear kernels, radial basis function (RBF), polynomial kernels, and sigmoid kernels.

Naïve Bayes (NB) algorithms are supervised learning algorithms based on Bayes' theorem. For a given number  $n$  dependent features  $x_{i=1}, \dots, x_n$  and target variable  $y$ , the objective function of NB can be denoted by:

$$y = \underset{y}{\operatorname{arg\,max}} P(y) \prod_{i=1}^n P(x_i|y) \quad (3)$$

Where  $P(y)$ ,  $P(x_i|y)$  represents the probability  $y$  and posterior probabilities of  $x_i$  given the values of  $y$  respectively. There exist different types of NB classifiers namely: Bernoulli NB (BNB), Gaussian NB (GNB), multinomial NB (MNB), and complement NB (CNB) [22]. K-Nearest Neighbour (k-NN), known for its simplicity and effectiveness is part of the popularly utilized classification techniques in machine learning. It is employed to categorize data in accordance with nearby or close-by training examples in a particular area by utilizing the Euclidean distance.

## 2.4 DEEP LEARNING MODELS

Many researchers in recent times are deploying deep learning (DL) algorithms for sentiment analysis due to the phenomenal successes it has achieved in NLP applications. DL models are based on the theory of Artificial Neural Networks (ANN). They utilize multiple layers of non-linear processing units that enable them to learn multiple representations and abstractions from raw data effectively [23].

Some common types of DL architectures are Deep Neural Networks (DNN), Convolution Neural Networks (CNN), and Recurrent Neural Networks (RNN) among several others [24]. CNN can learn the local response from temporal or spatial data. But it has the limitation of learning sequential correlation and long-distance context information since it is unable to identify the position of a word in a document.

In contrast with CNN, RNN is used for sequential modeling since they consider contextual information in the sentences which enables them to adequately handle the sequences better. Unfortunately, RNNs are not suitable for long sequences of sentences, because during training the components of the gradient vector can grow or vanish exponentially. Variants of RNN were developed by the researchers to overcome this drawback.

The Long Short-Term Memory (LSTM) neural network is one such variant that resolves this major RNN drawback. However, the LSTM, despite its long-lasting memory cannot hold information that is located too far from the current point. This problem is more prevalent when handling document-level sentiment classification.

For LSTMs to store longer information, various variant models have been proposed to enhance the capability of LSTMs to store long-range information. Bidirectional LSTM (Bi-LSTM) is another variant of RNN that solves the issues present in the LSTM [63].

## 2.5 METAHEURISTIC ALGORITHMS

Metaheuristic algorithms (MAs) refer to optimization techniques with global search capabilities that provide (near) optimal solutions for optimization problems. These algorithms are simple, flexible, derivative-free, and can avoid local optimal. They begin the optimization process by generating random solutions and exhibit stochastic behaviour.

A key characteristic of MA is its remarkable ability to prevent premature convergence. Metaheuristic algorithms are classified into four groups based on their behaviour namely: evolution-based, swarm intelligence-based, human behaviour-based, and physics-based algorithms. A detailed explanation of these groups can be found in [25], [41] as shown in Fig. 2.2.

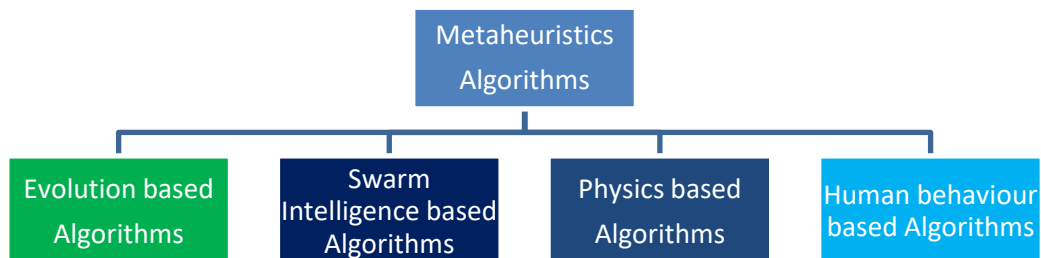


Fig. 2.2: Metaheuristic algorithm categories [41]

Over the past few years, several metaheuristic algorithms have been invented by researchers seeking to find solutions that are optimal or near optimal for optimization problems in different fields.

Interestingly, most of these MAs draw their inspiration from varied sources such as plants, insects, birds, sea creatures, reptiles, etc. while others tend to improve existing ones. Table 2.1, Table 2.2, and Table 2.3 show a classified list of some metaheuristic techniques that draw inspiration from nature. However, it must be mentioned that some of these metaphor-based algorithms have received their fair share of criticism from some researchers compared to some widely accepted ones like the genetic algorithm (GA), particle swarm optimization (PSO), differential evolution (DE), and others [104]. In this thesis, the PSO was

utilized due to its impressive performance in solving numerous feature selection tasks and widespread application in various disciplines [101].

Table 2.1 Metaheuristic methods inspired by insects and reptiles [65]

Method	Year
Butterfly Optimization Algorithm	2019
Gartener Snake Optimization	2017
Ant Star	2016
Improved Discrete Bees Algorithm	2016
Locust Swarm Algorithm	2015
Moth-Flame Optimization	2015
Alienated Ant Algorithm	2015
Dragonfly Algorithm	2015
Ant Lion Optimizer	2015
Bat Algorithm	2010
Social Spider Algorithm	2015
Dispersive Flies Optimization	2014
Fruit Fly Optimization	2012
Superbug Algorithm	2012
Bees Swarm Optimization	2012
Bacteria Colony Optimization	2012
Stochastic Diffusion Search	2011
Ant Colony Optimization	1999
Ant Colony System	1997
Firefly Algorithm	2009

Table 2.2 Metaheuristic methods inspired by birds and sea creatures [65]

Method	Year
Particle Swarm Optimization	1995
Haris Hawks Optimizer	2019
Cuckoo Search	2009
Dove Swarm Optimization	2009
Crow Search Algorithm	2018
Artificial Feeding Birds	2019
Chicken Swarm Optimization	2014
Laying Chicken Algorithm	2017
Eagle Strategy	2010
Emperor Penguins Colony	2019
Shark Search Algorithm	1998
Sail Fish Algorithm	2019
Bottlenose Dolphin Optimization	2017
Killer Whale Algorithm	2017
Whale Optimization Algorithm	2016
Artificial-Fish Swarm Optimization	2014

A metaheuristic algorithm is single solution-based (S-metaheuristics) or population size-based (P-metaheuristics) depending on the behaviour it exhibits during the exploration or exploitation phase. In the optimization phase, single solution-based metaheuristic algorithms only process one solution at a time whereas several processes can be processed at a time for population size-based metaheuristics algorithms [64].

Table 2.3 Metaheuristic methods inspired by plants, humans, and other animals [65]

Method	Year
Runner-Root Algorithm	2015
Flower Pollination Algorithm	2012
Artificial-Root Foraging Algorithm	2017
Invasive Weed Optimization	2010
Artificial Algae Algorithm	2015
Genetic Algorithm	1988
Human-behaviour-based Optimization	2017
Biogeography Based Optimization	2008
Queuing Search	2018
Focus Group	2018
Improvement of Position	2013
Krill Herd Algorithm	2017
Grey Wolf Optimizer	2014
Monkey Algorithm	2007
Camel Algorithm	2016
Jaguar Algorithm	2016
Lion Optimization Algorithm	2016
Artificial Buffalo Optimization	2015
Spotted Hyena Optimizer	2019

Sharma and Kaur report in [65] that metaheuristic techniques have been used more regularly to address various issues in robotics, education, and disease diagnostics (see Fig. 2.3). Nevertheless, sentiment analysis and fraud detection, however, are the least researched applications of metaheuristic algorithms.

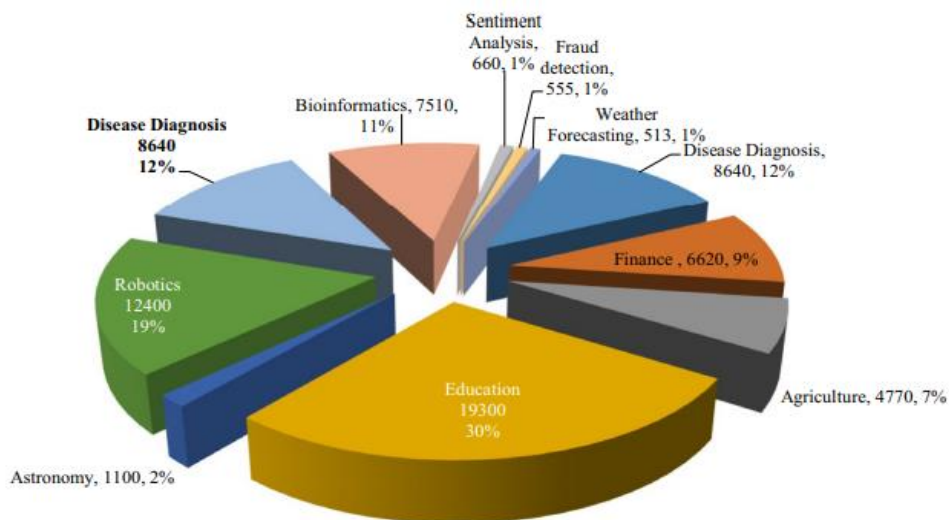


Fig. 2.3: Applications of metaheuristic techniques from various domains [65]

### 2.5.1 PARTICLE SWARM OPTIMIZATION (PSO)

In this segment, the Particle Swarm Optimization (PSO) algorithm together with its modified version called the BPSO is discussed. This is a bio-inspired evolutionary search technique originally formulated in [62] by Eberhart and Kennedy. This optimization technique is widely used in addressing research problems involving continuous and discrete search spaces. It is not influenced by the nonlinearity or size of the optimization problem and converges to the optimum solution in numerous problems where the analytical techniques mostly flop to converge. Consequently, PSO is efficiently incorporated into diverse optimization problems such as medical applications, power systems, feature selection, and many others.

Indeed, in a recent publication analysis conducted by the authors in [101], 65% of the 403 selected articles after analysing 3600 publications from reputable scientific databases from 2006 to 2020 confirms the popularity of the PSO. Thus, the authors aver that the overwhelming interest in the usage of the PSO across various disciplines may be related to the fact that it was the first swarm intelligence (SI) based algorithm proposed in SI literature coupled with its implementation simplicity. Hence, the above reasons motivated the author to utilize the PSO in this thesis.

A collection of individuals known as particles make up the PSO's swarm. Each swarm particle has a location and velocity section that identifies a unique solution and the direction of travel of that particle in the solution space. The PSO is structured into three steps as a repeating optimization process. Using random numbers, the PSO generates each particle's position and velocity sections in a first-stage process that initializes the population of the swarm. Next, the solutions corresponding to the particle positions are assessed. The final step is to update the particle locations and velocities using equations (4) and (5).

$$v_i^{t+1} = w * v_i^t + c_1 * r_1 * (P_b - x_i^t) + c_2 * r_2 * (G_b - x_i^t) \quad (4)$$

$$x_i^{t+1} = x_i^t + v_i^{t+1} \quad (5)$$

- $v_i^{t+1}$ : velocity component of a particle ( $p$ ) at  $(t + 1)$  iteration
- $x_i^{t+1}$ : position component of a particle ( $p$ ) at  $(t + 1)$  iteration
- $c_1, c_2$ : confidence coefficients
- $r_1, r_2$ : uniformly distributed random variables ranging from 0 to 1
- $w$ : inertia weight
- $P_b$ : individual position of particles ( $p$ 's) position
- $G_b$ : swarms global best position

A variant of the PSO called the binary particle swarm optimizer (BPSO) was once again invented by the authors in [39] which enabled the traditional PSO to solve discrete problems such as text classification. In the BPSO, Kennedy and Eberhart presented a particle's position as a vector with binary digits compared to the earlier case where it was considered as a vector containing continuous values. Consequently, the position is now updated using the following equation:

$$x_{id} = \begin{cases} 0, & \text{if } r \geq S(v_{id}) \\ 1, & \text{otherwise} \end{cases} \quad (6)$$

$$\text{where } S(v_{id}) = \frac{1}{1+e^{-v_{id}}} \quad (7)$$

### 2.5.1 GREY WOLF OPTIMIZATION

The intelligence, leadership abilities, and stalking strategies of grey wolves in the wild serve as the primary inspiration for GWO. Initially proposed by the authors in [74], the wolf packs usually have a rigorous hierarchy (see Fig. 2.4), with the alpha ( $\alpha$ ) serving as the pack's leader in charge of all group activities such as feeding, hunting, and migration.

A beta ( $\beta$ ) wolf, who takes over as pack leader if the alpha ( $\alpha$ ) wolf is injured or killed, is found in the second level of the hierarchy. Next is the delta wolves ( $\delta$ ) followed by the omegas ( $\omega$ ). Algorithm 1 describes the pseudocode of the grey wolf optimizer. The formulation of the GWO mathematical model entails encircling, stalking, and attacking the prey. Equation 8 describes the encircling behaviour of the GWO.

$$\vec{X}(t+1) = \vec{X}_p(t) + (\vec{A} * \vec{D}) \quad (8)$$

$$\vec{D} = |\{\vec{C} * \vec{X}_p(t)\} - \vec{X}(t)| \quad (9)$$

$\vec{X}$  represents the wolves vector location in dimensional space,  $d$ ,  $\vec{X}_p$  denotes the prey's vector position at iteration  $t$  with  $\vec{A}$  and  $\vec{C}$  as coefficient vectors. Equation (9) shows the distance between each wolf and the prey.

Equations 10 and 11 express coefficient vectors  $\vec{A}$  and  $\vec{C}$  mathematically:

$$\vec{A} = (2\vec{a} * \vec{r}_1) - \vec{a} \quad (10)$$

$$\vec{C} = 2 * \vec{r}_2 \quad (11)$$

A set vector  $\vec{a}$  linearly reduces over iterations from 2 to 0,  $\vec{r}_1$  and  $\vec{r}_2$  are defined as vectors randomly inside the [0,1] range. The stalking behaviour of the wolves is mathematically modeled using equations (12), (13), and (14). While beta and delta



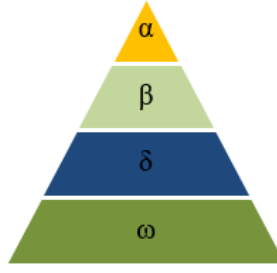


Fig. 2.4: Social hierarchy of grey wolves [74]

wolves are anticipated to possess enough expertise regarding the likely location of the prey, alpha wolves are considered to be the finest candidates for the solution during the hunting process.

---



---

Algorithm 1 Pseudocode of Grey Wolf Optimizer [74]

<p><b>input:</b> Number of grey wolves.  Total number of Iter.  <b>output:</b> Optimal grey wolf position (<math>x_\alpha</math>).  Best fitness value <math>f(x_\alpha)</math>.  Randomly initialize the population of grey wolves' positions.  Compute <math>\alpha</math>, <math>\beta</math>, and <math>\delta</math> solutions based on their fitness values  <b>while</b> (termination condition not satisfied) <b>do</b>      <b>for</b> (each wolf <math>W_i \in</math> pack) <b>do</b>          update current wolf position based on equation (7)      <b>end</b>  Update A, <math>a</math> and <math>c</math>  Evaluate the positions of the individual wolves  Update alpha(<math>\alpha</math>), beta(<math>\beta</math>), and delta (<math>\delta</math>)  <b>end</b></p>
---

Thus, the best three solutions obtained based on a given iteration are retained, which then compels others (such as Omega) to update their locations inside the search (decision) area using the following equations:

$$\begin{aligned}\vec{D}_\alpha &= |(\vec{C}_1 \cdot \vec{X}_\alpha) - \vec{X}|, \\ \vec{D}_\beta &= |(\vec{C}_2 \cdot \vec{X}_\beta) - \vec{X}|, \\ \vec{D}_\delta &= |(\vec{C}_3 \cdot \vec{X}_\delta) - \vec{X}| \\ \vec{X}_1 &= \vec{X}_\alpha - \{A_1 * (\vec{D}_\alpha)\},\end{aligned}\tag{12}$$

$$\vec{X}_2 = \vec{X}_\beta - \{A_2 * (\vec{D}_\beta)\},$$

$$\vec{X}_3 = \vec{X}_\delta - \{A_3 * (\vec{D}_\delta)\} \quad (13)$$

$$\vec{X}(t + 1) = \frac{\vec{X}_1 + \vec{X}_2 + \vec{X}_3}{3} \quad (14)$$

The vector  $\vec{a}$  in the GWO is one of the key parameters that must be tuned to ensure exploitation and exploration. It is a random vector with a value found inside  $[-a, a]$  range. Accordingly, it is recommended to decrease each vector's dimension from 2 to 0 so that it is linearly proportionate to the number of iterations. The updated equation is shown below:

$$\vec{a} = 2 - t * \left\{ \frac{2}{\max_i \mathcal{J}} \right\} \quad (15)$$

$t$  and  $\mathcal{J}$ , respectively, represent the iterations number per iteration and the overall number of iterations. According to the authors in [59], [60], the GWO has few parameters to tune, favourable convergence can be reached quite easily while a good balance between exploration and exploitation can also be achieved with simplicity. Furthermore, it is easy to use, scalable, simple, and flexible. According to [106], GWO demonstrated impressive results against metaheuristic algorithms such as particle swarm optimization (PSO), gravitational search algorithm(GSA), genetic algorithm(GA), differential evolution (DE), etc... Furthermore, it also showed competitive performance in the exploitation phase compared to other metaheuristic algorithms.

Facial recognition, EMG signal classification, disease diagnosis, gene selection and intrusion detection systems, parameter tuning, and economic dispatch are some of the fields where GWO has been widely utilized [66]. Emary et al., in [42] proposed a wrapper feature selection technique where the binary GWO was used for feature selection with the k-NN classifier and evaluated using 18 benchmark datasets from the UCI machine learning repository. The results obtained outperformed the GA and PSO in terms of accuracy and the number of features reduced. Besides the GWO's unique characteristics, it was chosen for use in this thesis by the author because it was originally part of the hybrid metaheuristic algorithm that was modified using angle modulated PSO.

### 2.5.2 WHALE OPTIMIZATION ALGORITHM

The whale optimization algorithm (WOA) is a member of the group of stochastic population-based algorithms that Mirjalili and Lewis developed [71].

In a recent study performed by [107], the authors employed a wrapper-based WOA for feature selection using 16 datasets culled from the UCI ML data repository. The results realized from the study show better WOA classification performance when compared to the GA and PSO which demonstrates WOA's ability to search the space for optimal feature subset. Again, [108] in their work

employed a wrapper-based WOA to determine the optimal feature subset from low sample medical datasets with high dimensionality. The wrapper-based WOA technique had greater success on three evaluation criteria viz classification accuracy, the best fitness value, and the number of selected features, when compared with some of the notable state-of-the-art algorithms (i.e. binary GWO, PSO, and GSA) from metaheuristic feature selection literature thereby confirming the “*No-Free-Lunch theory*” in optimization.

The WOA imitate how humpback whales forage by using bubble nets. WOA employs three phases: finding the prey, enclosing the victim, and bubble-net attacking tactics. The first phase, looking for prey, is used for exploration, whilst the second and third stages, circling prey, and bubble-net assault, are utilized for exploitation. WOA begins by initializing the parameters together with a collection of the initial search agents. Depending on the parameter values, the search method alternates between the global search and local search phases, and on each iteration, it determines the best ideal value.

A predetermined number of iterations will be completed before the procedure ends and the best search agent value is returned. WOA is simple to implement because there are not many parameters to tune [72]. WOA traverses between the exploration and exploitation phases while updating an ideal solution with a probability value, which results in higher randomness and faster convergence [73].

The exploration phase is modelled as follows:

Equations (16) and (17), a mathematical model of a whale's movement around a prey, are utilized to update a solution.

$$\vec{D} = |\vec{C} \cdot \vec{X}_t^* - \vec{X}_t| \quad (16)$$

$$\vec{X}_{t+1} = \vec{X}_t^* - \vec{A} \cdot \vec{D} \quad (17)$$

where the current iteration denotes  $t$ ,  $X^*$ , reflects the most successful solution so far, as determined by equations (18) and (19),  $A$  and  $C$  are coefficient vectors whereas the present solution is denoted as  $X$ .

$$\vec{A} = 2\vec{a} \cdot r - \vec{a} \quad (18)$$

$$\vec{C} = 2 * \vec{r} \quad (19)$$

$a$  is lowered linearly from 2 to 0, and  $r \in [0,1]$ , where  $r$  is a random vector. As determined by equation (17), the locations of the solutions are updated in accordance with the location of the prominent solution. The regions where a solution can be found in the vicinity of the optimal solution are controlled by varying the values of the  $A$  and  $C$  vectors. The humpback whales approach their prey in a spiraling motion and by reducing their encircling mechanism. By lowering the value of  $a$  in equation (18) in accordance with equation (19), WOA simulates the shrinking encircling behaviour.

$$\vec{a} = 2 - t * \left\{ \frac{2}{\max_i \mathcal{J}} \right\} \quad (20)$$

$\max_i \mathcal{J}$  denotes the highest number of iterations permitted. Computing the separation between solution  $X$  and the leading solution  $X^*$  results in a spiral-shaped path. Thus, as determined by equation (21), a spiral equation is formed between the existing solution and the ideal solution.

$$\vec{X}_{t+1} = D' \cdot e^{bl} \cdot \cos(2\pi l) + \vec{X}_t^* \quad (21)$$

$\vec{D}$  denotes the distance between a whale  $X$  and a prey,  $b$  defines the spiral's shape of the spiral, and  $l$  is a random number  $l \in [-1,1]$ .

In order to describe the two mechanisms; the ascending spiral-shaped path and the diminishing encircling mechanism; a probability of 50% is used in the optimization equation (22).

$$\vec{X}_{t+1} = \begin{cases} \vec{X}_t^* - \vec{A} \cdot \vec{D} & \text{if } p < 0.5 \\ D' \cdot e^{bl} \cdot \cos(2\pi l) + \vec{X}_t^* & \text{if } p \geq 0.5 \end{cases} \quad (22)$$

where  $p \in [0,1]$ ,  $p$  is a random number.

Below is a description of the exploration phase.

Instead of requiring the solutions to search randomly based on the location of the best solution discovered thus far, a randomly selected solution is utilized to update the position appropriately which improves the exploration in WOA. Thus, a vector  $A$  with random values higher than 1 or less than 1 is employed to shift a solution far from the most effective search agent. Mathematical models of this mechanism can be found in equations (23) and (24).

$$\vec{D} = |(\vec{C} \cdot \vec{X}_{rand}) - \vec{X}| \quad (23)$$

$$\vec{X}_{t+1} = \vec{X}_{rand} - \vec{A} * \vec{D} \quad (24)$$

$\vec{X}_{rand}$  is a randomly selected whale from the present population. The WOA was chosen for comparison in this thesis due to the limited WOA literature available on feature selection compared to the PSO and GWO.

### 2.5.3 ANGLE MODULATED PARTICLE SWARM OPTIMIZATION (AMPSO)

The angle modulated PSO (AMPSO) is a discrete optimization method that utilizes the standard particle swarm optimizer to optimize binary problems without making any amendments to the PSO algorithm. The technique was initially proposed by Franken [29] with Pamepara et al., [30]. introducing it

formally in their work. The coefficients of the trigonometric function indicated in equation (25) are optimized by the AMPSO using the PSO. This trigonometric function is also dubbed *generating function g*.

$$g(x) = \sin[2\pi(x - a)b \cos(2\pi(x - a)c)] + d \quad (25)$$

The coefficients  $a$ ,  $b$ ,  $c$ , and  $d$  are defined as follows:

- $a$ : regulates the horizontal shift.
- $b$ : regulates the sine wave's frequency as well as the cos wave's amplitude.
- $c$ : regulates the cos wave's frequency.
- $d$ : regulates the vertical shift.

The *generating function's* shape is controlled by its coefficients due to the way it acts on the function's displacement and frequency. When searching for optimal coefficients for  $g$ , the PSO, the location of a particle  $i$  is identified by  $x_i = (a, b, c, d)$ . Before a binary solution can be produced, the particle's position must be inserted into the function  $g$ .

Following that, the function  $g$  is subsequently sampled periodically at  $x = 1, 2, 3, \dots, n_b$ , where  $n_b$  defines the number of binary digits needed. By taking notice of the value of  $g(x)$  for every sampling position, a binary solution  $B \in \mathbb{B}^{n_b}$  is produced:

$$B_j = \begin{cases} 0 & \text{if } g(x) \leq 0 \\ 1 & \text{otherwise} \end{cases} \quad (26)$$

Fig. 2.5 illustrates the entire process. This, the binary solution is assessed to determine the fitness value  $f(B)$ , where  $f$  is the  $n_b$ -dimensional binary objective function, for the given particle.

Hence, 4-dimensional PSO is used to solve an  $n_b$ -dimensional binary problem. Since the coefficients are continuous, no modification is made to the PSO. For the illustration depicted in Fig. 2.5, the generating function  $g$  is sampled periodically, resulting in the creation of a five-dimensional binary solution. For this example  $x_i = (0.0, 0.5, 0.8, 0.0)$  where the coefficients  $a, b, c$  and  $d$  have the values  $0.0, 0.5, 0.8$  and  $0.0$  respectively.

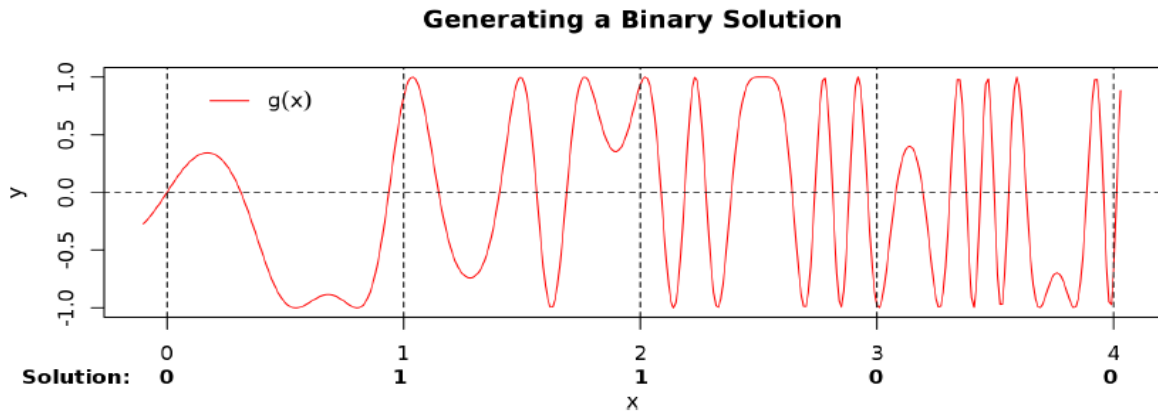


Fig. 2.5: A 5-dim binary solution generated by sampling  $g$  at regular intervals [31]

## 2.6 FEATURE SELECTION

Feature selection (FS) is one of the most critical and challenging problems in machine learning. It has two primary conflicting objectives namely, maximizing the classifier performance and reducing the number of features to overcome the curse of dimensionality. In the last few decades, several researchers have relied on metaheuristic algorithms to solve feature selection problems in various domains such as text mining, bioinformatics, industrial applications, computer vision, and others.

Generally, FS techniques are classified into three categories: filter, wrapper, and embedded methods [28], [29] (see Fig. 2.6). Filter methods are independent of learning or classification algorithms. It always focuses on the general characteristics of the data [16]. Wrapper methods always include the classification algorithm and interact with the classifier. These are computationally expensive methods than the filter and provide more accurate results as compared to filter methods. Embedded methods are a combination of filters and wrapper methods. According to Jovic et al., [33], generating features subset using the wrapper approach can be organized into the following three different search strategies namely, sequential, randomized, and exponential strategies [36].

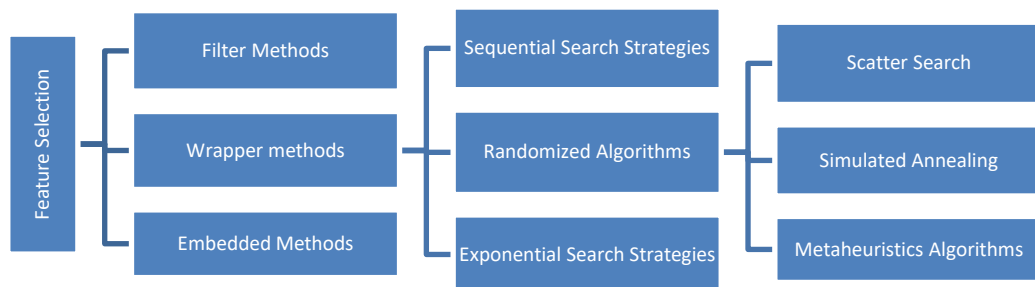


Fig. 2.6: Classification of Feature Selection methods [41]

Embedded models embed feature selection within an ML classifier. Such models combine the strengths of both filter and wrapper-based methods.

## 2.7 HYBRID METAHEURISTIC ALGORITHMS FOR FEATURE SELECTION

Search space exploration and optimal solution exploitation are two essential contradictory principles usually considered when utilizing metaheuristic algorithms. A good option will be to consider a hybrid technique (also referred to as the memetic method) which entails combining two or more metaheuristic algorithms to enhance the performance of the algorithms involved [37].

Example applications include a hybrid Genetic Algorithm (GA) with PSO (GPSO) as a wrapper for feature selection with the SVM in [34], [35] using the microarray data and digital mammography datasets. Again, a hybrid Differential Evolution (DE) and Artificial Bee Colony (ABC) were proposed (DEABC) for feature selection [38]. In [37], Mafarja and Mirjarlili designed a hybrid wrapper feature selection method by integrating the simulated annealing (SA) algorithm with the whale optimization algorithm (WOASAT-2) based on the low-level teamwork (LTH) and high-level relay (HRH) hybrid models. During this hybridization process, the tournament mechanism was applied to improve the diversity of the WOA.

More recently in [28], a continuous hybrid GWO and PSO (GWOPSO) was proposed. In their work, the authors sought to improve the GWOPSO's ability to efficiently exploit optimal solutions and explore the search space using the PSO and GWO. The author of this thesis proposes the hybrid angle modulated algorithm (described in more detail in section 5.2) which generates a binary solution in the search space using equations 25 and 26, as opposed to the binary GWOPSO (BGWOPSO) [40], which employed a sigmoid function to update the locations of the search agents. A detailed discussion and mathematical formulations of the continuous and proposed hybrid method are presented in chapter 5.

In this thesis, the AMGWOPSO is used as a wrapper for feature selection with the k-NN classifier. The k-NN classifier was chosen due to its impressive track record in FS literature involving wrapper approaches coupled with its relative simplicity and speed during training and validation [41].

As the literature shows, the feature selection problem has two competing goals. That is, minimizing the number of features by removing redundant features in the feature set while maximizing our classification accuracy. To realize both objectives I adopted the fitness function in [42] shown below (27):

$$fitness_{am} = \varphi\gamma(B) + \mu \frac{|N_f|}{|N_T|} \quad (27)$$

Where  $|N_f|$  depicts the size of attributes in the features subset,  $|N_T|$ , the size of attributes in the given dataset,  $\varphi = [0,1]$  and  $\mu = (1 - \varphi)$  are parameters adapted from [42] while  $\gamma(B)$  depicts the error rate of the k-NN classifier.



### 3. OBJECTIVES OF THE THESIS

This thesis employs the relevant text mining and evolutionary computation techniques for sentiment analysis and feature selection tasks.

Consequently, the main objective of this thesis along with the specific steps to be taken is formulated based on the above ideas:

1. Modification of the sentiment analysis pipeline to improve classification:
  - a. To determine the best lexicon-based technique based on classification performance and also identify tweet (text) contents most illustrative of positive and negative value user contribution.
2. Design a metaheuristic-based solution for sentiment analysis using the binary PSO (BPSO) given its impressive performance in solving numerous feature selection tasks.
3. Develop a new Angle Modulated-based metaheuristic memetic method for wrapper feature selection. The proposed method utilizes the GWO and AMPSO (AMGWOPSO).
4. Test and evaluate the proposed metaheuristic and memetic method on some selected publicly available benchmark datasets from the University of California, Irvine (UCI) ML repository [32].

The following steps will be pursued to help achieve the aims captured in the thesis:

Literature review and analysis:

1. Of available published literature related to sentiment analysis and metaheuristic algorithms
2. Of current hybrid metaheuristic approaches and examine their potential modifications as it applies to feature selection.

Experiments are conducted:

1. To verify the effectiveness of the proposed approaches for sentiment analysis tasks and metaheuristic algorithm-based feature selection (see results section).
2. To test the newly developed hybrid AMGWOPSO.

Evaluation:

1. Of the selected lexicons for sentiment analysis using tweets crawled from the Twitter handle of an entity of interest (financial institution) to identify insights most illustrative of positive and negative value-user contribution (see section 6.1).
2. Applicability of the proposed metaheuristic technique for feature selection on some selected benchmark datasets (see sections 6.2 and 6.3).

## 4. WORKFLOW

A brief description of the sentiment lexicons utilized together with the sentiment analysis workflow adopted in this thesis is presented in this section.

### 4.1 SENTIMENT ANALYSIS WORKFLOW

The diagram in Fig. 4.1 shows the workflow adopted for sentiment analysis in this study. Like all other social media research, data collection precedes data preprocessing, data splitting and transformation, model training, and evaluation. Data collection involves collecting tweets using a tool called Twint [102] which allows the utilization of specific keywords or the social media handle of the entity of interest. Data preprocessing helps to filter out and remove noise in the dataset after which I generate labels for the dataset based on each lexicon. Furthermore, the author splits the dataset into training and testing sets which help in verifying if the model has learned the generalized features to enable it to handle unseen data effectively. Details regarding the implementation of this workflow/pipeline are shown in the case study illustrated in chapter 6.

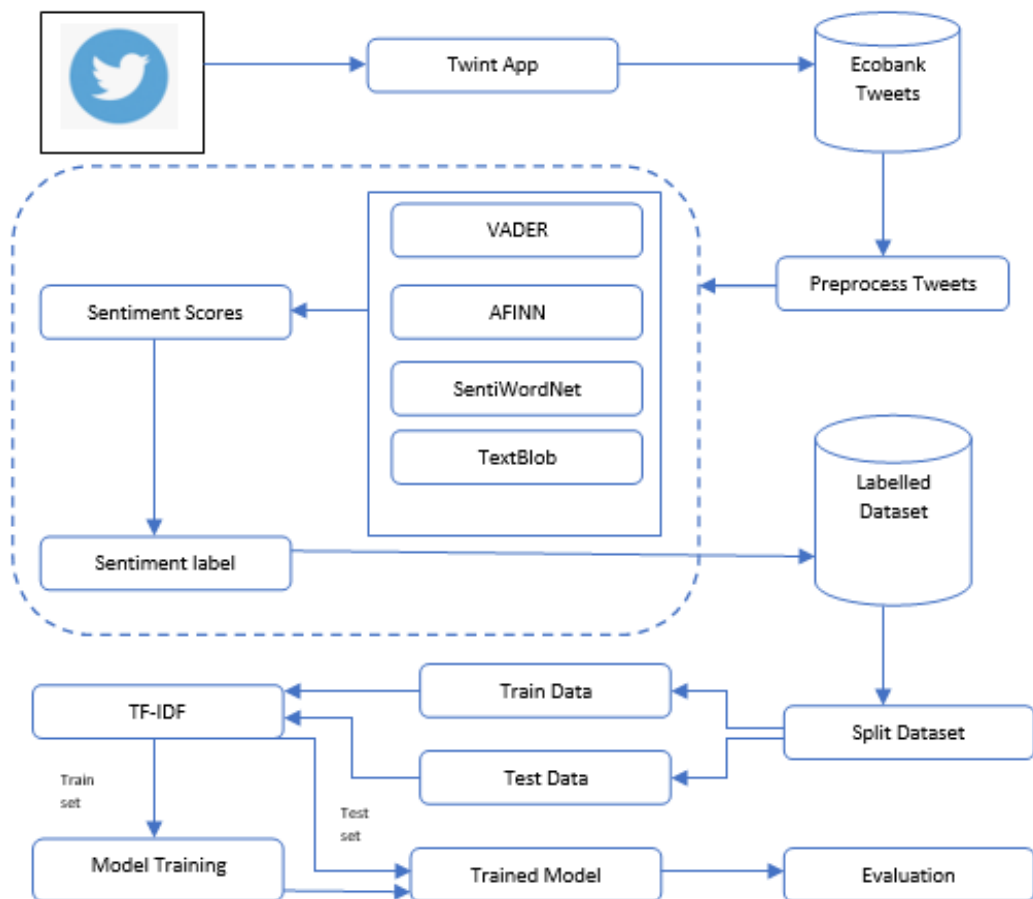


Fig. 4.1: Overview of the sentiment analysis workflow

A brief description of the lexicons shown in the workflow is presented below.

- VADER [15]: a lexicon and rule-based sentiment analysis tool that is specifically attuned to sentiments expressed via social media. It is fast and computationally efficient without compromising accuracy. Training data is not required and works very well on social media text. Additionally, it does not suffer extremely from a speed-performance trade-off. Given a document, VADER examines its lexical features to determine an initial sentiment score before applying five different rules based on grammatical conventions and syntax to amend that score. These rules handle capitalization and exclamation marks as sentiment amplifiers. Again, they also handle negations and contrastive conjunctions very well. VADER produces positive, negative, neutral, and compound scores for each tweet in the dataset. The positive, negative, and neutral scores are ratios for the proportions of text that lie in these categories. The compound score is a metric that sums up all the lexicon ratings which have been normalized between -1(most extreme negative) and +1(most extreme positive) [15].
- AFINN [44]: The AFINN lexicon created by Finn contains a list of English terms manually rated for valence with an integer ranging between -5 (negative) and +5 (positive) by Finn Arup Nielsen [46]. This lexicon is better equipped to handle tweets expressed using internet slang and obscene words.
- SentiWordNet [13]: a lexical resource for opinion mining and an extension of WordNet in which 147,306 synonym sets (synsets) are annotated with three numerical scores relating to positivity, negativity, and objectivity [13]. They describe how positive, negative, and objective the terms in the synset are. Each of the sentiment scores ranges from 0.0 to 1.0 and sums up to 1.0 for each synset.
- TextBlob: a library for natural language processing that works very well with python [46]. It assigns both polarity and subjectivity scores to each text in the dataset. The sentiment property returns a named tuple of the form sentiment (polarity, subjectivity). The polarity score is a float that ranges within the interval [-1,1]. The subjectivity is a value within the range [0.0, 1.0] where 0.0 and 1.0 are considered very objective and very subjective respectively.

## 4.2 FEATURE SELECTION WORKFLOW

The diagram below (Fig. 4.2) illustrates the working mechanism of feature selection as utilized in this study. The original data set, a subset of the chosen attributes, the evaluation mechanism, the selection criterion, and ultimately the validations constitute the five main components of the feature selection process.

Feature selection (FS) may appear to be an uncomplicated issue, but that is not the case in reality. The feature selection dilemma presents some difficulties. Managing high dimensional data, feature relevance, and feature redundancy are some of the main problems with feature selection as indicated earlier.

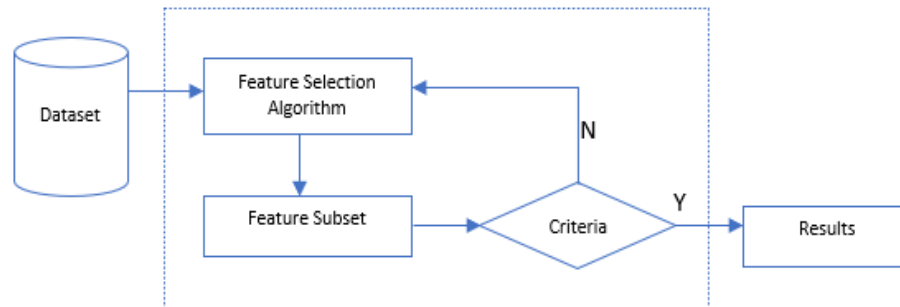


Fig. 4.2: Feature selection process [41]

Thanks to the advent of Web 2.0, access to incredibly huge volumes daily for ML tasks is no longer a challenge. Moreover, it is possible to store such a huge volume of data at a relatively low cost because of improvements in hardware technology. Again, the majority of text processing and biomedical applications deal with high dimensional data leading to the development of ML models that are difficult to understand and deploy.

It is possible to assess the importance of a feature by looking at its contribution to the classification procedure. A feature can be entirely irrelevant or weakly irrelevant. When choosing features, all insignificant features will be outrightly ignored. Nonetheless, the prevailing challenge and circumstances dictate whether a feature is included or not.

Even though relevance represents a key consideration in the feature selection process, redundancy is also another crucial concern. There are situations where more than one feature may contribute to the same piece of data. In choosing the best features subset, all attributes with potential redundancy-causing qualities will be discarded.

## 5. HYBRID BIO-INSPIRED FEATURE SELECTION TECHNIQUE USING ANGLE MODULATION

The goal of this section is to succinctly present some basic concepts, definitions, and mathematical formulations associated with the continuous hybrid PSOGWO, and the proposed AMGWOPSO.

Instructively, the “*No Free Lunch*” doctrine in optimization suggests that there exists no metaheuristic technique equipped with the capability to provide optimal solutions to every optimization problem. Indeed, it is worth mentioning that while an algorithm may perform well on some datasets, its performance may suffer greatly when applied to a different dataset [57], [68].

Thus, it is important to create novel or hybrid methods to optimally address a particular problem or set of problems. Additionally, it proves that a particular optimization technique can only effectively handle a limited set of optimization issues. Consequently, utilizing a hybrid technique in which two or more algorithms are integrated to increase how well each algorithm performs is one sure way to establish a healthy balance between their exploitative and exploratory abilities.

### 5.1 CONTINUOUS HYBRID PSOGWO

Hybridization of PSO with GWO algorithm (PSOGWO) was put forth in [28]. The primary goal of the PSOGWO is to increase the strength of both variants (i.e., PSO and GWO) by enhancing the GWO’s exploration and PSO’s exploitation capabilities. By using equation (28), the hybrid PSOGWO updates the locations of the first three agents inside the search area. Again, instead of employing conventional mathematical formulas to regulate the exploration and exploitation of the grey wolf in the search area, the inertia parameter is employed. Equations (29) and (30) are employed when updating the velocity and positions of the agents inside the search space when hybridizing PSO and GWO variants.

$$\begin{aligned}\vec{D}_\alpha &= |(\vec{C}_1 \cdot \vec{X}_\alpha) - (w * \vec{X})| \\ \vec{D}_\beta &= |(\vec{C}_2 \cdot \vec{X}_\beta) - (w * \vec{X})|, \\ \vec{D}_\delta &= |(\vec{C}_3 \cdot \vec{X}_\delta) - (w * \vec{X})|\end{aligned}\tag{28}$$

$$v_i^{k+1} = w * (v_i^k + c_1 \cdot r_1 (x_1 - x_i^k) + c_2 \cdot r_2 (x_2 - x_i^k) + c_3 \cdot r_3 (x_3 - x_i^k))\tag{29}$$

$$x_i^{k+1} = x_i^k + v_i^{k+1}\tag{30}$$

## 5.2 PROPOSED HYBRID ANGLE MODULATED GWOPSO (AMGWOPSO)

The hybrid PSOGWO's search space is continuous. The binary nature of the FS task requires the development of a binary system to transform the solutions from the continuous search space to a discrete search space. Al-Tashi et al., in [40] developed a binary variation of the hybrid PSOGWO called the binary GWOPSO (BGWOPSO) for FS. In this text, the author of the thesis hybridizes the GWO with the AMPSO [30] to form a new hybrid algorithm called the Angle Modulated GWOPSO (AMGWOPSO) for feature selection tasks.

A low-level coevolutionary mixed hybrid approach is used in this dissertation when combining AMPSO and GWO. The resulting hybrid algorithm is regarded as low-level coevolutionary because the two algorithms' functionalities are merged and applied concurrently rather than sequentially. Again, it is largely referred to as a mixed hybrid algorithm because two separate variants are used in solving the problem. Based on this modification, the author enhances the exploitative capabilities in the angle modulated PSO with the Grey Wolf Optimizer's exploration ability to enhance the strength of both variants.

In the proposed AMGWOPSO, a hybrid PSOGWO is used to optimize the trigonometric generating function's coefficients. As indicated earlier, the *generating function's* shape is controlled by its coefficients. For a binary solution to be created, the coefficients found by the hybrid PSOGWO are initially inserted into the generating function  $g$ . Following that, the function  $g$  is subsequently sampled periodically. Then, a binary digit is assigned to represent the value of  $g$  at each sampling interval. Hence, in a  $g_c$ -dimensional real-valued space, AMGWOPSO is capable of solving  $n_b$ -dimensional binary task. It must be stated, however, that the number of coefficients of  $g$  is denoted by  $g_c$ .

## 5.3 MODELING THE HYBRID AMGWOPSO

Due to the continuous domain of their location vectors, agents in the continuous PSOGWO constantly move across the search space. The wolf update mechanism according to [42] is a function of the three vector positions  $X_1$ ,  $X_2$  and  $X_3$ , which elevates each wolf to the top three solutions. Consequently, the proposed Angle Modulated GWOPSO generates a binary solution in the search space using equations (31) and (32) as opposed to the binary GWOPSO, which employs a sigmoid function to update the locations of the search agents. Algorithm 2 describes the proposed AMGWOPSO technique's pseudocode.

$$g(\lambda) = \sin[2\pi(\lambda - a)b \cos(2\pi(\lambda - a)c)] + d \quad (31)$$

$$X_d(t + 1) = \begin{cases} 0, & g(\lambda) \leq 0 \\ 1, & otherwise \end{cases} \quad (32)$$

$$\text{where } \lambda = \frac{X_1 + X_2 + X_3}{3}$$

$X_d(t + 1)$ , represents the binary updating location mechanism at iteration  $t$  in the dimension  $d$ ,  $g(\lambda)$  refers to the generating function shown in equation (31).

$$X_1^d(t + 1) = \begin{cases} 1, & \text{if } (X_\alpha^d + \mathbb{B}_\alpha^d) \geq 1 \\ 0, & \text{otherwise} \end{cases}$$

$$X_2^d(t + 1) = \begin{cases} 1, & \text{if } (X_\beta^d + \mathbb{B}_\beta^d) \geq 1 \\ 0, & \text{otherwise} \end{cases}$$

$$X_3^d(t + 1) = \begin{cases} 1, & \text{if } (X_\delta^d + \mathbb{B}_\delta^d) \geq 1 \\ 0, & \text{otherwise} \end{cases} \quad (33)$$

Where  $X_\alpha^d$ ,  $X_\beta^d$ , and  $X_\delta^d$  represents the  $\alpha$ ,  $\beta$ , and  $\delta$  wolves' position vectors respectively in the  $d$ -dimension.

$$\mathbb{B}_{\alpha,\beta,\delta}^d = \begin{cases} 1, & \text{if } r \leq \mathbb{C}_{\alpha,\beta,\delta}^d \\ 0, & \text{otherwise} \end{cases} \quad (34)$$

$\mathbb{B}_{\alpha,\beta,\delta}^d$  is a  $d$ -dimension binary step defined using equation (34), and random value  $r$ , derived from the uniform distributed  $\in [0,1]$ .  $\mathbb{C}_{\alpha,\beta,\delta}^d$  refers to the dimension  $d$ 's continuous value computed using equation (35) below [66]:

---

Algorithm 2 Pseudocode of proposed AMGWOPSO algorithm

---

**input:** Population of search agents  
Total number of Iter.  
**output:** Elite positions and their optimal fitness value  
Randomly initialize the population of search agents  
Initialize parameters  $t$ ,  $a$ ,  $A$ ,  $C$ ,  $w$   
Compute the search agent's fitness values  $X_\alpha$ ,  $X_\beta$  and  $X_\delta$   
**while** ( $t < \text{Total number of Iter.}$ ) **do**  
  **for** (each agent population) **do**  
    Update the velocity of agents using equation (31)  
    Update the search agent positions by transforming the new positions into a binary form using equation (32)  
  **end**  
  **I.** Update parameters  $a$ ,  $A$ ,  $C$ ,  $w$   
  **II.** Use the objective function to evaluate the particle positions  
  **III.** Update locations of the best three search agents ( $\alpha$ ,  $\beta$ ,  $\delta$ )  
   $t = t + 1$   
**end while**

---

$$\mathbb{C}_{\alpha,\beta,\delta}^d = \begin{cases} 0, & g(\xi) \leq 0 \\ 1, & \text{otherwise} \end{cases} \quad (35)$$

$$\begin{aligned} & \text{where } \xi = \mathbb{A}_1^d \cdot \mathbb{D}_{\alpha,\beta,\delta}^d \\ g(\xi) &= \sin[2\pi(\xi - a)b \cos(2\pi(\xi - a)c)] + d \end{aligned} \quad (36)$$

In the proposed AMGWOPSO algorithm, equation (33) is employed to update the positions of the best three solutions. Synonymous with the continuous hybrid PSOGWO [28], the inertia weight is utilized to regulate the exploration and exploitation modeled mathematically by equation (38). Finally, equations (38) and (39) are utilized for velocity and position updates respectively.

$$\vec{D}'_{\alpha} = |\vec{C}_1 \cdot \vec{X}_{\alpha} - w * \vec{X}|,$$

$$\vec{D}'_{\beta} = |\vec{C}_2 \cdot \vec{X}_{\beta} - w * \vec{X}|,$$

$$\vec{D}'_{\delta} = |\vec{C}_3 \cdot \vec{X}_{\delta} - w * \vec{X}| \quad (37)$$

$$\begin{aligned} v_i^{(k+1)'} &= w * (v_i^{k'} + c_1 \cdot r_1 (x_1 - x_i^{k'}) + c_2 \cdot r_2 (x_2 - x_i^{k'}) + \\ & c_3 \cdot r_3 (x_3 - x_i^{k'})) \end{aligned} \quad (38)$$

$$x_i^{(k+1)'} = x_d^{(t+1)'} + v_i^{(k+1)'} \quad (39)$$

It is worth mentioning that  $x_d^{(t+1)'}$  and  $v_i^{(k+1)'}$  are computed according to equations (32) and (38) respectively.

## 5.4 EXPERIMENTAL SETUP FOR PROPOSED AMGWOPSO

This segment presents the experiments designed to assess the potency of the new hybrid AMGWOPSO technique. Since the author of this thesis aims to decrease the number of attributes while maximizing the classification accuracy, the proposed AMGWOPSO is wrapped with the k-NN classifier for this task. To realize both objectives, the fitness function in equation (27) was utilized. For this study, the entire set of solutions referred to as the solution set  $\mathbf{s}$ , is represented by a binary string of length  $\mathbf{M}$ , where  $\mathbf{s} = (s_1, s_2, s_3, \dots, s_M)$ . A bit  $s_i$ , in the solution set  $\mathbf{s}$  is denoted as 1 if the associated feature  $s_i$  is chosen and 0 otherwise. As stated earlier, the k-NN classifier was chosen due to its impressive track record in FS literature involving wrapper approaches coupled with its relative simplicity and speed during training and validation.



## 5.5 DATASETS

Eighteen (18) benchmark freely available datasets from the UCI ML storehouse were selected to test and validate the proposed hybrid AMGWOPSO. These selected datasets encompass different domains and also vary in terms of instances and attributes. The experiments were performed using Python 3.8 running on a Windows 10 Professional 64bit computer system equipped with a 3.4 GHz Intel® Core™ i7-7700 Processor and 32 GB memory. The benchmark datasets utilized are described in Table 5.1.

Table 5.1: UCI datasets used [32]

Number	Dataset	# Features	#Instances
1	Breastcancer	9	699
2	BreastEW	30	569
3	CongressEW	16	435
4	Exactly	13	1000
5	Exactly2	13	1000
6	HeartEW	13	270
7	IonosphereEW	34	351
8	KrVsKpEW	36	3196
9	Lymphography	18	148
10	M-of-n	13	1000
11	PenglungEW	325	73
12	SonarEW	60	208
13	SpectEW	22	268
14	Tic-tac-toe	9	958
15	Vote	16	300
16	WaveformEW	40	5000
17	WineEW	13	178
18	Zoo	16	101

## 6. RESULTS

In this section, the author of this thesis presents and discusses the results obtained in chronological order during the development of methods as part of the thesis preparation process. Specifically, some functional examples using the lexicon-based sentiment analysis approach and the metaheuristic algorithms for feature selection are provided in the following sub-sections.

### 6.1 DEDUCTIONS FROM A SUB-SAHARAN AFRICAN BANK: A SENTIMENT ANALYSIS APPROACH

The upsurge in social media websites has no doubt triggered a huge source of data for mining interesting expressions on a variety of subjects. These expressions on social media websites empower firms and individuals to discover varied interpretations regarding the opinions expressed. In Sub-Saharan Africa, financial institutions are making the needed technological investments required to remain competitive in today's challenging global business environment.

Twitter as one of the digital communication tools has in recent times been integrated into the marketing communication tools of banks to augment the free flow of information. In this light, the author of this thesis conducted sentiment analysis on a large dataset of tweets associated with the Ecobank Group, a prominent pan-African bank in sub-Saharan Africa using four different sentiment lexicons to determine the best lexicon based on its performance.

The results of Valence Aware Dictionary and sEntiment Reasoner (VADER) outperform all the other three lexicons based on accuracy and computational efficiency. Additionally, a word cloud is generated to visually examine the terms in the positive and negative sentiment categories based on VADER. This approach demonstrates that in today's world of empowered customers, firms need to focus on customer engagement to enhance customer experience via social media channels (e.g., Twitter) since the meaning of competitive advantage has shifted from purely competing over price and product to building loyalty and trust. In theory, the study contributes to broadening the scope of online banking given the interplay of consumer sentiments via the social media channel.

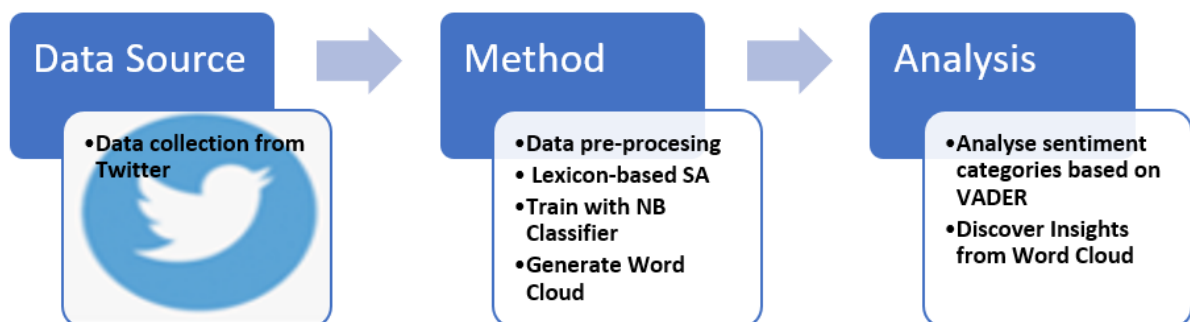


Fig. 6.1: Three-phase methodology deployed [47]

The methodology described in chapter 4 was adopted and implemented in three phases as shown in Fig. 6.1 below. A detailed description of these processes can be found in [47]. The dataset used for the study consists of 7,730 English tweets collected between January 1, 2015, and December 31, 2019, using the “*ecobank*” keyword.

The study sought to answer three research questions (RQs):

- Which lexicon produces the best output that describes the opinions of bank clients on Twitter?
- What insights can be gained from the expressions garnered from the Ecobank group's Twitter handle?
- Since Fintech has come to stay, what lessons can the bank learn from social media to improve its service delivery?

Table 6.1 and Table 6.2 shows the classification of the Tweets annotated by the various lexicons and some sample tweets respectively.

Table 6.1 Classification of Tweets by various lexicons

LEXICON	CLASSIFICATION		
	POSITIVE	NEGATIVE	NEUTRAL
VADER [15]	4317	1233	2180
AFINN [44]	4038	1567	2125
SentiWordNet [13]	2982	3148	1600
TextBlob [46]	3400	1159	3171

In response to *research question 1*, a sample of 773 tweets constituting 10% of the tweets is drawn from the dataset and manually labelled using two independent annotators from the A.I.Lab research group at Faculty of Applied Informatics, Tomas Bata University in Zlin. Out of 773 tweets that were hand-labelled, 525 were classified as positive with 46 and 202 categorized as negative and neutral respectively. The inter-annotator agreement [103] (Cohen’s Kappa) [48] reached 0.75 which indicates a reasonable agreement level. Cohen's Kappa is a statistic that is used to gauge how closely two raters agree on the classification of items into mutually exclusive groups. To select the best lexicon, the labelled tweets of each lexicon were trained using the Naive Bayes classifier to obtain the respective classification scores. The classification accuracy (AC) score is defined in equation (46).

From Table 6.3, the VADER lexicon outperforms all the other lexicons used in this study. This indeed confirms what is stated in the literature [15] since it is accustomed to sentiments expressed on social media. Fig. 6.2 comprises 1,233 tweets expressing negative sentiments visualized as a word cloud after removing stop words. Hence, one can deduce that the tweets contained discussions

regarding frustrations (pain points see Fig. 6.3) customers face concerning the usage of the e-banking solutions provided by the bank.

Table 6.2 Sample Tweets from the Ecobank Tweets dataset [47]

<p><b>Sofiat Lawal</b> @iamsophielawal · Dec 30, 2019 @ecobank_nigeria Thanks for the 20% for buying airtime with *326#. ❤️❤️</p> <p><b>Ecobank Nigeria</b> @ecobank_nigeria · Dec 30, 2019 Hello, we appreciate your feedback. Keep transacting and enjoy exciting rewards with us. Thank you.</p> <p><b>Sofiat Lawal</b> @iamsophielawal Replying to @ecobank_nigeria You're welcome, 🙏</p>	<p><b>Blitz-ng</b> @blitz_ng Ecobank organizes Business Bootcamp for 100 SME's in Africa <a href="https://blitznewsng.wordpress.com/2019/12/04/eco...">blitznewsng.wordpress.com/2019/12/04/eco...</a></p> 
<p><b>Ecobank Group</b> @GroupEcobank · Jun 29, 2019 Are you using Africa's favourite way to send money home? #Rapidtransfer lets you send money to a loved one wherever they are in 33 countries. #Ecobank makes cross-border remittances quick, convenient and affordable. Download the Ecobank Mobile and Rapidtransfer apps today.</p> 	<p><b>HayWhy</b> @Lam2Lio · Dec 30, 2019 @ecobank_nigeria @ecobankhelp_ng please may I know when the issue with the Mobile App will be resolved? Does this affect online banking too?</p> <p><b>Ecobank Nigeria</b> @ecobank_nigeria · Dec 30, 2019 Hello @Lam2Lio, we acknowledge your message and sincerely empathize with you regarding your experience. Our mobile app is currently up and running....1/2</p> <p><a href="#">Send us a private message</a></p>
<p><b>MrKerryMartin</b> @MrKerryMartin · Dec 3, 2019 I need a new bank, my current one have disappointed me enough times. Which is the best Bank 🏦 to be in?</p> <p><b>G</b> @Th3George Replying to @MrKerryMartin EcoBank is the best</p> <p>11:57 AM · Dec 4, 2019 · Twitter for Android</p>	<p><b>Ecobank Group</b> @GroupEcobank · Dec 1, 2018 With almost 20 million people living with HIV in Africa, the epidemic remains high on the health agenda. On #WorldAIDSDay, Ecobank is committed to fight the stigma that still surrounds the disease on the continent. #KnowYourStatus</p> 
<p><b>Ecobank Group</b> @GroupEcobank · Aug 31, 2018 #Mobile money app Nala from #Tanzania was crowned as overall winner of the Ecobank Fintech Challenge securing the top prize of US\$10,000 #EcobankFintech #FinancialInclusion</p> 	<p><b>Ecobank Group</b> @GroupEcobank · Apr 26, 2019 The first winner of the Sustainability Award is Ecobank Ghana, which trained instructors from the Ministry of Education on financial literacy and inclusion. The trainers are now empowering over 500,000 ghanaians, including women traders and cocoa farmers.</p> 
<p><b>Myra Rae</b> @miraACE @ecobank_nigeria what exactly have i done to you? All i did was commend your service and I am suffering like a thief on the cross.</p> <p>2:59 PM · Dec 4, 2019 · Twitter for Android</p> <p><b>Ecobank Nigeria</b> @ecobank_nigeria · Dec 4, 2019 Replying to @miraACE Hello @miraACE, we acknowledge your tweet and empathize with you regarding your experience. Please engage us via DM and share the exact challenge encountered so we can assist adequately. Thank you.</p> <p><a href="#">Send us a private message</a></p>	<p><b>Ecobank Group</b> @GroupEcobank · Apr 29, 2019 @GroupEcobank continued to put great emphasis on promoting women's empowerment in 2018. 44% of Ecobank's workforce is female &amp; 30% are in management positions. 2 experienced female non-executive directors were recently admitted. Senegal's Aichatou Agne Pouye &amp; Nigeria's @aoteh.</p> 
<p><b>Ecobank Group</b> @GroupEcobank · Apr 29, 2019 @GroupEcobank continued to put great emphasis on promoting women's empowerment in 2018. 44% of Ecobank's workforce is female &amp; 30% are in management positions. 2 experienced female non-executive directors were recently admitted. Senegal's Aichatou Agne Pouye &amp; Nigeria's @aoteh.</p>	<p><b>Ecobank Group</b> @GroupEcobank · Apr 29, 2019 @GroupEcobank continued to put great emphasis on promoting women's empowerment in 2018. 44% of Ecobank's workforce is female &amp; 30% are in management positions. 2 experienced female non-executive directors were recently admitted. Senegal's Aichatou Agne Pouye &amp; Nigeria's @aoteh.</p>



Kotzias et. al., in the year 2015. Fig. 6.6 shows an n-dimensional view of the dataset reduced to 2 dimensions using principal component analysis (PCA).

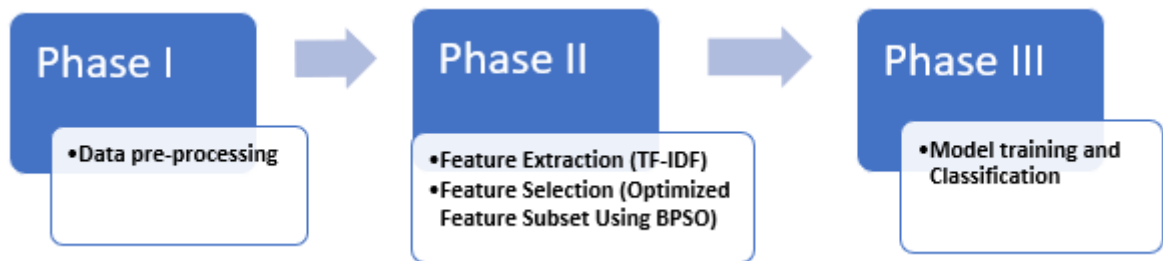


Fig. 6.4: Conceptual framework of the study

The data preprocessing stage comprises data cleaning, tokenization, and stemming. To this end, the TF-IDF feature extraction technique was employed to generate a feature matrix (see Table 6.4). In the BPSO for FS, each particle denotes one candidate solution having a dimension represented by a vector with values 0 or 1 (see Fig. 6.5). A value of 1 at position  $k^{th}$  means the  $k^{th}$  feature has been selected and 0 otherwise.

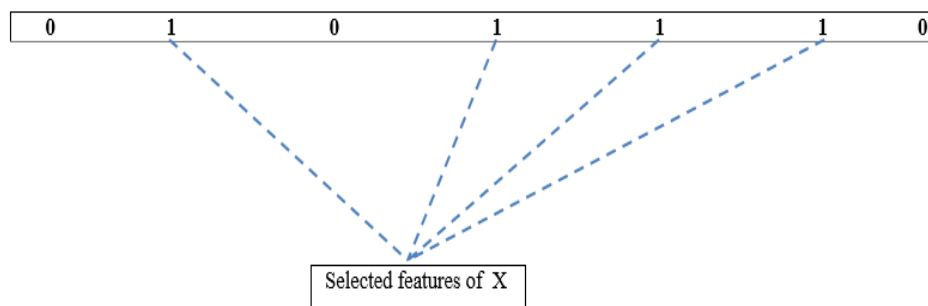


Fig. 6.5: Particle Swarm Representation

The optimal feature subset is produced during the feature selection phase by utilizing the BPSO algorithm for the TF-IDF feature matrix. The optimized feature subset is then trained followed by sentiment classification. Guided by the enhancement of the classifier performance objective, the objective/fitness function utilized by the authors of [53] was adopted in this work.

Table 6.4 TF-IDF scores for some pre-processed text

bar	batteri	case	day	great	work
0.0368	0.2207	0.1103	0.0736	0.2943	0.0736

Table 6.5 Model accuracy scores

Method	Accuracy Score (%)	BPSO-based (%)	Accuracy Gain (%)
k-NN	68.66	69.57	1.0
NB	3.67	85.27	11.6
SVM	78.67	87.10	8.43

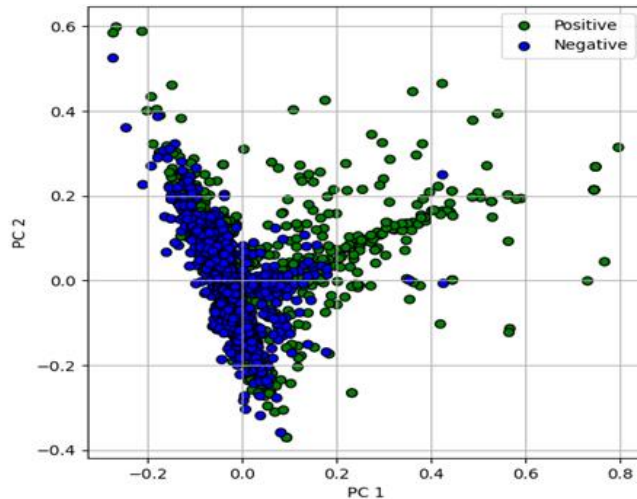


Fig. 6.6: Two-dimensional view of data using PCA

### 6.2.1 EXPERIMENTAL SETUP

The study was conducted using Python 3.8 running on a Windows 10 Professional 64bit operating system machine furnished with an Intel Core i7 processor and 8 Gigabyte memory.

The ML classifiers (k-NN, NB, SVM) were used as a baseline whereas the proposed text feature selection method was designed and implemented using open-source Python libraries available via sklearn [54] and pyswarms [55]. The experiment is assessed using the accuracy evaluation metric shown in equation (46).

### 6.2.2 RESULTS REALIZED

The results of the text-based feature selection for sentiment classification using the optimized and non-optimized techniques on the *sentiments labelled sentences dataset* are presented in Table 6.5. From Table 6.5, the best score with BPSO was realized using SVM followed by the NB and k-NN on the sentiment labelled dataset. A graphical illustration of the results depicted in Table 6.5 is shown in Fig. 6.7. In terms of accuracy gain (i.e., the difference between BPSO-based accuracy score and baseline accuracy score), NB recorded the highest followed by SVM and the k-NN.

In conclusion, this work demonstrates the significance of text feature selection for sentiment classification from a metaheuristic perspective with the view to enhancing the classifier accuracy. The results of the evaluation with and without the BPSO on the baseline models prove the superiority of the metaheuristic approach in text feature selection. In the future, other metaheuristic algorithms can be explored within the framework of multimodal sentiment analysis.

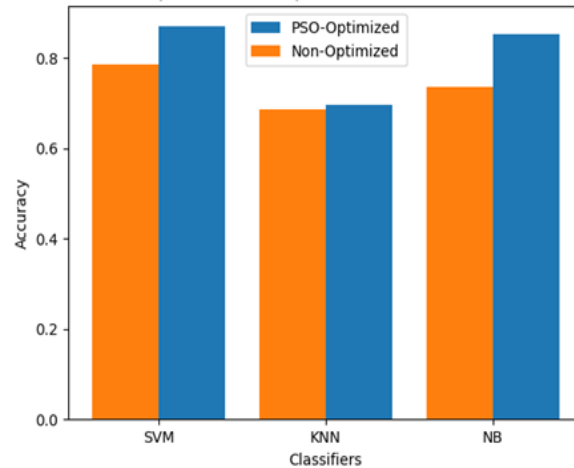


Fig. 6.7: Optimized/Non-Optimised plot using SVM, k-NN, and NB

### 6.3 HYBRID BIO-INSPIRED FEATURE SELECTION USING ANGLE MODULATION

In this section, the grey wolf optimization algorithm (GWO) is combined with the Angle Modulated PSO (AMPSO) to create a new hybrid Angle Modulated GWO and PSO (AMGWOPSO) for wrapper feature selection with the k-NN classifier given the k-NN's implementation simplicity and popularity in hybrid metaheuristic literature [109].

#### 6.3.1 PARAMETERS UTILIZED

The parameters utilized for this work are shown in Table 6.6. As stated earlier, this study aims to decrease the number of selected attributes while enhancing the accuracy of classification. Besides, the AMGWOPSO is utilized as a wrapper for FS with a k-NN classifier to produce the optimum results.

By using trial and error, the value of  $k = 5$  (in k-NN) is obtained. Moreover, the  $k$  value of 5 and population size of 10 was adopted because it produces the best results across all datasets. K-fold cross-validation, as described in [40], was used to divide the datasets into  $K$  segments. The experimental findings were duly compared with some related work methods indicated below:

- Angle modulated particle swarm optimizer (AMPSO) [30]
- Binary particle swarm optimizer [39]
- Binary whale optimization algorithm [72]



- WOASAT-2 [37]

Moreover, the author of this thesis downloaded and implemented the native Python code of the algorithms used for comparison from github [67]. Furthermore, the necessary modifications were made using helper functions to support the implementation of the proposed techniques.

### 6.3.2 METRICS FOR EVALUATION

To make certain that the experimental outcomes are stable and statistically relevant, the partitioned data is repeated 30 times with the following statistical metrics acquired from the validation data in each run.

Table 6.6 Parameter settings

Parameter	Value
Agents pop.	10
Maximum num of Iter.	200
Problem scope	No. features in dataset
Search domain	[0,1]
No. of repeated exec.	30
k-NN classifier	K=5
$\alpha, \beta$ in fitness function	0.99,0.01
AMGWOPSO accel. Coeff.	$c_1 = c_2 = c_3 = 0.5$
AMGWOPSO inert. weight	$0.5 + \text{rand}()/2$
BPSO	[8]

The Average Classification Accuracy, Average Feature Selection Size, Mean Fitness Function, Best Fitness Function, Worst Fitness function, and Average Computational Time metrics were adopted to compare the proposed and state-of-the-art algorithms [40].

- **Classification Average Accuracy (CAA):** measures the classifier accuracy of the selected feature set after  $R$  executions of the algorithm. The  $CAA$  is formulated mathematically in equation (40):

$$CAA = \frac{1}{R} \sum_{k=1}^R CA^k \quad (40)$$

where  $CA^k$  refers to the accuracy obtained at the  $k - th$  run.

- **Mean Selection Size:** formulated mathematically in equation (41) measures the average number of attributes selected to the overall size of attributes after  $R$  executions of the algorithm.

$$AvgSS = \frac{1}{R} \sum_{k=1}^R \frac{FS^k}{M} \quad (41)$$

where  $FS^k$  refers to the number of attributes chosen at the  $k - th$  execution, and  $M$  connotes the overall number of attributes in the dataset.

- **Mean Fitness:** refers to the fitness function's average value that is produced after  $R$  executions of the algorithm. It can be computed using equation (42) as follows:

$$MF = \frac{1}{R} \sum_{k=1}^R g_k^* \quad (42)$$

$g_k^*$  : mean fitness score achieved during the  $k - th$  run.

- **Best Fitness:** describes the fitness function's minimal value following  $R$  executions of the algorithm.

$$BF = \min_k g_k^* \quad (43)$$

$g_k^*$  : best fitness score recorded during the  $k - th$  run.

- **Worst Fitness:** represents the fitness function's maximum value following  $R$  executions of the algorithm.

$$WF = \max_k g_k^* \quad (44)$$

$g_k^*$  : worst fitness score recorded during the  $k - th$  run.

- **Mean Execution Time:** represents the average computing time (in seconds) following  $R$  executions of the algorithm. Usually computed using equation (45):

$$MET = \frac{1}{R} \sum_{k=1}^R CT^k \quad (45)$$

$CT^k$ : computation time value obtained at the  $k - th$  run.

- **The Accuracy metric (AC)** (46) depends on the True Positives (TP), False Positives (FP), True Negatives (TN), and False Negatives (FN) [19].

$$AC = \frac{TP + TN}{TP + TN + FP + FN} \quad (46)$$

### 6.3.3 EXPERIMENTAL RESULTS AND DISCUSSION

The outputs of the tests performed to evaluate the effectiveness of the proposed AMGWOPSO are presented and discussed in this section. Additionally, several different work-related contemporary methods are chosen for comparison.

### 6.3.4 PROPOSED AMGWOPSO

A summary of results obtained by the authors' proposed hybrid AMGWOPSO is shown in Table 6.7. By utilizing the proposed approach, the CongressEW and M-of-n datasets from the UCI repository obtained the highest average classification score of 100%. Following closely are the WineEW, Exactly, Zoo, Breastcancer, KrVsKpEW, and Vote datasets with average accuracies of 99.9%, 98.4%, 98.3%, 97.9%, 97.5%, and 97.4% respectively. Furthermore, the BreastEW, IonosphereEW, SonarEW, PenglungEW, and Lymphography datasets each obtained an average accuracy score of 96.5%, 95.8%, 95.6%, 93.8%, and 91.8% respectively.

In the feature selection tests, the Exactly2, Breastcancer, Vote, HeartEW, and Zoo datasets recorded the most reduced number of attributes with 2.4, 3.8, 4.2, 4.6,

Table 6.7 Summary results of proposed AMGWOPSO compared to other related state-of-the-art algorithms.

Dataset	Mean Accuracy	Mean Fitness	Selected Features	Computational times(s)
Breastcancer	0.979	0.020	3.80	12.01
BreastEW	0.965	0.032	12.50	12.76
CongressEW	1.000	0.000	5.60	15.77
Exactly	0.984	0.012	6.20	15.96
Exactly2	0.764	0.007	2.40	14.24
HeartEW	0.868	0.010	4.60	15.37
IonosphereEW	0.958	0.042	13.20	11.49
KrVsKpEW	0.975	0.022	19.30	15.26
Lymphography	0.918	0.020	5.60	12.69
M-of-n	1.000	0.000	5.80	13.15
PenglungEW	0.938	0.064	132.20	14.22
SonarEW	0.956	0.032	28.20	15.38
SpectEW	0.891	0.006	8.80	14.39
Tic-tac-toe	0.805	0.010	5.60	16.44
Vote	0.974	0.020	4.20	15.41
WaveformEW	0.812	0.010	16.20	44.22
WineEW	0.986	0.020	6.80	15.67
Zoo	0.983	0.020	5.20	15.64
<b>Average</b>	<b>0.931</b>	0.019	15.900	16.115

Table 6.8. Average classification and features results of proposed AMGWPSO compared to other related state-of-the-art algorithms.

Dataset	Average Accuracy					Average Features Selected				
	AMGWOPSO	AMPSO	BPSO	BWOA	WOASAT-2	AMGWOPSO	AMPSO	BPSO	BWOA	WOASAT-2
Breastcancer	<b>0.979</b>	0.967	0.967	0.957	0.970	<b>3.80</b>	6.20	5.70	6.38	4.20
BreastEW	0.965	0.926	0.935	0.955	<b>0.980</b>	12.50	16.80	16.60	23.80	<b>11.60</b>
CongressEW	<b>1.000</b>	0.934	0.938	0.929	0.980	<b>5.60</b>	7.20	6.80	10.20	6.40
Exactly	0.984	0.694	0.684	0.758	<b>1.000</b>	6.20	7.00	9.80	9.20	<b>6.00</b>
Exactly2	<b>0.764</b>	0.746	0.752	0.698	0.750	<b>2.40</b>	5.30	6.20	4.78	2.80
HeartEW	<b>0.868</b>	0.810	0.778	0.763	0.850	<b>4.60</b>	4.90	7.90	9.40	5.40
IonosphereEW	0.958	0.839	0.837	0.890	<b>0.960</b>	13.20	17.80	19.20	22.40	<b>12.80</b>
KrVsKpEW	0.975	0.952	0.958	0.915	<b>0.980</b>	19.30	19.80	20.80	24.20	<b>18.40</b>

Lymphography	<b>0.918</b>	0.678	0.689	0.786	0.890	<b>5.60</b>	6.80	9.00	10.80	7.20
M-of-n	<b>1.000</b>	0.852	0.857	0.854	<b>1.000</b>	<b>5.80</b>	6.60	9.10	6.02	6.00
PenglungEW	0.938	0.708	0.719	0.729	<b>0.940</b>	132.20	179.20	178.80	188.60	<b>127.40</b>
SonarEW	0.956	0.738	0.741	0.854	<b>0.970</b>	28.20	34.20	32.20	46.20	<b>26.40</b>
SpectEW	<b>0.891</b>	0.767	0.769	0.788	0.880	<b>8.80</b>	9.20	12.50	9.40	9.40
Tic-tac-toe	<b>0.805</b>	0.742	0.731	0.751	0.790	<b>5.60</b>	6.20	6.60	8.40	6.00
Vote	<b>0.974</b>	0.919	0.889	0.939	0.970	<b>4.20</b>	4.80	8.80	9.40	5.20
WaveformEW	<b>0.812</b>	0.748	0.758	0.713	0.760	<b>16.20</b>	23.20	22.70	33.60	20.60
WineEW	0.986	0.946	0.946	0.928	<b>0.990</b>	6.80	10.60	8.40	7.38	<b>6.40</b>
Zoo	<b>0.983</b>	0.824	0.830	0.965	0.970	<b>5.20</b>	6.20	9.70	8.80	5.60
<b>Average</b>	<b>0.931</b>	0.822	0.821	0.843	0.920	<b>15.90</b>	20.67	21.66	24.39	15.99

and 5.2 respectively. The average number of features reduced is the average of the features selected after  $R$  runs for each dataset as described in the metrics for evaluation section. Hence, the final value (float in this case) depends on whether the selected feature for each run is either odd or even [66]. Moreover, the least execution times (in seconds) achieved after running the proposed method for 200 iterations for each dataset are as follows: IonosphereEW (11.49), Breastcancer (12.01), Lymphography (12.69), and BreastEW (12.76). Overall, the average classification accuracy (93.1%), fitness (0.019), selected attributes (15.9), and execution time (16.115 in seconds) were achieved using the proposed AMGWPSO method on the 18 benchmark datasets.

Thus, I can conclude that the impressive and competitive results obtained demonstrate an improvement in the exploitation and exploration abilities of the proposed hybrid AMGWPSO by combining the GWO and AMPSO.

Table 6.9 Mean fitness results of proposed AMGWPSO compared to other related state-of-the-art algorithms.

Dataset	AMGWPSO	AMPSO	BPSO	BWOA	WOASAT-2
Breastcancer	<b>0.020</b>	0.034	0.030	0.035	0.040
BreastEW	0.032	0.032	<b>0.030</b>	0.035	<b>0.030</b>
CongressEW	<b>0.000</b>	0.034	0.040	0.042	0.030
Exactly	0.012	0.278	0.280	0.186	<b>0.010</b>
Exactly2	<b>0.007</b>	0.240	0.250	0.210	0.250
HeartEW	<b>0.010</b>	0.152	0.150	0.192	0.160
IonosphereEW	0.042	0.148	0.140	0.122	<b>0.040</b>
KrVsKpEW	0.022	0.055	0.050	0.067	<b>0.020</b>
Lymphography	<b>0.020</b>	0.195	0.190	0.152	0.110
M-of-n	<b>0.000</b>	0.232	0.110	0.081	0.010
PenglungEW	0.064	0.126	0.220	0.137	<b>0.060</b>
SonarEW	0.032	0.131	0.130	0.147	<b>0.030</b>
SpectEW	<b>0.006</b>	0.132	0.130	0.148	0.130
Tic-tac-toe	<b>0.010</b>	0.247	0.240	0.238	0.210
Vote	<b>0.020</b>	0.049	0.050	0.057	0.040
WaveformEW	<b>0.010</b>	0.242	0.220	0.257	0.250
WineEW	0.020	0.024	0.020	0.051	<b>0.010</b>
Zoo	<b>0.020</b>	0.166	0.100	0.048	0.040
<b>Total</b>	<b>0.347</b>	<b>2.517</b>	<b>2.380</b>	<b>2.204</b>	<b>1.470</b>

### 6.3.5 COMPARING THE PROPOSED HYBRID AMGWOPSO WITH OTHER METHODS

The findings of the proposed AMGWOPSO are compared in this section to some contemporary methods available in metaheuristic feature selection scholarly works. The average classification (see Fig. 6.8 to Fig. 6.25 ) and the number of attributes reduced (see Fig. 6.26 to Fig. 6.43) using the proposed metaheuristic algorithm based on the chosen 18 benchmark datasets are presented in Table 6.8.

Similarly, the numerical statistical results of the AMPSO, BPSO, BWOA, and the hybrid WOASAT-2 metaheuristic algorithms used for comparisons are also presented in the same table. As the outcomes illustrate, the proposed hybrid AMGWOPSO obtained the finest average accuracy results on 11 out of the 18 datasets followed by the hybrid WOASAT-2 (8 out of 18 datasets), BWOA, and BPSO.

Furthermore, the performance of the different contemporary metaheuristic methods with reference to the average features reduced over all the datasets is also outlined in Table 6.8. I can state that the impressive performance of the proposed hybrid AMGWOPSO is asserted on 11 out of the 18 UCI datasets in respect of the average number of attributes reduced as depicted. Again, the suggested hybrid method obtains the best reduction (regarding features selected) compared to the native AMPSO, BPSO, and BWOA across all the 18 benchmark UCI machine learning (ML) datasets.

However, the hybrid WOASAT-2 follows closely by obtaining the best-reduced features across 7 out of the 18 benchmark UCI ML datasets with the AMPSO, BPSO, and BWOA concluding the list. Table 6.9 and Table 6.10 outline the mean, best, and worst fitness function scores based on different execution times (see Fig. 6.44 to Fig. 6.61) averaged over all the 18 benchmark UCI machine learning datasets using the fitness function defined in equation (27).

It can be seen from Table 6.9 and Table 6.10 that the proposed hybrid AMGWOPSO shows enhanced performance with reference to mean fitness function on 11 out of the 18 datasets, whereas, within the best fitness function results, AMGWOPSO surpasses the other contemporary FS metaheuristic techniques on 11 out of the 18 benchmark datasets when compared to AMPSO, BPSO, BGOA, and the hybrid WOASAT-2.

Table 6.10 Best and Worst fitness results of proposed AMGWOPSO compared to other related state-of-the-art methods.

Dataset	Best					Worst				
	AMGWOPSO	AMPSO	BPSO	BWOA	WOASAT-2	AMGWOPSO	AMPSO	BPSO	BWOA	WOASAT-2
Breastcancer	<b>0.010</b>	0.024	0.030	0.034	0.030	0.030	<b>0.044</b>	0.030	0.039	0.040
BreastEW	0.022	0.022	<b>0.020</b>	0.028	<b>0.020</b>	0.042	0.042	0.050	<b>0.065</b>	0.040
CongressEW	<b>0.000</b>	0.024	0.030	0.024	0.020	0.000	0.044	0.040	0.045	<b>0.050</b>
Exactly	<b>0.002</b>	0.268	0.210	0.162	0.010	0.022	0.288	0.320	<b>0.327</b>	0.010
Exactly2	<b>0.007</b>	0.230	0.220	0.180	0.230	0.017	0.250	<b>0.310</b>	0.252	0.270
HeartEW	<b>0.010</b>	0.142	0.130	0.132	0.130	0.020	0.162	0.180	<b>0.198</b>	0.180
IonosphereEW	0.032	0.138	0.120	0.108	<b>0.030</b>	0.052	0.158	<b>0.170</b>	0.144	0.050

KrVsKpEW	<b>0.012</b>	0.045	0.030	0.044	0.020	0.032	0.065	0.070	<b>0.120</b>	0.020
Lymphography	<b>0.010</b>	0.185	0.140	0.132	0.090	0.030	0.205	<b>0.270</b>	0.157	0.140
M-of-n	<b>0.000</b>	0.222	0.060	0.048	0.010	0.000	<b>0.242</b>	0.160	0.181	0.010
PenglungEW	0.054	0.116	0.130	0.093	<b>0.030</b>	0.074	0.136	<b>0.290</b>	0.232	0.110
SonarEW	0.022	0.121	0.070	0.114	<b>0.010</b>	0.042	0.141	<b>0.220</b>	0.189	0.050
SpectEW	<b>0.006</b>	0.122	0.100	0.129	0.110	0.016	0.142	0.160	<b>0.166</b>	0.150
Tic-tac-toe	<b>0.010</b>	0.237	0.210	0.160	0.200	0.020	0.257	<b>0.270</b>	0.242	0.230
Vote	<b>0.010</b>	0.039	0.030	0.057	<b>0.020</b>	0.030	0.059	<b>0.080</b>	0.067	0.040
WaveformEW	<b>0.010</b>	0.232	0.162	0.128	0.230	0.020	0.252	0.230	<b>0.276</b>	0.260
WineEW	0.010	0.014	<b>0.000</b>	0.037	<b>0.000</b>	0.030	0.034	0.030	<b>0.057</b>	0.030
Zoo	0.010	0.156	0.030	0.038	<b>0.000</b>	0.030	0.176	<b>0.210</b>	0.105	0.100
Total	<b>0.337</b>	2.337	1.722	1.649	1.190	<b>0.507</b>	2.697	3.090	2.862	1.780

Additionally, for all 18 datasets, the proposed hybrid AMGWOPSO is not inferior in terms of performance (worst solution) to any of the other approaches used in this study. Once more, these impressive and competitive results highlight the capabilities of the proposed hybrid AMGWOPSO in efficiently striking a healthy balance between exploitation and exploration as it hybridizes the AMPSO with GWO.

Besides, a comparison of computation time (in seconds) as shown in Table 6.12 between the hybrid WOASAT-2 and the proposed hybrid AMGWOPSO demonstrates its amazing performance. Whereas the hybrid WOASAT-2 takes 2820.21 seconds to run all the datasets, the proposed method takes 290.07 seconds to run across all datasets in its entirety, demonstrating that AMGWOPSO is more capable of providing better solutions within a rational execution time. This in essence is largely attributable to the few parameters utilized by both the GWO and AMPSO and the enhanced strength of the resulting hybrid AMGWOPSO algorithm.

### 6.3.6 STATISTICAL SIGNIFICANCE ANALYSES

To evaluate the statistical relevance of the variations in the mean fitness values acquired by the AMGWOPSO and the other contemporary optimizers, the Wilcoxon signed-rank test [58] was utilized. The test aims to determine whether the findings of the two methods are independent.

Table 6.11: Wilcoxon signed-rank test for mean fitness evaluation metric.

Evaluation metric	Comparison	p-Value	Hypothesis	Significant difference
Mean Fitness	AMGWOPSO vs AMPSO	2.9E-04	Reject $h_0$ at 5%	Yes
	AMGWOPSO vs BPSO	3.5E-04	Reject $h_0$ at 5%	Yes
	AMGWOPSO vs BWOA	7.6E-06	Reject $h_0$ at 5%	Yes
	AMGWOPSO vs WOASAT2	1.04E-02	Reject $h_0$ at 5%	Yes

To formulate the null hypothesis, the author assumes that there is no significant variation in the mean fitness scores of the AMGWOPSO and the other optimizers.

The null hypothesis is accepted when the significance level is greater than 5% implying no significant improvement when the proposed hybrid AMGWOPSO

was used and vice versa. Table 6.11 depicts the scores of the Wilcoxon signed-ranked statistical test after adopting the analytical procedures from the authors in [69].

The author can remark that the enhancement achieved by the proposed hybrid AMGWOPSO was substantial when compared to the other optimizers (i.e., AMPSO, BPSO, BWOA, hybrid WOASAT-2) given that all the p-values obtained are lower than 5%. This means that there are statistically significant differences between the mean fitness obtained by the AMGWOPSO and the other methods. Consequently, the null hypothesis suggesting that there is no significant variation in the mean fitness scores of the AMGWOPSO and the other optimizers at 5% significant level is duly rejected.

### 6.3.7 COMPUTATIONAL COMPLEXITY

The FS task is acknowledged as an NP-hard problem with a huge combinatorial search space [70]. When the number of attributes grows, the number of potential solutions in the search space increases exponentially.

Table 6.12 Execution time (in seconds) of proposed AMGWOPSO compared with the hybrid WOASAT-2

Dataset	AMGWOPSO	WOASAT-2
Breastcancer	<b>12.01</b>	41.74
BreastEW	<b>12.76</b>	44.3
CongressEW	<b>15.77</b>	35.67
Exactly	<b>15.96</b>	51.79
Exactly2	<b>14.24</b>	54.88
HeartEW	<b>15.37</b>	29.79
IonosphereEW	<b>11.49</b>	30.84
KrVsKpEW	<b>15.26</b>	589.56
Lymphography	<b>12.69</b>	26.17
M-of-n	<b>13.15</b>	51.54
PenglungEW	<b>14.22</b>	30.49
SonarEW	<b>15.38</b>	27.76
SpectEW	<b>14.39</b>	31.38
Tic-tac-toe	<b>16.44</b>	56.89
Vote	<b>15.41</b>	30.79
WaveformEW	<b>44.22</b>	1633.27
WineEW	<b>15.67</b>	26.33
Zoo	<b>15.64</b>	27.02
<b>Total</b>	<b>290.07</b>	2820.21

Three major processes involved in the AMGWOPSO optimization procedure are solution initialization, fitness function evaluation, and search agent updates. Given that  $n$  denotes the population of search agents, then  $O(n)$  defines the initialization step's computational complexity. Equations modelled in the updating process are used to explore the best positions that ensure the optimal

solution and appropriately update the other solutions' positions. As such,  $O(\mathcal{M} \times n) + O(\mathcal{M} \times n \times \mathcal{B})$  describes the computational complexity for the updating process where  $\mathcal{M}$  describes the iterations number,  $\mathcal{B}$  the boundary of the decision space. Consequently, the computational complexity of the entire optimization exercise engaged in AMGWOPSO is  $O(n \times (\mathcal{M} + \mathcal{M} \times \mathcal{B} + 1))$ .

### 6.3.8 ACHIEVED RESULTS FOR AMGWOPSO

In this study, an efficient and novel angle-modulated metaheuristic method called AMGWOPSO was proposed and used to investigate the FS problem with the dual goal of increasing classification accuracy while diminishing the number of attributes chosen. As indicated earlier, a low-level coevolutionary mixed hybrid approach was adopted in hybridizing the AMPSO with the GWO. By utilizing 18 benchmark UCI machine learning datasets, experiments were administered to assess the effectiveness and efficiency of the proposed hybrid AMGWOPSO.

The overall performance of the proposed technique across all the datasets was compared with the hybrid WOASAT-2, AMPSO, BPSO, and BWOA metaheuristic techniques available in the feature selection literature. With reference to accuracy and the number of attributes reduced, the findings portrayed that the proposed hybrid AMGWOPSO surpasses most metaheuristic algorithms on the majority of classical benchmark datasets utilized in the study.

Furthermore, juxtaposing an execution or computation time analysis of the proposed method with the other metaheuristic techniques used in this work across all the datasets demonstrates that the newly proposed approach is reliable and better placed in providing outstanding solutions within a respectable computation or execution timeframe. Moreover, mean fitness, best fitness, and worst fitness statistical tests were conducted where the impressive and competitive results reaffirmed the proposed AMGWOPSO's ability to effectively guarantee a reasonable balance between exploitation and exploration.

Despite the successes chalked by my novel method, the fixed amplitude of the generating function constitutes a drawback to the proposed approach given that it is a sine wave. In the future, the author will consider modifying the amplitude of the generating function to potentially scale the effect of the vertical shift coefficient. While this work represents the first attempt in introducing the concept of angle modulation into hybrid metaheuristics FS literature as far as the author can tell, this concept can further be experimented with other non-hybrid/hybrid metaheuristic algorithms to assess their efficacy and stability.

Next is a graphical illustration of the classification results achieved, the average number of features reduced and average fitness obtained for all 18 benchmark datasets using the five different metaheuristic algorithms (i.e. Fig. 6.8 to Fig. 6.25, Fig. 6.26 to Fig. 6.43 and Fig. 6.44 to Fig. 6.61).



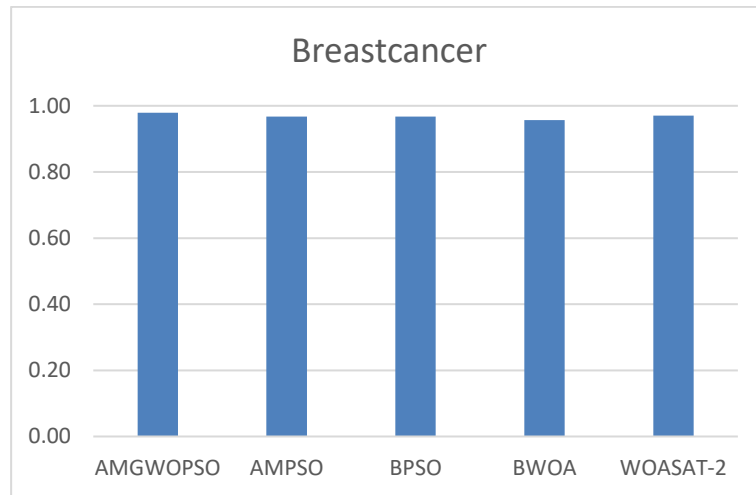


Fig. 6.8: Classification accuracy results of five different MAs on the Breastcancer datasets

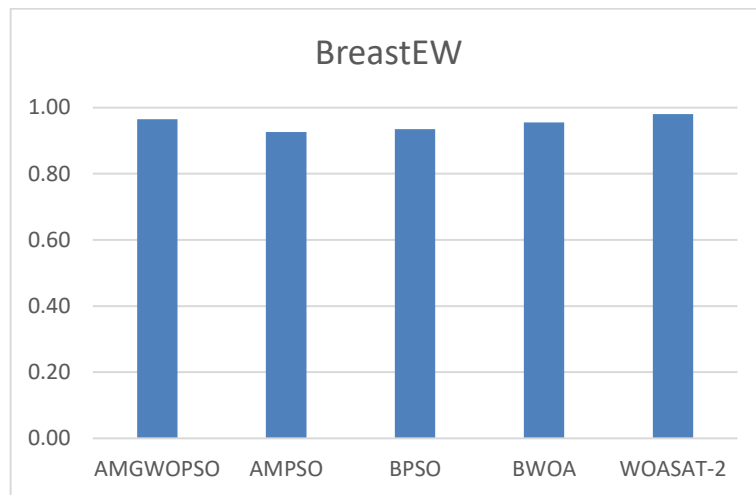


Fig. 6.9: Classification accuracy results of five different MAs on the BreastEW dataset

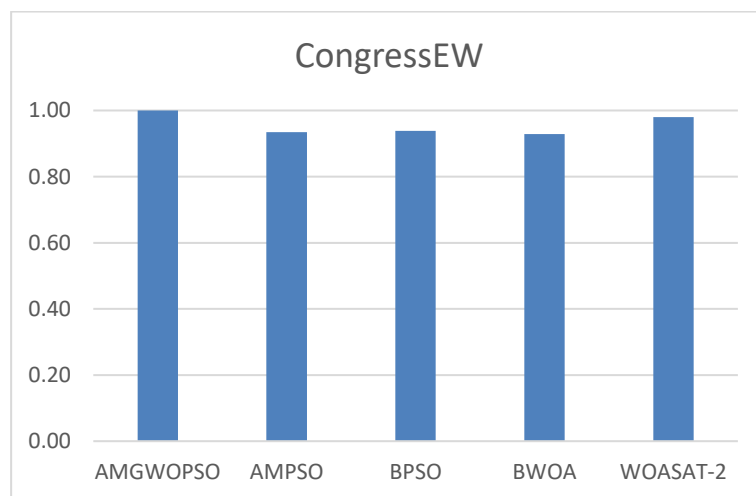


Fig. 6.10: Classification accuracy results of five different MAs on the CongressEW dataset

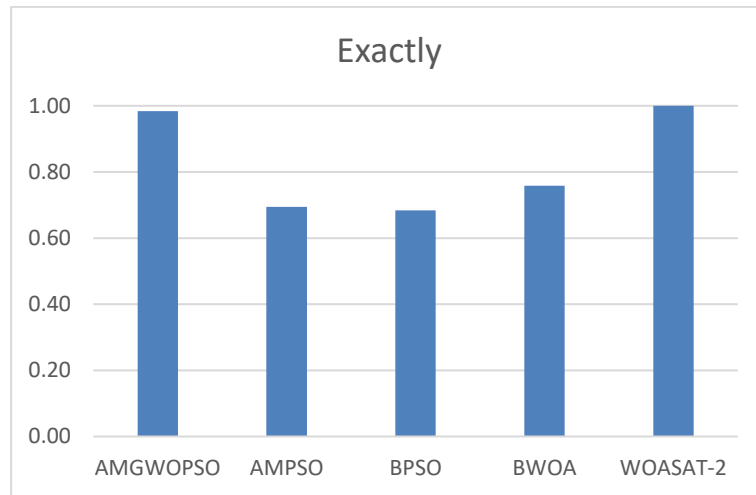


Fig. 6.11: Classification accuracy results of five different MAs on the Exactly dataset

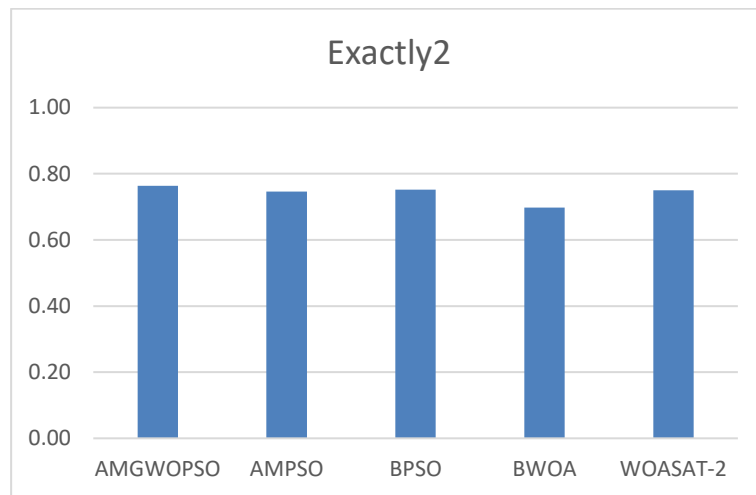


Fig. 6.12: Classification accuracy results of five different MAs on the Exactly2 dataset

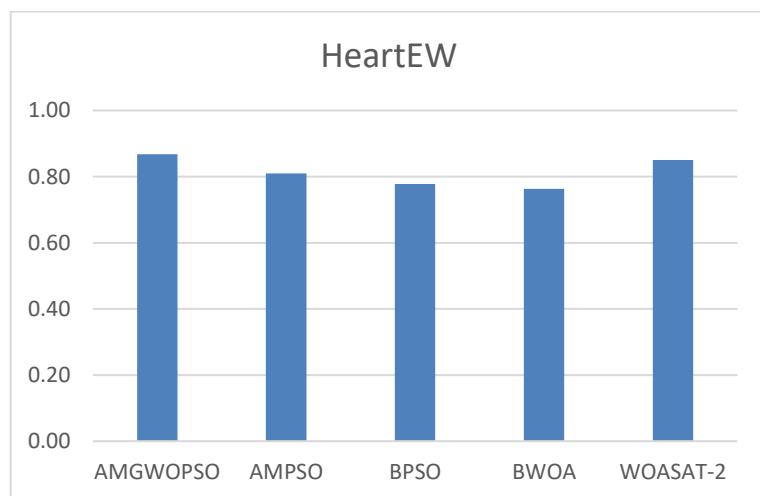


Fig. 6.13: Classification accuracy results of five different MAs on the HeartEW dataset

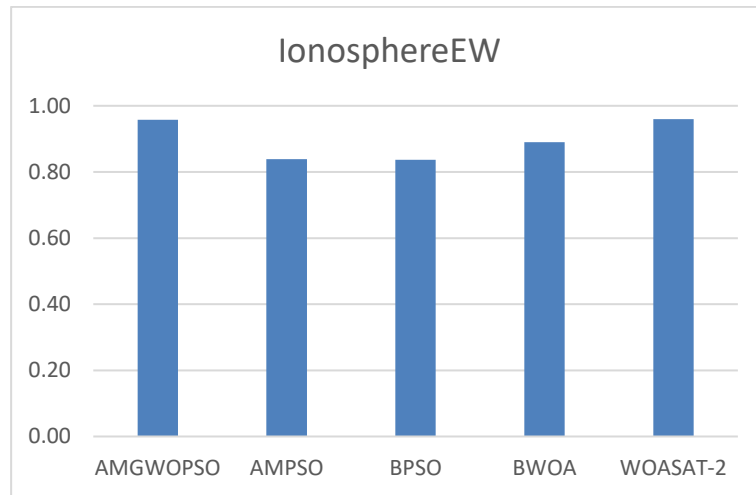


Fig. 6.14: Classification accuracy results of five different MAs on the IonosphereEW dataset

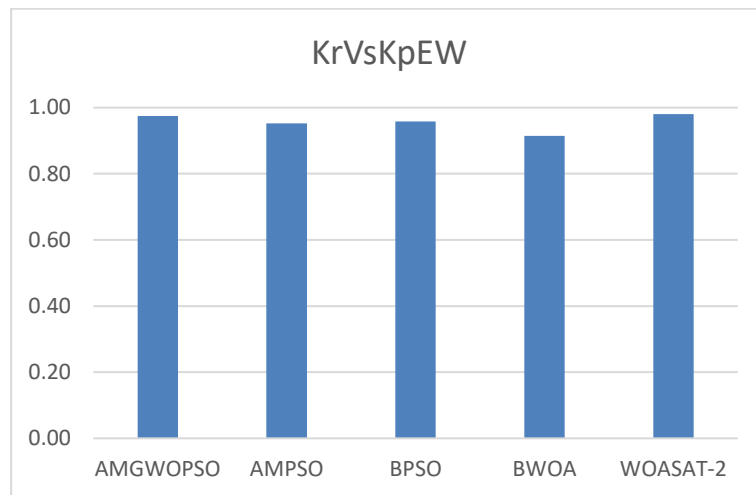


Fig. 6.15: Classification accuracy results of five different MAs on the KrVsKpEW dataset

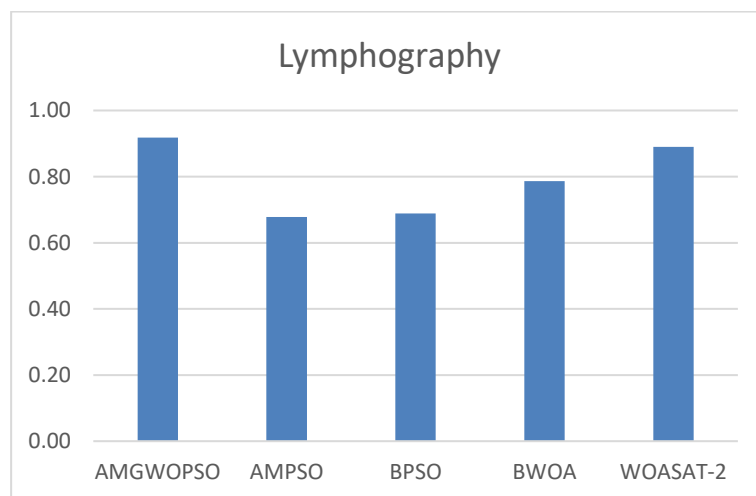


Fig. 6.16: Classification accuracy results of five different MAs on the Lymphography dataset

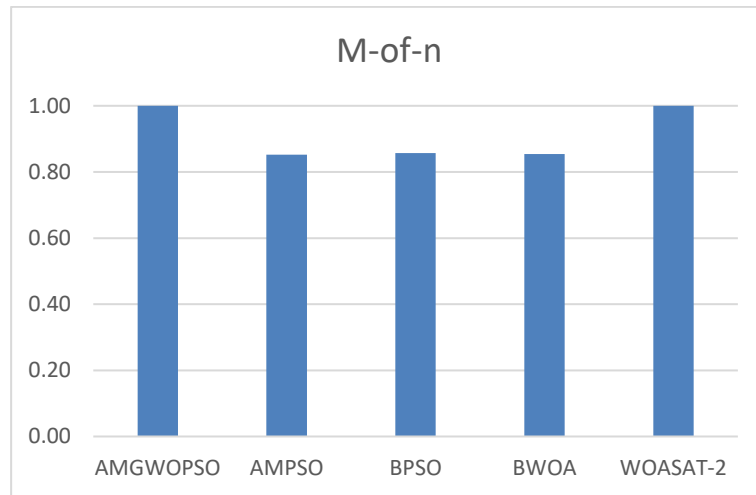


Fig. 6.17: Classification accuracy results of five different MAs on the M-of-n dataset

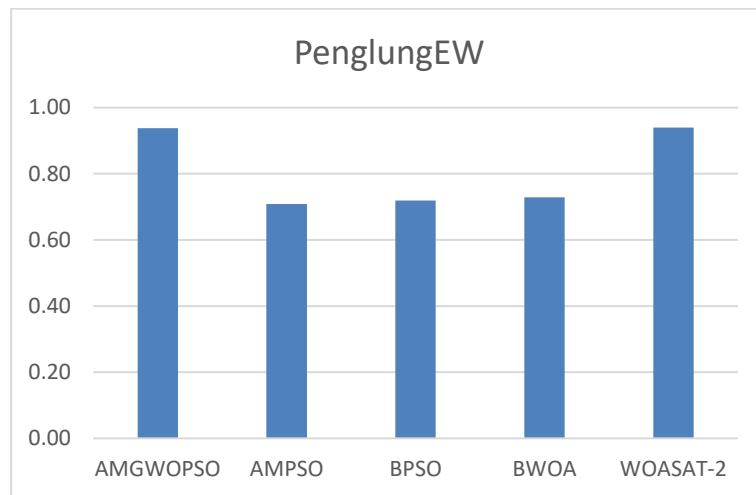


Fig. 6.18: Classification accuracy results of five different MAs on the PenglungEW dataset

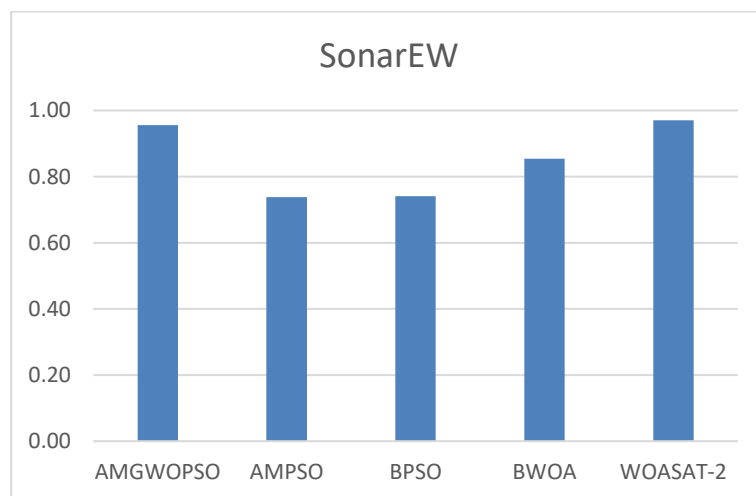


Fig. 6.19: Classification accuracy results of five different MAs on the SonarEW dataset

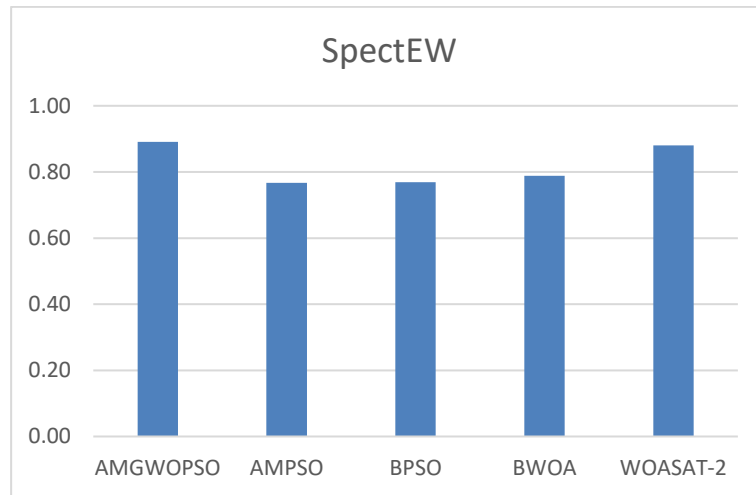


Fig. 6.20: Classification accuracy results of five different MAs on the SpectEW dataset

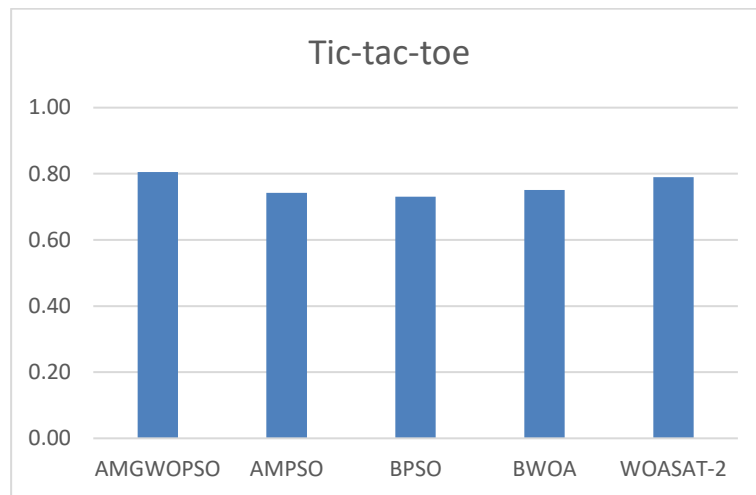


Fig. 6.21: Classification accuracy results of five different MAs on the Tic-tac-toe dataset

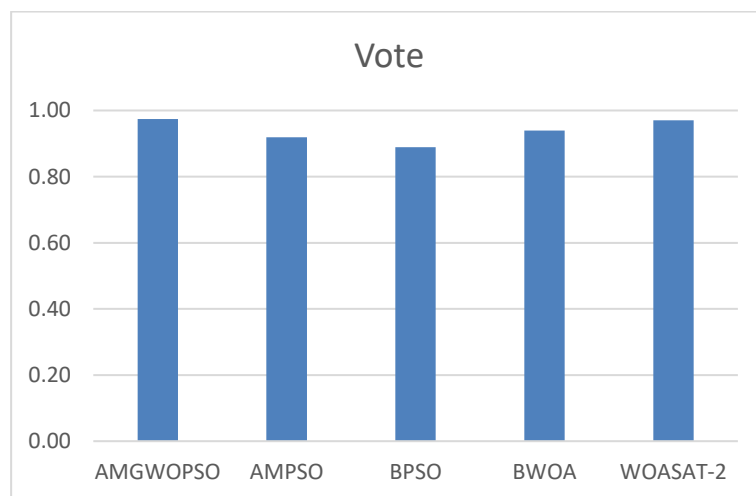


Fig. 6.22: Classification accuracy results of five different MAs on the Vote dataset

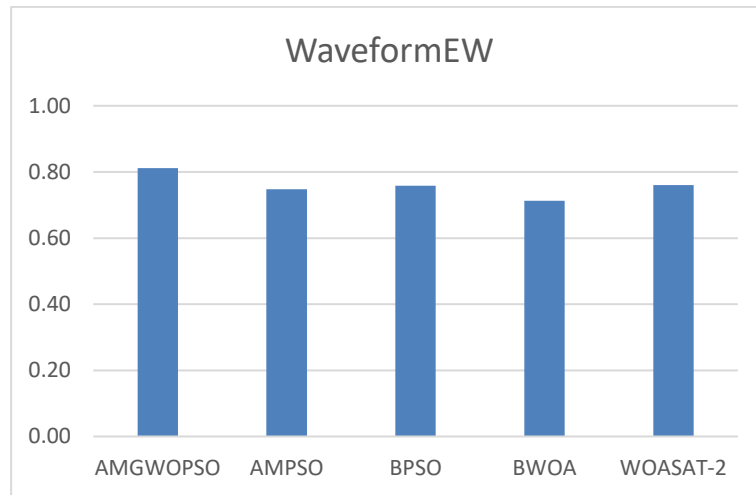


Fig. 6.23: Classification accuracy results of five different MAs on the WaveformEW dataset

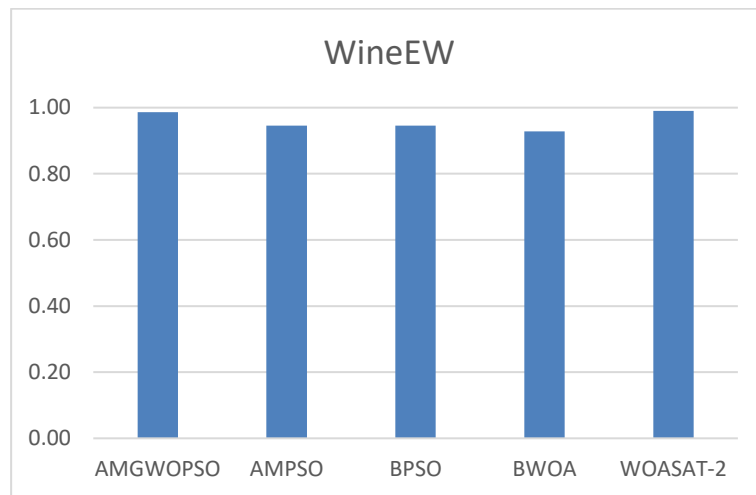


Fig. 6.24: Classification accuracy results of five different MAs on the WineEW dataset

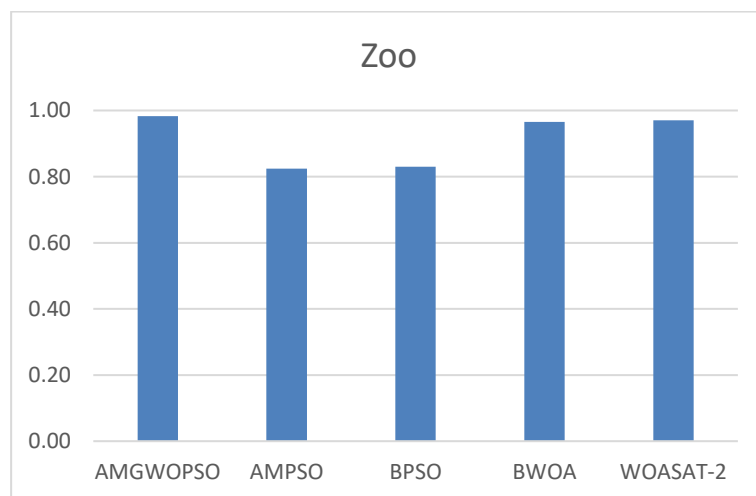


Fig. 6.25: Classification accuracy results of five different MAs on the Zoo dataset

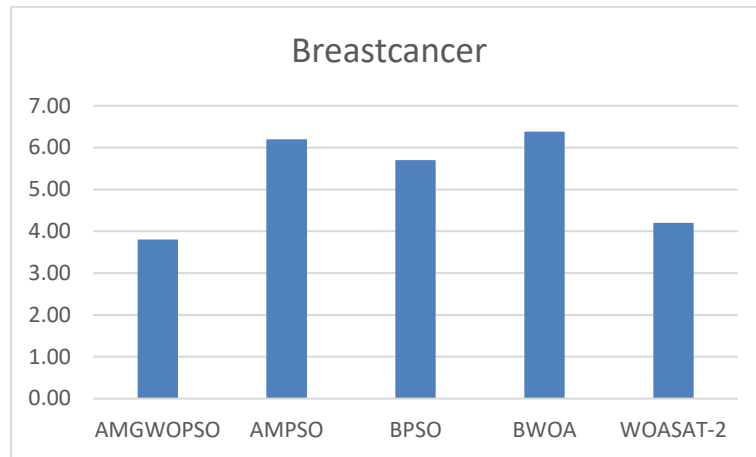


Fig. 6.26: Average features selected using five different MAs on the Breastcancer dataset

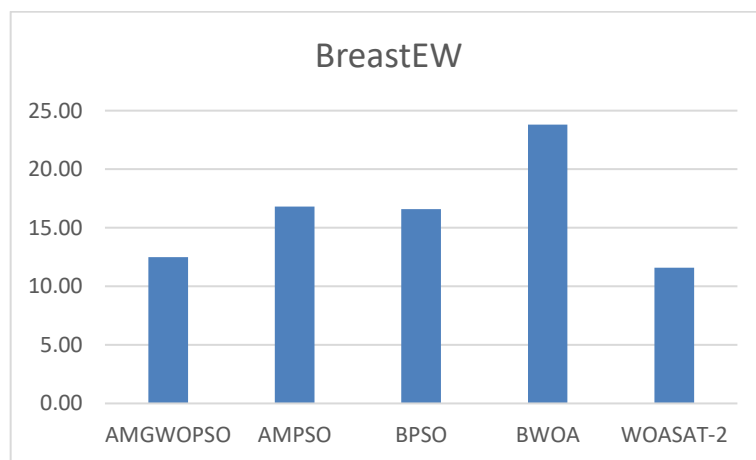


Fig. 6.27: Average features selected using five different MAs on the BreastEW dataset

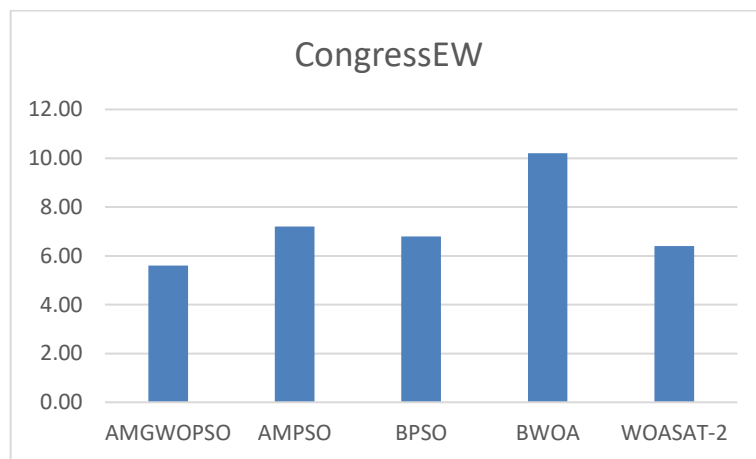


Fig. 6.28: Average features selected using five different MAs on the CongressEW dataset

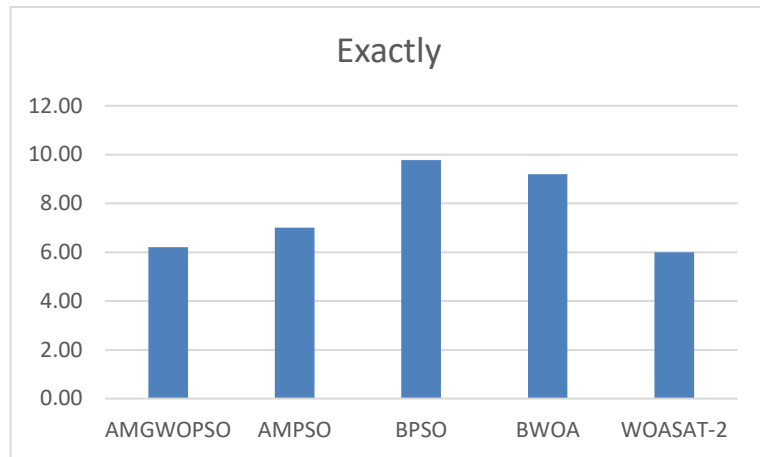


Fig. 6.29: Average features selected using five different MAs on the Exactly dataset

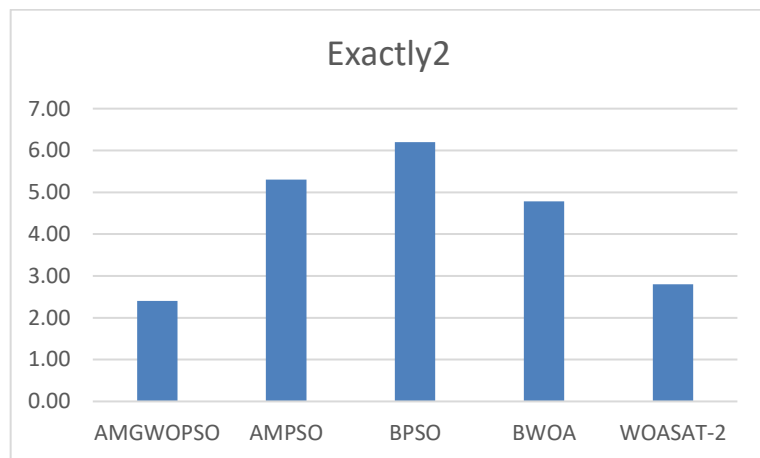


Fig. 6.30: Average features selected using five different MAs on the Exactly2 dataset

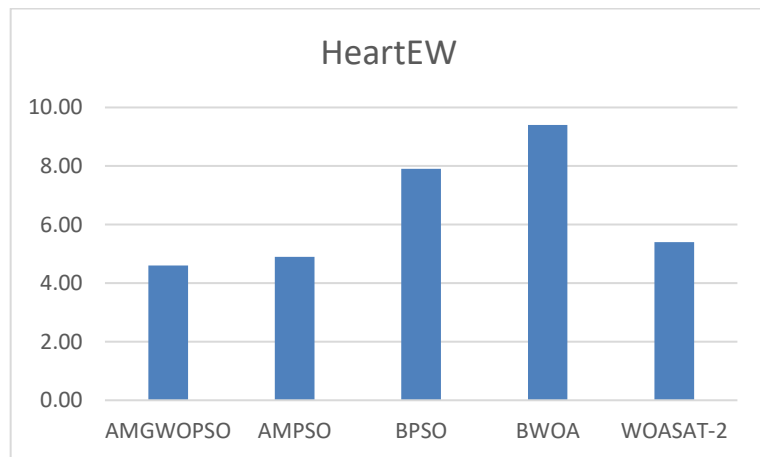


Fig. 6.31: Average features selected using five different MAs on the HeartEW dataset



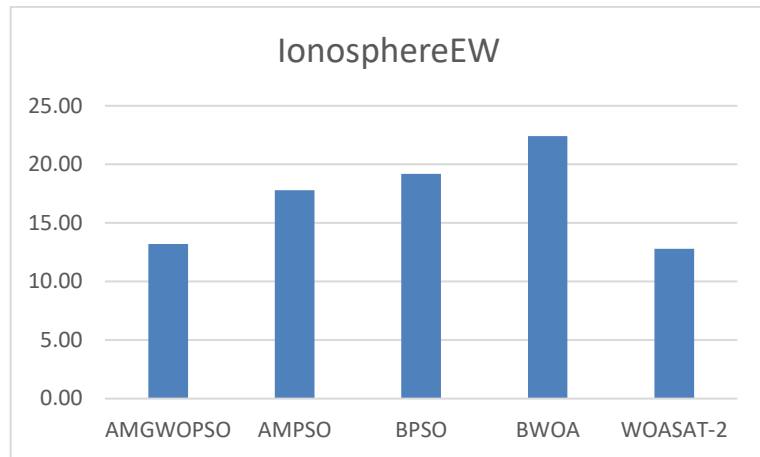


Fig. 6.32: Average features selected using five different MAs on the IonosphereEW dataset

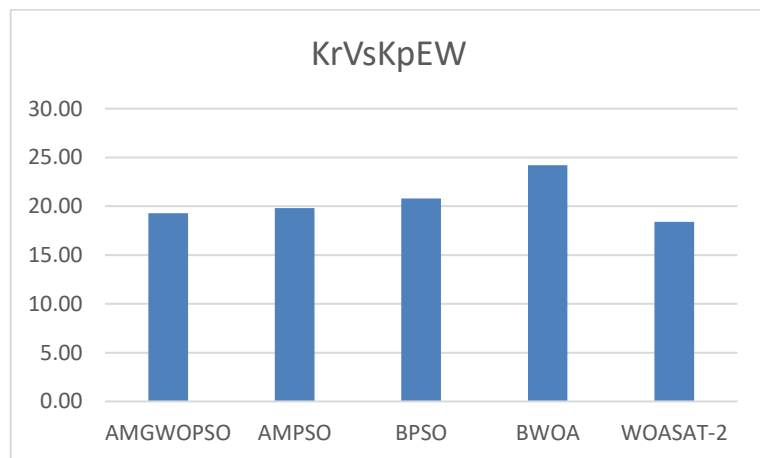


Fig. 6.33: Average features selected using five different MAs on the KrVsKpEW dataset

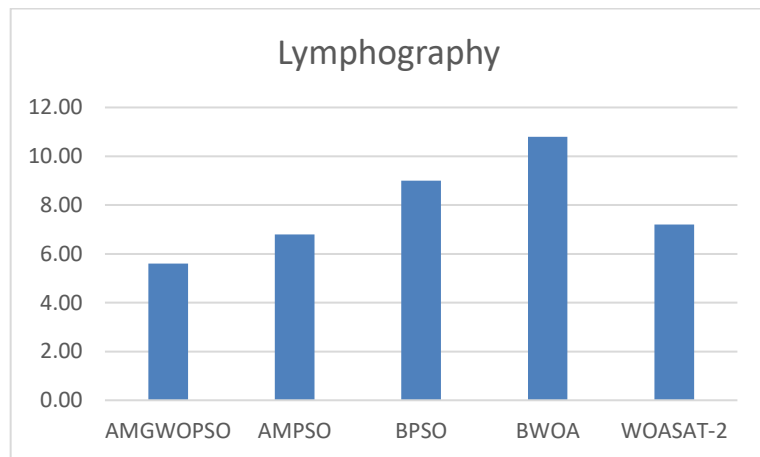


Fig. 6.34: Average features selected using five different MAs on the Lymphography dataset

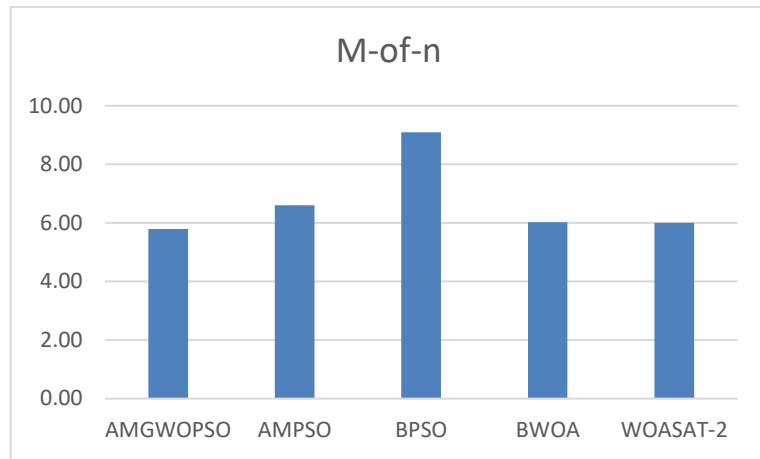


Fig. 6.35: Average features selected using five different MAs on the M-of-n dataset

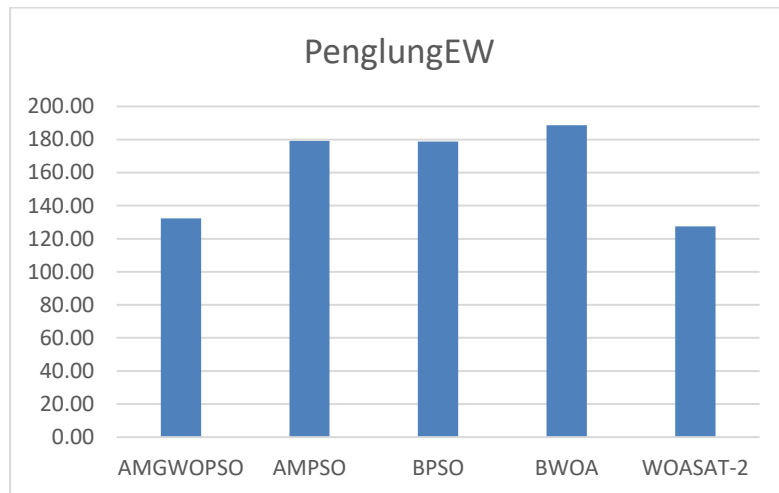


Fig. 6.36: Average features selected using five different MAs on the PenglungEW dataset

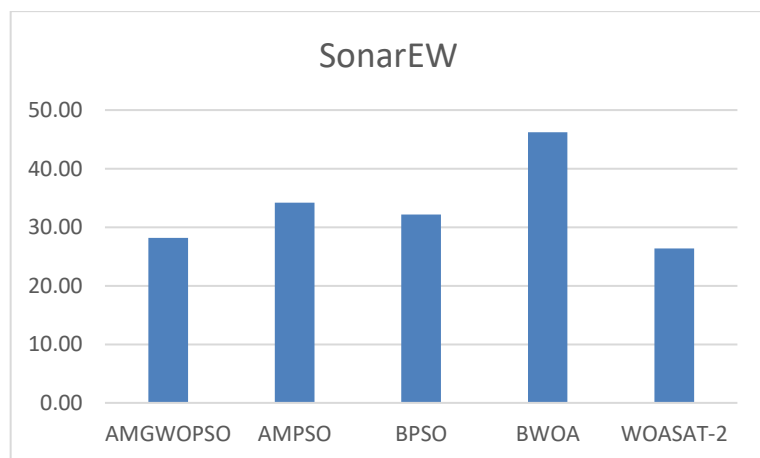


Fig. 6.37: Average features selected using five different MAs on the SonarEW dataset

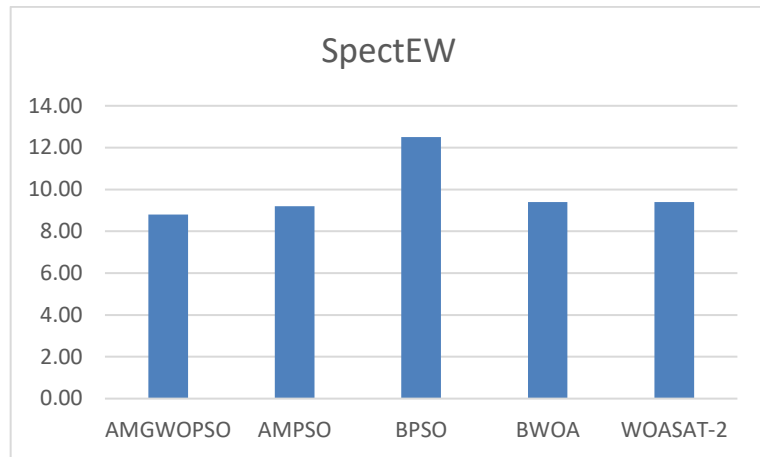


Fig. 6.38: Average features selected using five different MAs on the SpectEW dataset

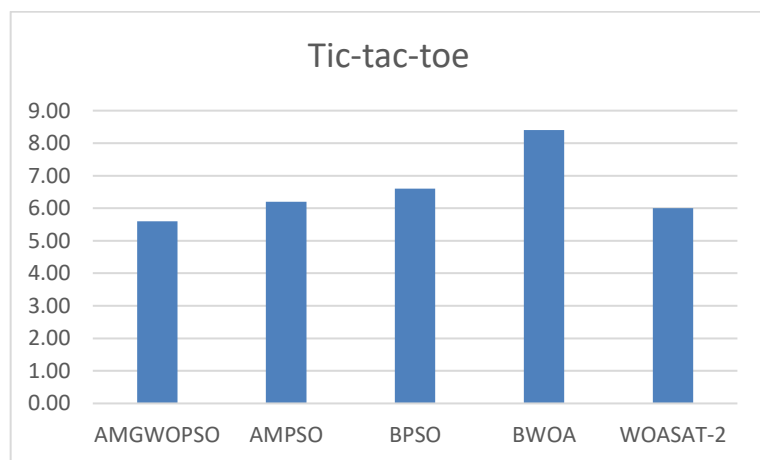


Fig. 6.39: Average features selected using five different MAs on the Tic-tac-toe dataset

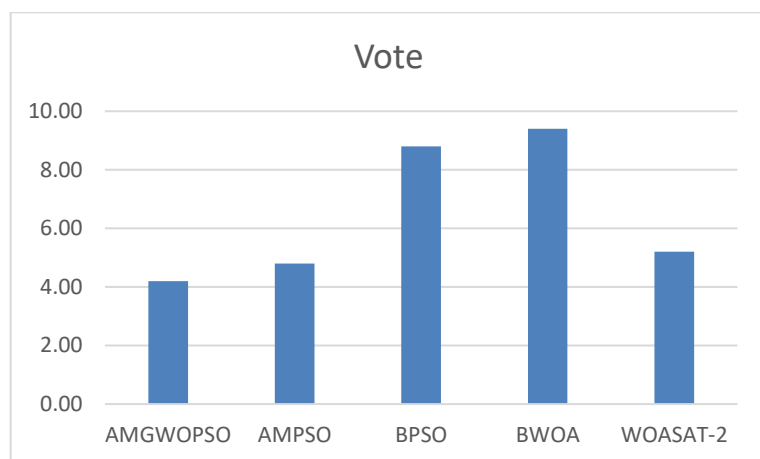


Fig. 6.40 Average features selected using five different MAs on the Vote dataset

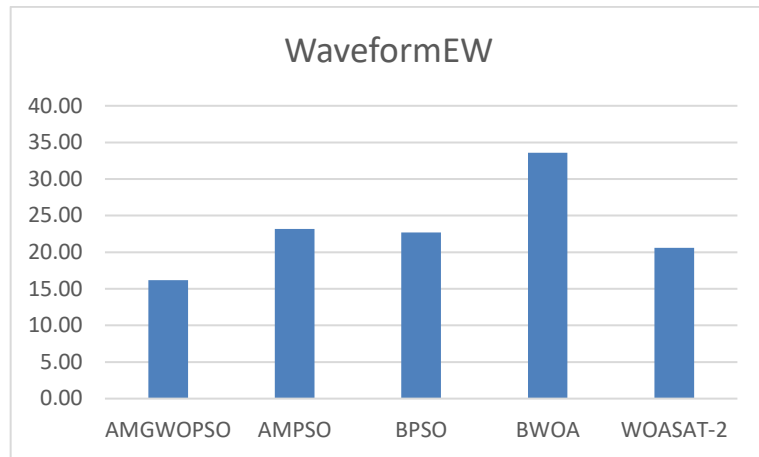


Fig. 6.41 Average features selected using five different MAs on the WaveformEW dataset

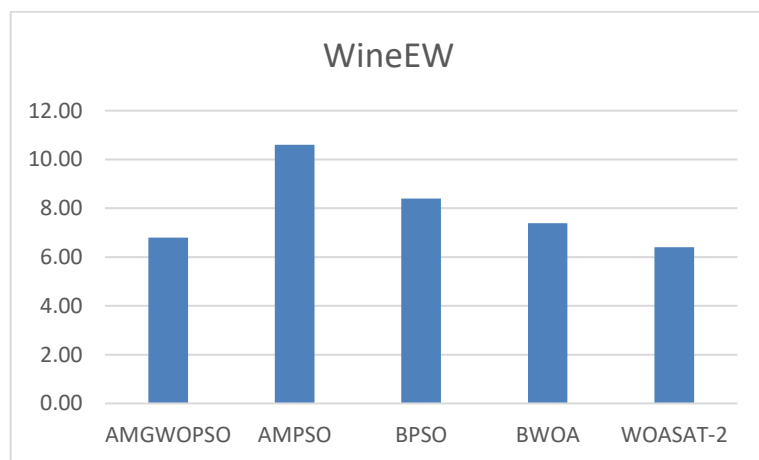


Fig. 6.42: Average features selected using five different MAs on the WineEW dataset

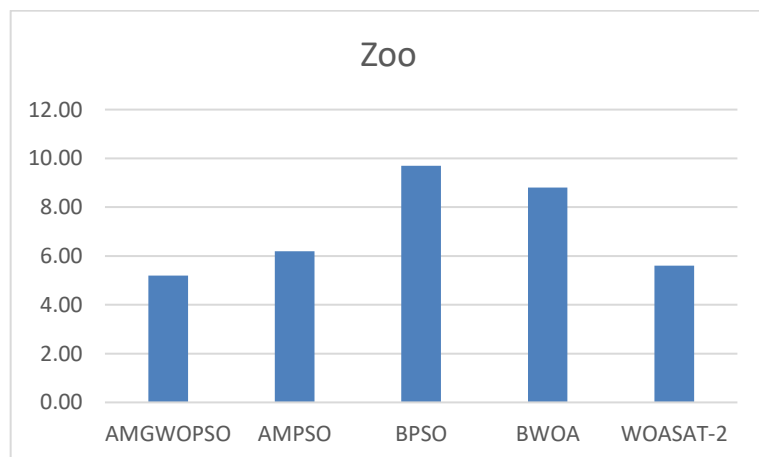


Fig. 6.43: Average features selected using five different MAs on the Zoo dataset

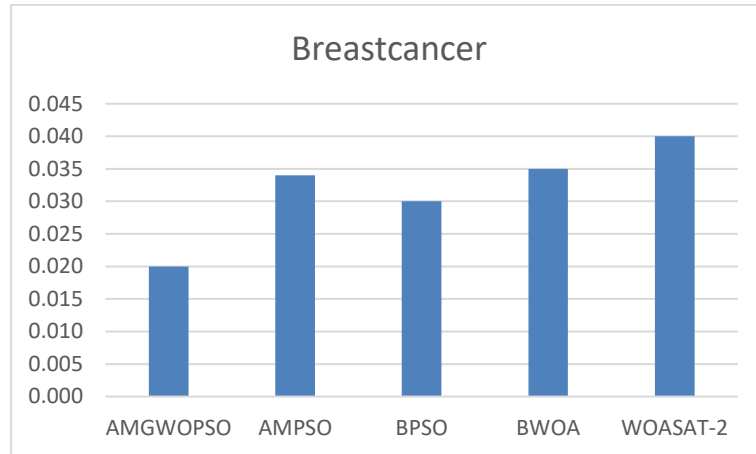


Fig. 6.44: Average fitness obtained for the Breastcancer dataset using five different MAs

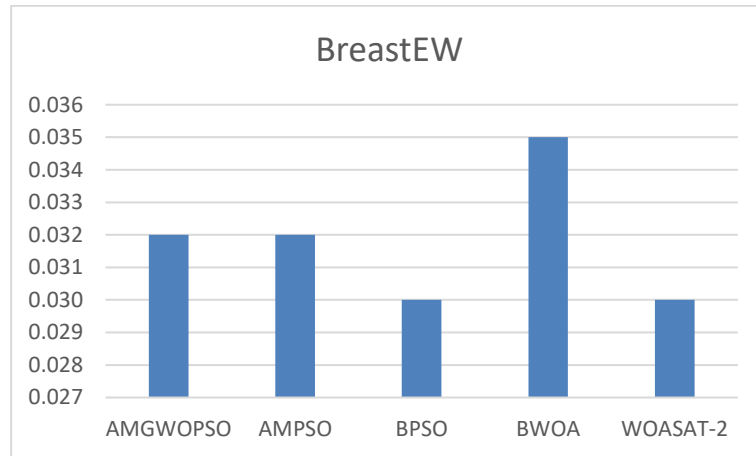


Fig. 6.45: Average fitness obtained for the BreastEW dataset using five different MAs

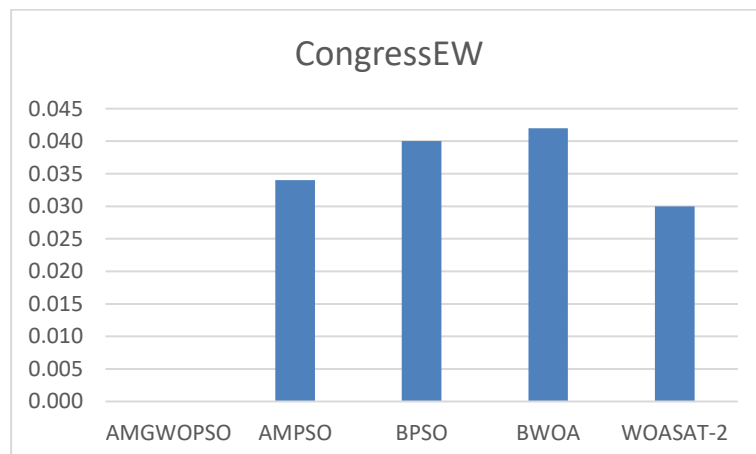


Fig. 6.46: Average fitness obtained for the CongressEW dataset using five different MAs

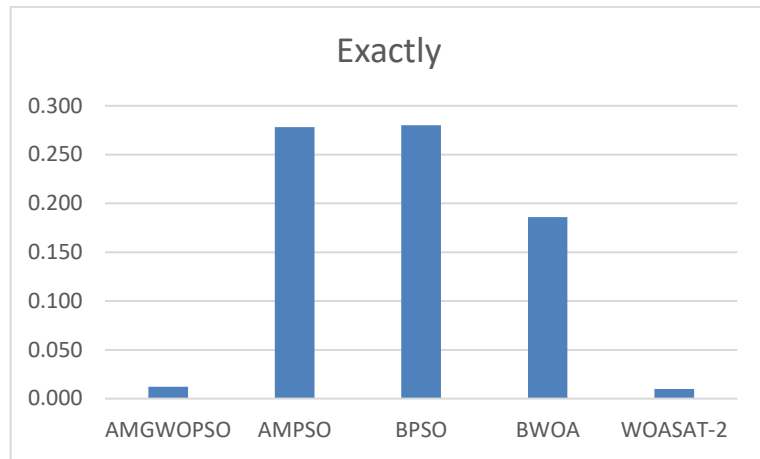


Fig. 6.47: Average fitness obtained for the Exactly dataset using five different MAs

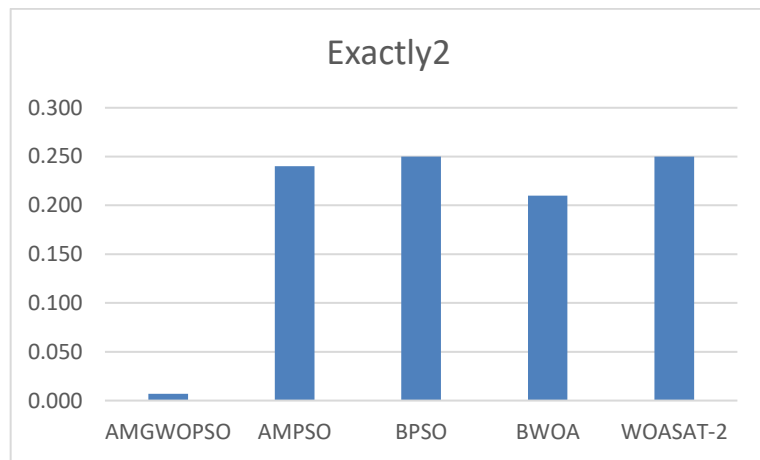


Fig. 6.48: Average fitness obtained for the Exactly2 dataset using five different MAs

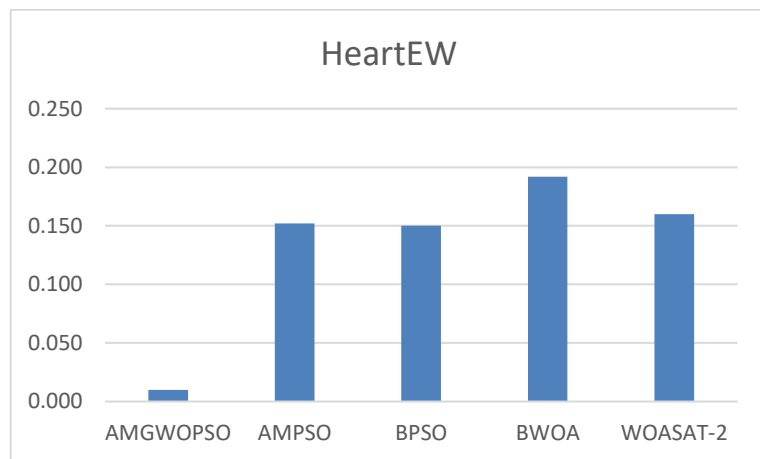


Fig. 6.49: Average fitness obtained for the HeartEW dataset using five different MAs

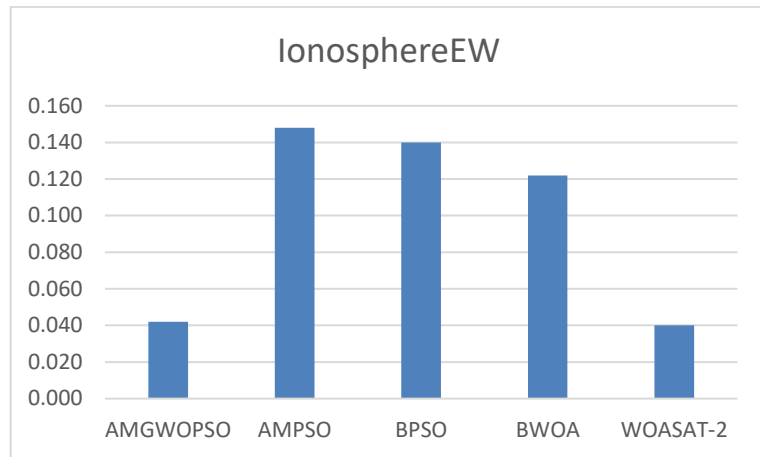


Fig. 6.50: Average fitness obtained for the IonosphereEW dataset using five different MAs

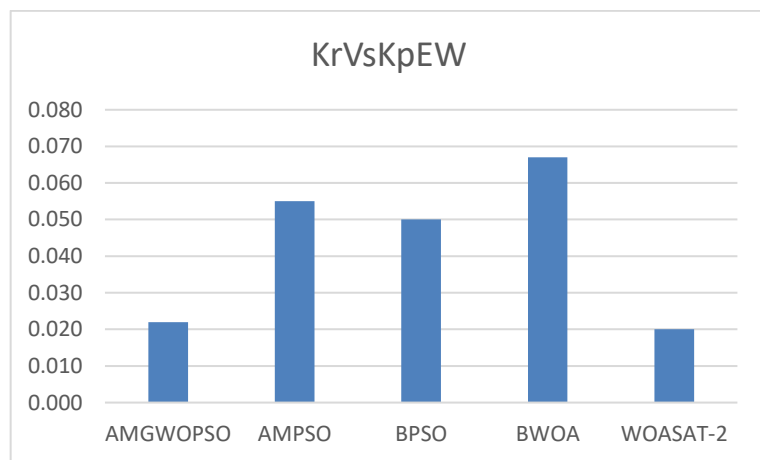


Fig. 6.51: Average fitness obtained for the KrVsKpEW dataset using five different MAs

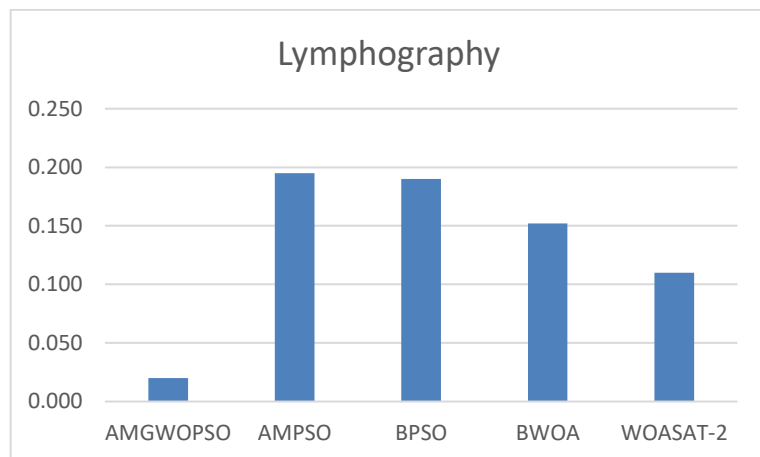


Fig. 6.52: Average fitness obtained for the Lymphography dataset using five different MAs

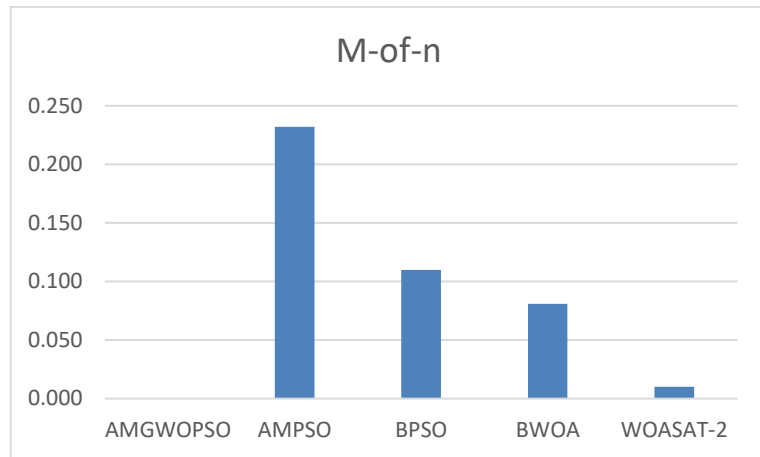


Fig. 6.53: Average fitness obtained for the M-of-n dataset using five different MAs

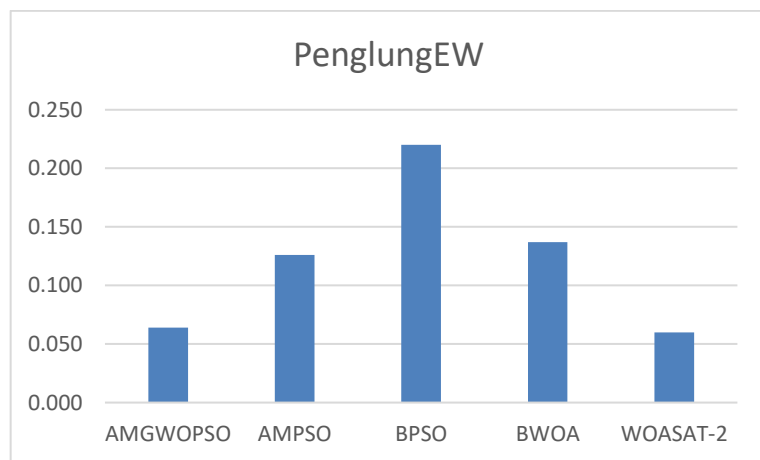


Fig. 6.54: Average fitness obtained for the PenglungEW dataset using five different MAs

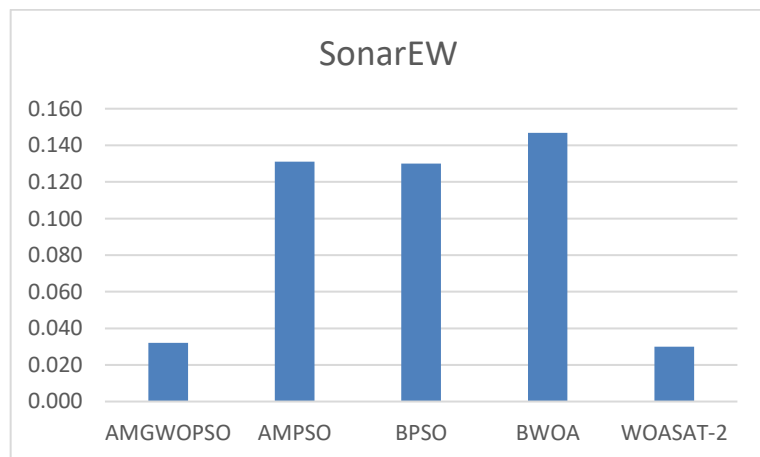


Fig. 6.55: Average fitness obtained for the SonarEW dataset using five different MAs



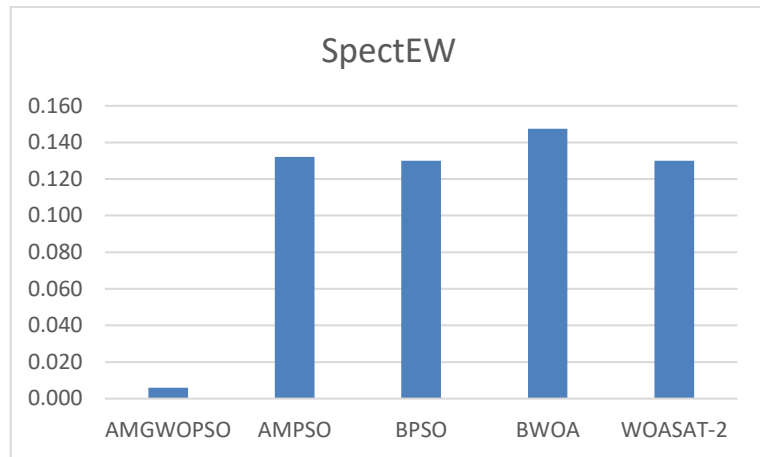


Fig. 6.56: Average fitness obtained for the SpectEW dataset using five different MAs

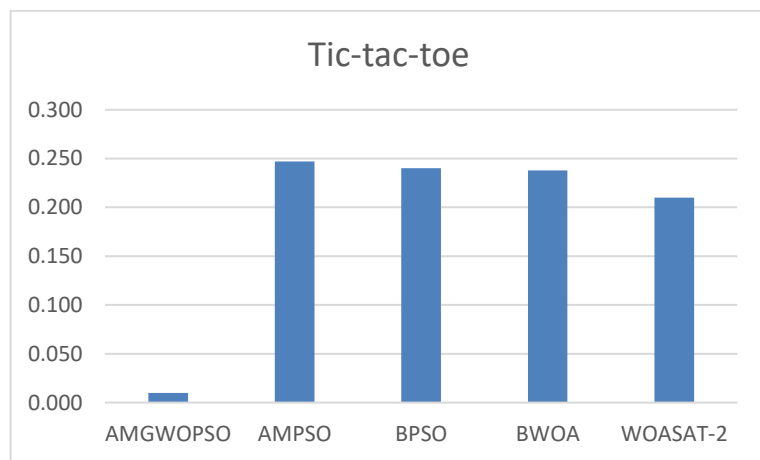


Fig. 6.57: Average fitness obtained for the Tic-tac-toe dataset using five different MAs

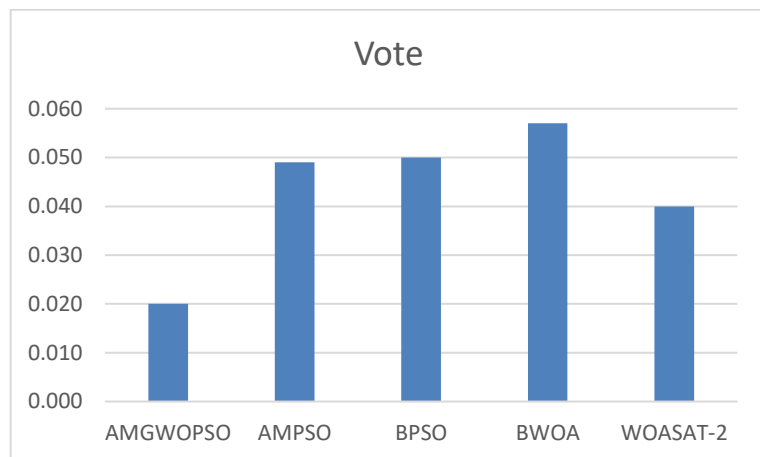


Fig. 6.58: Average fitness obtained for the Vote dataset using five different MAs

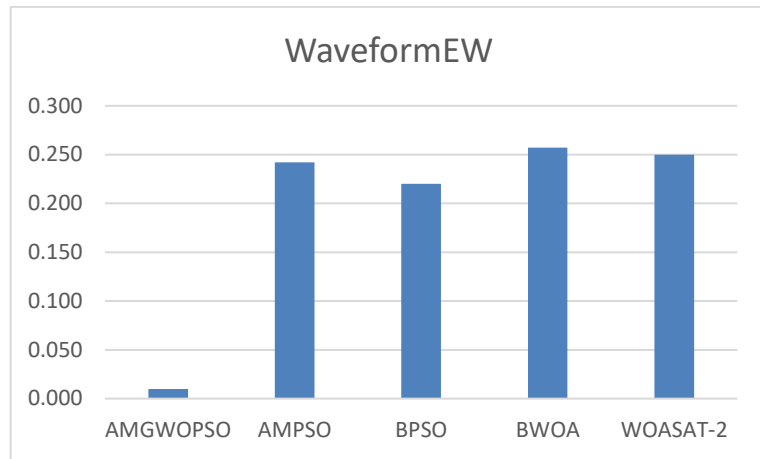


Fig. 6.59: Average fitness obtained for the WaveformEW dataset using five different MAs

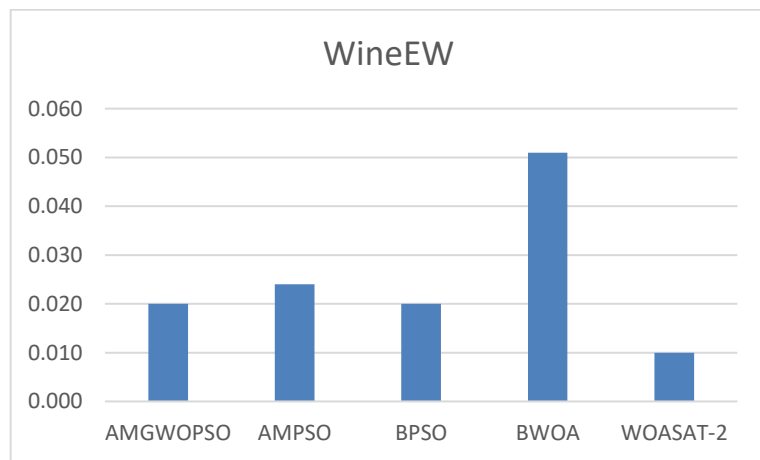


Fig. 6.60: Average fitness obtained for the WineEW dataset using five different MAs

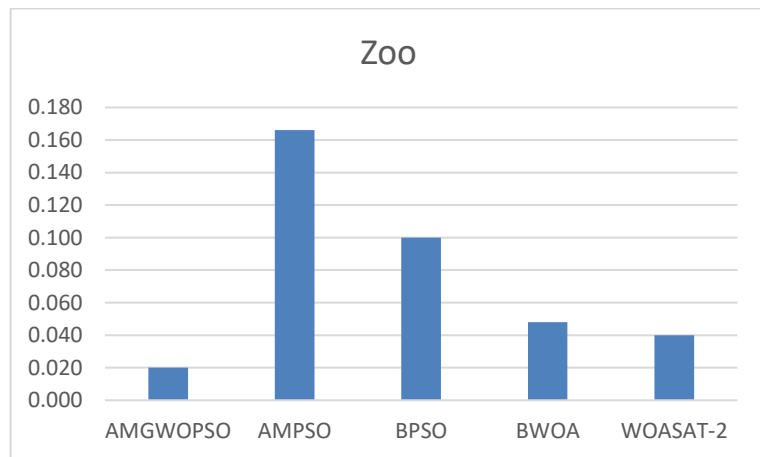


Fig. 6.61: Average fitness obtained for the Zoo dataset using five different MAs

## 7. SUMMARY OF RESULTS AND DISCUSSION

The recent proliferation of new metaheuristic algorithms after the introduction of metaphor-based development has elicited what researchers refer to as the novel algorithm dilemma. Consequently, the Evolutionary Computation (EC) Bestiary project was launched in 2018 to track and categorize these new metaphor-based metaheuristic algorithms [105]. However, issues such as an overly competitive and novel mindset, bio-inspired lingo, and algorithmic dualities among several others were highlighted by the authors in [104] as part of the defects identified in the new metaphor-based algorithms.

On the other hand, proponents of these new metaphor-based metaheuristic algorithms such as GWO cite the “*No Free Lunch*” doctrine in optimization as the main motivation for their metaheuristic algorithmic creativity [57]. Indeed, it is worth mentioning that while an algorithm may perform well on some datasets, its performance may suffer greatly when applied to a different dataset [68]. Thus, it is important to create novel or hybrid methods to optimally address a particular problem or set of problems. While it is important to acknowledge that the proliferation of these metaphor-based metaheuristics algorithms will not decline despite facing generic opposition, every effort must be made to avoid repetitive flaws usually associated with the new metaheuristic methods.

The aims of this dissertation were achieved.

1. Modification of the sentiment analysis pipeline to improve sentiment classification:
  - a. To determine the best lexicon-based technique based on classification performance and also identify tweet (text) contents most illustrative of positive and negative value user contribution.

The author adopted the sentiment analysis workflow in chapter 4. The lexicons were selected based on the literature review and the authors’ interest. Results of this work showcasing the best lexicon based on classification as well as tweets contents most illustrative of positive and negative value-user contribution have since been published in a journal [47].

2. Design a metaheuristic-based solution for sentiment analysis using the binary PSO given the BPSO’s impressive performance in feature selection.

A metaheuristic-based approach for optimal selection of features subset via the binary particle swarm optimization (BPSO) metaheuristic algorithm with the view to improve sentiment classification accuracy on the sentiment labelled sentences benchmark dataset was conducted by the author of this dissertation. The study results were presented at the International Conference on Electrical, Computer, and Energy Technologies (ICECET) held in Prague July 20-22, 2022, and have since been published in IEEE Xplore [56].

3. Develop a new Angle Modulated-based metaheuristic memetic method for wrapper feature selection. The proposed method utilizes the GWO AMPSO.

The author developed the hybrid Angle Modulated GWO and PSO (AMGWOPSO) for feature selection tasks (see chapters 5 and 6). Possibilities for publishing the results in a reputable scientific journal will be explored and submitted in due course.

4. Test and evaluate the proposed metaheuristic and memetic method on some selected publicly available benchmark datasets from UCI Invine [32].

The statistical tests and evaluation processes together with the visualization of results where applicable were presented in the results described in chapter 6.

As a recap, the dual segmentation of the entire dissertation is organised into chapters as follows. The evolving nature of the social media landscape that has redefined the way our society works leading to an upsurge in sentiments analysis research as well as some concepts in soft computing techniques were covered in chapter one. The second chapter reviews (state-of-the-art) literature focusing on approaches that help in understanding and gaining valuable insights from the huge amount of unstructured social media data (Twitter) available as well as the role of soft computing techniques in sentiment analysis.

Furthermore, a brief history and state-of-the-art solutions covering core concepts related to NLP, sentiment analysis, and metaheuristics algorithms for feature selection are discussed. Particular emphasis is laid on the application of angle modulation in hybrid metaheuristic algorithms for feature selection.

A summary of the proposed thesis objectives is presented in chapter three. The next chapter describes the sentiment analysis and feature selection workflows adopted respectively. The essential steps involved in each of the workflows are highlighted. Chapter five presents the novel low-level coevolutionary mixed hybrid AMGWOPSO technique. Chapter six showcases the relevant experimental results obtained and published (in a journal and a conference) by the author of this thesis where the following observations were made.

Besides enhancing the existing literature on social media analytics, the sentiment analysis approach adopted in this thesis demonstrates that adopting a data-driven approach produces robust and generalizable outputs compared to conventional marketing approaches such as customer surveys, focus groups, and interviews.

An implementation of text feature selection using the BPSO to enhance sentiment classification and analysis is also demonstrated. The results affirm the generalization power of SC methods given that social media data such as tweets, reviews, etc... serves as a good data source used in attesting the reasoning and search capabilities of SC techniques. Furthermore, the author of this thesis employs a new hybrid metaheuristic algorithm to solve feature selection tasks on eighteen selected UCI benchmark datasets.

The authors' findings confirm the competitive and better performance of the AMGWOPSO when juxtaposed with other work-related metaheuristics methods available in feature selection literature. Further statistical tests also confirm AMGWOPSO as a potent technique for resolving binary optimization problems

across different domains. Chapter seven presents the contribution of the dissertation to science and industry with chapter eight concluding the work.

## 8. CONTRIBUTION OF THESIS TO SCIENCE AND PRACTICE

As the literature suggests, the upsurge in social media websites has triggered a huge data source for mining interesting expressions on a variety of subjects. Social media data offers great insights for firms and prospective customers in general.

The results of sentiment analysis in this work demonstrate that in today's world of empowered customers, firms need to focus on customer engagement to enhance customer experience via social media channels (e.g., Twitter) since the meaning of competitive advantage has shifted from purely competing over price and product to building loyalty and trust.

Furthermore, adopting a data-driven approach for this work produced robust and generalizable outputs compared to conventional marketing approaches such as customer surveys, focus groups, and interviews. From a broader managerial perspective, the study findings can make firms responsive to customer needs and think strategically while focusing on areas of service provision that are vital to business growth. Theoretically, the study contributes to broadening the scope of online banking given the interplay of consumer sentiments via the social media channel.

Metaheuristic algorithms are considered by researchers in recent times as reliable and effective tools for providing optimal solutions to optimization tasks such as feature selection. For feature selection problems, maximizing the classifier performance and reducing the number of features to overcome the curse of dimensionality remains a key priority.

It is in this light that a metaheuristic-based Binary Particle Swarm Optimization (BPSO) algorithm is utilized to demonstrate textual FS for effective sentiment analysis/classification on the UCI benchmark sentiment labelled sentences dataset. The results of the evaluation with and without the BPSO on the baseline models prove the superiority of the metaheuristic approach in text feature selection.

Despite the successes of the BPSO in solving binary tasks like feature selection, available literature suggests that it has received enormous criticism for its extreme modification compared to the original PSO [26]. As such, the AMPSO is hailed for its ability to utilize the standard PSO to solve binary problems without making any changes to the standard PSO algorithm [30].

By taking inspiration from the BGWOPSO and creating a new (novel) hybrid AMGWOPSO, the concept of employing a trigonometric fitness function as a bit string generator is extended to hybrid metaheuristic algorithms. In other words, the resulting hybrid metaheuristic algorithm has embraced the angle modulation technique used in the domain of signal processing within the telecommunication industry [27]. Indeed, this ability to use the AMGWOPSO as a wrapper feature selection method for feature selection constitutes a major contribution to this

work. In sum, the principal theoretical contributions of this new proposed hybrid AMGWOPSO are chronicled below:

- Introduction of angle modulation to the literature on memetic metaheuristic methods for feature selection.
- Propose a hybrid binary AMGWOPSO to solve binary optimization problems.
- Extend the concept of angle modulation from non-hybrid metaheuristic methods to the memetic metaheuristic paradigm.
- Testing and validating the proposed AMGWOPSO's performance on 18 UCI benchmark datasets and other selected metaheuristic algorithms for comparison.

## 9. CONCLUSION

Understanding the significance of social media has attracted academic attention in recent times. As a prominent scholar once put it, social media is no longer a passing sensation or fad. Customer opinions expressed on social media can convey important messages that businesses can use to build strong relationships with customers. As social media usage among the general population grows, so are its uses in the business world as more businesses turn to social media as a cost-effective and efficient way to connect with many clients.

This dissertation was structured into two segments. The first segment of this dissertation utilizes the abundance of social media data available online as leverage to explore the use of soft computing techniques for sentiment analysis. During this process, text mining techniques are employed to analyze UGC from social media posts (tweets) to support consumer decision-making and marketing communications.

The results show that firms should be more proactive in learning about their customers' behaviour by analysing their social media messages and also focusing on the factors that influence how customers perceive specific products or services.

The second part of the dissertation builds on the earlier segment by extending the use of evolutionary computation techniques to solve feature selection problems. In this phase, a metaheuristic-based solution using the Particle Swarm Optimization (PSO) algorithm for optimal subset text feature selection during sentiment analysis is implemented. Furthermore, a low-level coevolutionary mixed hybrid approach is adopted to develop a new hybrid metaheuristic algorithm by hybridizing the GWO with the AMPSO for wrapper feature selection.

Despite the successes chalked by the novel hybrid method, the fixed amplitude of the generating function constitutes a drawback to the proposed approach given that it is a sine wave. In the future, the author of this thesis will consider modifying the amplitude of the generating function to potentially scale the effect of the vertical shift coefficient. While the authors' proposed AMGWOPSO represents the first attempt at introducing the concept of angle modulation into hybrid metaheuristics FS literature as far as the author can tell, this concept can be further experimented with other non-hybrid/hybrid metaheuristic algorithms to assess their efficacy and stability.



## BIBLIOGRAPHY

- [1] DiMaggio, P., Hargittai, E., Neuman, W. R., & Robinson, J. P. (2001). Social implications of the Internet. *Annual review of sociology*, 27(1), 307-336.
- [2] Kumar, A., & Sebastian, T. M. (2012). Sentiment analysis on twitter. *International Journal of Computer Science Issues (IJCSI)*, 9(4), 372.
- [3] Park, E. (2019). Motivations for customer revisit behavior in online review comments: Analyzing the role of user experience using big data approaches. *Journal of Retailing and Consumer Services*, 51, 14-18.
- [4] Guo, J., Wang, X., & Wu, Y. (2020). Positive emotion bias: Role of emotional content from online customer reviews in purchase decisions. *Journal of Retailing and Consumer Services*, 52, 101891.
- [5] Collomb, A., Costea, C., Joyeux, D., Hasan, O., & Brunie, L. (2014). A study and comparison of sentiment analysis methods for reputation evaluation. *Rapport de recherche RR-LIRIS-2014-002*.
- [6] Kumar, A., & Teeja, M. S. (2012). Sentiment analysis: A perspective on its past, present and future. *International Journal of Intelligent Systems and Applications*, 4(10), 1.
- [7] Sivanandam, S. N., & Deepa, S. N. (2007). *Principles of soft computing (with CD)*. John Wiley & Sons.
- [8] Balas, V. E., & Fodor, J. (2013). *New Concepts and Applications in Soft Computing*. A. R. Várkonyi-Kóczy (Ed.). Springer.
- [9] Chowdhary, K. (2020). Natural language processing. *Fundamentals of artificial intelligence*, 603-649.
- [10] Liu, B. (2012). Sentiment analysis and opinion mining. *Synthesis lectures on human language technologies*, 5(1), 1-167.
- [11] Fellbaum, W. (1998). *An Electronic Lexical Database (Language, Speech, and Communication)*.
- [12] Kamps, J., Marx, M., Mokken, R. J., & De Rijke, M. (2004, May). Using WordNet to measure semantic orientations of adjectives. In *LREC* (Vol. 4, pp. 1115-1118).
- [13] Ibrahim, N. F., Wang, X., & Bourne, H. (2017). Exploring the effect of user engagement in online brand communities: Evidence from Twitter. *Computers in Human Behavior*, 72, 321-338.
- [14] Thelwall, M., Buckley, K., & Paltoglou, G. (2012). Sentiment strength detection for the social web. *Journal of the American Society for Information Science and Technology*, 63(1), 163-173.
- [15] Hutto, C., & Gilbert, E. (2014, May). Vader: A parsimonious rule-based model for sentiment analysis of social media text. In *Proceedings of the international AAAI conference on web and social media* (Vol. 8, No. 1, pp. 216-225).

- [16] Ho, T. T., & Huang, Y. (2021). Stock Price Movement Prediction Using Sentiment Analysis and CandleStick Chart Representation. *Sensors*, 21(23), 7957.
- [17] Kumar, A., & Jaiswal, A. (2020). Systematic literature review of sentiment analysis on Twitter using soft computing techniques. *Concurrency and Computation: Practice and Experience*, 32(1), e5107.
- [18] Kumar, A., Dabas, V., & Hooda, P. (2020). Text classification algorithms for mining unstructured data: a SWOT analysis. *International Journal of Information Technology*, 12(4), 1159-1169.
- [19] Jordan, M. I., & Mitchell, T. M. (2015). Machine learning: Trends, perspectives, and prospects. *Science*, 349(6245), 255-260.
- [20] Drucker, H., Burges, C. J., Kaufman, L., Smola, A., & Vapnik, V. (1996). Support vector regression machines. *Advances in neural information processing systems*, 9.
- [21] Zhang, J., Jin, R., Yang, Y., & Hauptmann, A. (2003). Modified logistic regression: An approximation to SVM and its applications in large-scale text categorization.
- [22] Sulzmann, J. N., Fürnkranz, J., & Hüllermeier, E. (2007, September). On pairwise naive bayes classifiers. In *European Conference on Machine Learning* (pp. 371-381). Springer, Berlin, Heidelberg.
- [23] Dash, S., Pani, S. K., Rodrigues, J. J., & Majhi, B. (Eds.). (2022). *Deep Learning, Machine Learning and IoT in Biomedical and Health Informatics: Techniques and Applications*. CRC Press.
- [24] Gu, J., Wang, Z., Kuen, J., Ma, L., Shahroudy, A., Shuai, B., ... & Chen, T. (2018). Recent advances in convolutional neural networks. *Pattern Recognition*, 77, 354-377.
- [25] Mohamed, A. W., Hadi, A. A., & Mohamed, A. K. (2020). Gaining-sharing knowledge based algorithm for solving optimization problems: a novel nature-inspired algorithm. *International Journal of Machine Learning and Cybernetics*, 11(7), 1501-1529.
- [26] Leonard, B. J., & Engelbrecht, A. P. (2015, December). Frequency distribution of candidate solutions in angle modulated particle swarms. In *2015 IEEE Symposium Series on Computational Intelligence* (pp. 251-258). IEEE.
- [27] Proakis, J. G., & Salehi, M. (2002). *Communication systems engineering* Prentice-Hall. Inc. Upper Saddle River, New Jersey.
- [28] Singh, N., & Singh, S. B. (2017). Hybrid algorithm of particle swarm optimization and grey wolf optimizer for improving convergence performance. *Journal of Applied Mathematics*, 2017.
- [29] Franken, C. J. (2005). *PSO-based coevolutionary game learning* (Doctoral dissertation, University of Pretoria).
- [30] Pampara, G., Franken, N., & Engelbrecht, A. P. (2005, September). Combining particle swarm optimisation with angle modulation to solve

- binary problems. In *2005 IEEE congress on evolutionary computation* (Vol. 1, pp. 89-96). IEEE.
- [31] Leonard, B. J. (2017). *Critical analysis of angle modulated particle swarm optimisers* (Doctoral dissertation, University of Pretoria).
- [32] Blake, C. (1998). UCI repository of machine learning databases.[Online]. Available: <http://www.ics.uci.edu/~mlearn/MLRepository.html>.
- [33] Jović, A., Brkić, K., & Bogunović, N. (2015, May). A review of feature selection methods with applications. In *2015 38th international convention on information and communication technology, electronics and microelectronics (MIPRO)* (pp. 1200-1205). IEEE.
- [34] Jona, J., & Nagaveni, N. (2012). A hybrid swarm optimization approach for feature set reduction in digital mammograms. *WSEAS Trans Inf Sci Appl*, 9(11), 340-349.
- [35] Talbi, E. G., Jourdan, L., Garcia-Nieto, J., & Alba, E. (2008, March). Comparison of population based metaheuristics for feature selection: Application to microarray data classification. In *2008 IEEE/ACS International Conference on Computer Systems and Applications* (pp. 45-52). IEEE.
- [36] Sun, Z., Bebis, G., & Miller, R. (2004). Object detection using feature subset selection. *Pattern recognition*, 37(11), 2165-2176.
- [37] Mafarja, M. M., & Mirjalili, S. (2017). Hybrid whale optimization algorithm with simulated annealing for feature selection. *Neurocomputing*, 260, 302-312.
- [38] Zorarpacı, E., & Özel, S. A. (2016). A hybrid approach of differential evolution and artificial bee colony for feature selection. *Expert Systems with Applications*, 62, 91-103.
- [39] Kennedy, J., & Eberhart, R. C. (1997, October). A discrete binary version of the particle swarm algorithm. In *1997 IEEE International conference on systems, man, and cybernetics. Computational cybernetics and simulation* (Vol. 5, pp. 4104-4108). IEEE.
- [40] Al-Tashi, Q., Kadir, S. J. A., Rais, H. M., Mirjalili, S., & Alhussian, H. (2019). Binary optimization using hybrid grey wolf optimization for feature selection. *Ieee Access*, 7, 39496-39508.
- [41] Agrawal, P., Abutarboush, H. F., Ganesh, T., & Mohamed, A. W. (2021). Metaheuristic algorithms on feature selection: A survey of one decade of research (2009-2019). *IEEE Access*, 9, 26766-26791.
- [42] Emary, E., Zawbaa, H. M., & Hassanien, A. E. (2016). Binary grey wolf optimization approaches for feature selection. *Neurocomputing*, 172, 371-381.
- [43] Dua, D. and Graff, C. (2019). UCI Machine Learning Repository.[Online]. Available: [<http://archives.ics.uci.edu/ml>]. Irvine, CA: University of California, School of Information and Computer Science.
- [44] Nielsen, F. Å. (2017). afinn project.

- [45] Baccianella, S., Esuli, A., & Sebastiani, F. (2010, May). Sentiwordnet 3.0: An enhanced lexical resource for sentiment analysis and opinion mining. In *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10)*.
- [46] Loria, S. (2018). Textblob Documentation. *Release 0.15*, 2(8).
- [47] Botchway, R. K., Jibril, A. B., Oplatková, Z. K., & Chovancová, M. (2020). Deductions from a Sub-Saharan African Bank's Tweets: A sentiment analysis approach. *Cogent Economics & Finance*, 8(1), 1776006.
- [48] Pustejovsky, J., & Stubbs, A. (2012). Natural Language Annotation for Machine Learning: A guide to corpus-building for applications. " O'Reilly Media, Inc."
- [49] Grégoire, Y., Salle, A., & Tripp, T. M. (2015). Managing social media crises with your customers: The good, the bad, and the ugly. *Business Horizons*, 58(2), 173-182.
- [50] Ernst & Young (2017). Customer Experience: Innovate Like a FinTech," Retrieved from [https://www.ey.com/Publication/vwLUAssets/ey-gcbs-customerexperience/\\$FILE/ey-gcbs-customer-experience.pdf](https://www.ey.com/Publication/vwLUAssets/ey-gcbs-customerexperience/$FILE/ey-gcbs-customer-experience.pdf).
- [51] Ibrahim, N. F., & Wang, X. (2019). Decoding the sentiment dynamics of online retailing customers: Time series analysis of social media. *Computers in Human Behavior*, 96, 32-45.
- [52] Kotzias, D., Denil, M., De Freitas, N., & Smyth, P. (2015, August). From group to individual labels using deep features. In *Proceedings of the 21th ACM SIGKDD international conference on knowledge discovery and data mining* (pp. 597-606).
- [53] Vieira, S. M., Mendonça, L. F., Farinha, G. J., & Sousa, J. M. (2013). Modified binary PSO for feature selection using SVM applied to mortality prediction of septic patients. *Applied Soft Computing*, 13(8), 3494-3504.
- [54] Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., ... & Duchesnay, E. (2011). Scikit-learn: Machine learning in Python. *the Journal of machine Learning research*, 12, 2825-2830.
- [55] Miranda, L. J. (2018). PySwarms: a research toolkit for Particle Swarm Optimization in Python. *Journal of Open Source Software*, 3(21), 433.
- [56] Botchway, R. K., Yadav, V., Komínková, Z. O., & Senkerik, R. (2022, July). Text-based feature selection using binary particle swarm optimization for sentiment analysis. In *2022 International Conference on Electrical, Computer and Energy Technologies (ICECET)* (pp. 1-4). IEEE.
- [57] Ji, B., Lu, X., Sun, G., Zhang, W., Li, J., & Xiao, Y. (2020). Bio-inspired feature selection: An improved binary particle swarm optimization approach. *IEEE Access*, 8, 85989-86002.
- [58] Wilcoxon, F. (1992). Individual comparisons by ranking methods. In *Breakthroughs in statistics* (pp. 196-202). Springer, New York, NY.

- [59] Faris, H., Aljarah, I., Al-Betar, M. A., & Mirjalili, S. (2018). Grey wolf optimizer: a review of recent variants and applications. *Neural computing and applications*, 30(2), 413-435.
- [60] Chantar, H., Mafarja, M., Alsawalqah, H., Heidari, A. A., Aljarah, I., & Faris, H. (2020). Feature selection using binary grey wolf optimizer with elite-based crossover for Arabic text classification. *Neural Computing and Applications*, 32(16), 12201-12220.
- [61] Liu, X., & Shang, L. (2013, June). A fast wrapper feature subset selection method based on binary particle swarm optimization. In *2013 IEEE congress on evolutionary computation* (pp. 3347-3353). IEEE.
- [62] Kennedy, J., & Eberhart, R. (1995, November). Particle swarm optimization. In *Proceedings of ICNN'95-international conference on neural networks* (Vol. 4, pp. 1942-1948). IEEE.
- [63] Li, W., Qi, F., Tang, M., & Yu, Z. (2020). Bidirectional LSTM with self-attention mechanism and multi-channel features for sentiment classification. *Neurocomputing*, 387, 63-77.
- [64] Talbi, E.G. (2009). *Metaheuristics: from design to implementation*. John Wiley & Sons.
- [65] Sharma, M., & Kaur, P. (2021). A comprehensive analysis of nature-inspired meta-heuristic techniques for feature selection problem. *Archives of Computational Methods in Engineering*, 28, 1103-1127.
- [66] Al-Tashi, Q., Md Rais, H., Abdulkadir, S. J., Mirjalili, S., & Alhussian, H. (2020). A review of grey wolf optimizer-based feature selection methods for classification. *Evolutionary machine learning techniques*, 273-286.
- [67] Nguyen, T. (2020). A collection of the state-of-the-art meta-heuristics algorithms in python: Mealpy. Zenodo.
- [68] Agushaka, J. O., Ezugwu, A. E., & Abualigah, L. (2022). Gazelle Optimization Algorithm: A novel nature-inspired metaheuristic optimizer. *Neural Computing and Applications*, 1-33.
- [69] Al-Wajih, R., Abdulkadir, S. J., Aziz, N., Al-Tashi, Q., & Talpur, N. (2021). Hybrid binary grey wolf with Harris hawks optimizer for feature selection. *IEEE Access*, 9, 31662-31677.
- [70] Wäldchen, S., Macdonald, J., Hauch, S., & Kutyniok, G. (2021). The computational complexity of understanding binary classifier decisions. *Journal of Artificial Intelligence Research*, 70, 351-387.
- [71] Mirjalili, S., & Lewis, A. (2016). The whale optimization algorithm. *Advances in engineering software*, 95, 51-67
- [72] Zhang, Q., & Liu, L. (2019). Whale optimization algorithm based on lamarckian learning for global optimization problems. *Ieee Access*, 7, 36642-36666.
- [73] Chakraborty, S., Saha, A. K., Sharma, S., Chakraborty, R., & Debnath, S. (2023). A hybrid whale optimization algorithm for global

- optimization. *Journal of Ambient Intelligence and Humanized Computing*, 14(1), 431-467.
- [74] Mirjalili, S., Mirjalili, S. M., & Lewis, A. (2014). Grey wolf optimizer. *Advances in engineering software*, 69, 46-61.
- [75] Dave, K., Lawrence, S., & Pennock, D. M. (2003, May). Mining the peanut gallery: Opinion extraction and semantic classification of product reviews. In *Proceedings of the 12th international conference on World Wide Web* (pp. 519-528).
- [76] Singh, P., Dwivedi, Y. K., Kahlon, K. S., Sawhney, R. S., Alalwan, A. A., & Rana, N. P. (2020). Smart monitoring and controlling of government policies using social media and cloud computing. *Information Systems Frontiers*, 22, 315-337.
- [77] Festinger, L. A Theory of Cognitive Dissonance; Stanford University Press: Palo Alto, CA, USA, 1957; Volume 2, ISBN 0-8047-0911-4.
- [78] Morvan, C., & O'Connor, A. J. (2017). *An analysis of Leon Festinger's A theory of cognitive dissonance*. Macat Library.
- [79] Kim, A. J., & Johnson, K. K. (2016). Power of consumers using social media: Examining the influences of brand-related user-generated content on Facebook. *Computers in human behavior*, 58, 98-108.
- [80] Wu, J. J., & Chang, S. T. (2020). Exploring customer sentiment regarding online retail services: a topic-based approach. *Journal of Retailing and Consumer Services*, 55, 102145.
- [81] Nisar, T. M., Prabhakar, G., Ilavarasan, P. V., & Baabdullah, A. M. (2020). Up the ante: Electronic word of mouth and its effects on firm reputation and performance. *Journal of Retailing and Consumer Services*, 53, 101726.
- [82] Shareef, M. A., Kapoor, K. K., Mukerji, B., Dwivedi, R., & Dwivedi, Y. K. (2020). Group behavior in social media: Antecedents of initial trust formation. *Computers in Human Behavior*, 105, 106225.
- [83] Grover, P., Kar, A. K., Dwivedi, Y. K., & Janssen, M. (2019). Polarization and acculturation in US Election 2016 outcomes—Can twitter analytics predict changes in voting preferences. *Technological Forecasting and Social Change*, 145, 438-460.
- [84] Bughin, J., Chui, M., & Miller, A. (2009). How companies are benefiting from Web 2.0. *McKinsey Quarterly*, 9 (4), 84–85. <https://www.mckinsey.com/businessfunctions/mckinsey-digital/ourinsights/how-companies-are-benefiting-from-web-20-mckinsey-globalsurvey-results>
- [85] Statista 2019a Statista. (2019a). Social media - Statistics & Facts. In.
- [86] Statista 2019b Statista. (2019b). Social media usage in the United States - Statistics & Facts.
- [87] Choi, J., Yoon, J., Chung, J., Coh, B. Y., & Lee, J. M. (2020). Social media analytics and business intelligence research: A systematic review. *Information Processing & Management*, 57(6), 102279.

- [88] Bhimani, H., Mention, A. L., & Barlatier, P. J. (2019). Social media and innovation: A systematic literature review and future research directions. *Technological Forecasting and Social Change*, 144, 251-269.
- [89] Jeong, B., Yoon, J., & Lee, J. M. (2019). Social media mining for product planning: A product opportunity mining approach based on topic modeling and sentiment analysis. *International Journal of Information Management*, 48, 280-290.
- [90] Trappey, A. J., Trappey, C. V., Fan, C. Y., & Lee, I. J. (2018). Consumer driven product technology function deployment using social media and patent mining. *Advanced Engineering Informatics*, 36, 120-129.
- [91] Burke-Garcia, A. (2017). *Opinion leaders for health: formative research with bloggers about health information dissemination* (Doctoral dissertation, George Mason University).
- [92] Jain, S., & Sinha, A. (2020). Identification of influential users on Twitter: A novel weighted correlated influence measure for Covid-19. *Chaos, solitons & fractals*, 139, 110037.
- [93] Featherstone, J. D., Barnett, G. A., Ruiz, J. B., Zhuang, Y., & Millam, B. J. (2020). Exploring childhood anti-vaccine and pro-vaccine communities on twitter—a perspective from influential users. *Online Social Networks and Media*, 20, 100105.
- [94] Fernández-Gavilanes, M., Álvarez-López, T., Juncal-Martínez, J., Costa-Montenegro, E., & González-Castaño, F. J. (2016). Unsupervised method for sentiment analysis in online texts. *Expert Systems with Applications*, 58, 57-75.
- [95] Xing, F. Z., Pallucchini, F., & Cambria, E. (2019). Cognitive-inspired domain adaptation of sentiment lexicons. *Information Processing & Management*, 56(3), 554-564.
- [96] San Vicente, I., Agerri, R., & Rigau, G. (2014, April). Simple, robust and (almost) unsupervised generation of polarity lexicons for multiple languages. In *Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics* (pp. 88-97).
- [97] Deng, S., Sinha, A. P., & Zhao, H. (2017). Adapting sentiment lexicons to domain-specific social media texts. *Decision Support Systems*, 94, 65-76.
- [98] Hamilton, W. L., Clark, K., Leskovec, J., & Jurafsky, D. (2016, November). Inducing domain-specific sentiment lexicons from unlabeled corpora. In *Proceedings of the conference on empirical methods in natural language processing. conference on empirical methods in natural language processing* (Vol. 2016, p. 595). NIH Public Access.
- [99] Bandhakavi, A., Wiratunga, N., Massie, S., & Luhar, R. (2018). Context Extraction for Aspect-Based Sentiment Analytics: Combining Syntactic, Lexical and Sentiment Knowledge. In *Artificial Intelligence XXXV: 38th SGAI International Conference on Artificial Intelligence, AI 2018*,

- Cambridge, UK, December 11–13, 2018, *Proceedings 38* (pp. 357-371). Springer International Publishing.
- [100] Darwich, M., Mohd, S. A., Omar, N., & Osman, N. A. (2019). Corpus-Based Techniques for Sentiment Lexicon Generation: A Review. *J. Digit. Inf. Manag.*, 17(5), 296.
- [101] Macedo, M., Siqueira, H., Figueiredo, E., Santana, C., Lira, R. C., Gokhale, A., & Bastos-Filho, C. (2021). Overview on binary optimization using swarm-inspired algorithms. *IEEE Access*, 9, 149814-149858.
- [102] Poldi, F. (2020). Twint-twitter intelligence tool. URL: <https://github.com/twintproject/twint> [Accessed 08/02/2020].
- [103] Artstein, R. (2017). Inter-annotator agreement. *Handbook of linguistic annotation*, 297-313.
- [104] Sörensen, K. (2015). Metaheuristics—the metaphor exposed. *International Transactions in Operational Research*, 22(1), 3-18.
- [105] Kudela, J. (2022). A critical problem in benchmarking and analysis of evolutionary computation methods. *Nature Machine Intelligence*, 1-8.
- [106] Mohammed, H., & Rashid, T. (2020). A novel hybrid GWO with WOA for global numerical optimization and solving pressure vessel design. *Neural Computing and Applications*, 32(18), 14701-14718.
- [107] Sharawi, M., Zawbaa, H. M., & Emary, E. (2017, February). Feature selection approach based on whale optimization algorithm. In 2017 Ninth international conference on advanced computational intelligence (ICACI) (pp. 163-168). IEEE.
- [108] Mafarja, M., Jaber, I., Ahmed, S., & Thaher, T. (2021). Whale optimisation algorithm for high-dimensional small-instance feature selection. *International Journal of Parallel, Emergent and Distributed Systems*, 36(2), 80-96.
- [109] Alwajih, R., Abdulkadir, S. J., Al Hussian, H., Aziz, N., Al-Tashi, Q., Mirjalili, S., & Alqushaibi, A. (2022). Hybrid binary whale with harris hawks for feature selection. *Neural Computing and Applications*, 34(21), 19377-19395.



## LIST OF FIGURES

Fig. 1.1: Overview of Soft computing techniques [17] .....	12
Fig. 2.1: Sentiment analysis approaches .....	16
Fig. 2.2: Metaheuristic algorithm categories [41] .....	20
Fig. 2.3: Applications of metaheuristic techniques from various domains [65] .....	22
Fig. 2.4: Social hierarchy of grey wolves [74] .....	25
Fig. 2.5: A 5-dim binary solution generated by sampling g at regular intervals [31] .....	30
Fig. 2.6: Classification of Feature Selection methods [41] .....	30
Fig. 4.1: Overview of the sentiment analysis workflow .....	34
Fig. 4.2: Feature selection process [41] .....	36
Fig. 6.1: Three-phase methodology deployed [47].....	42
Fig. 6.2: Negative sentiments word cloud .....	45
Fig. 6.3: Tweet illustrating a pain-point. ....	45
Fig. 6.4: Conceptual framework of the study .....	46
Fig. 6.5: Particle Swarm Representation .....	46
Fig. 6.6: Two-dimensional view of data using PCA .....	47
Fig. 6.7: Optimized/Non-Optimised plot using SVM, k-NN, and NB classifiers. .....	48
Fig. 6.8: Classification accuracy results of five different MAs on the Breastcancer datasets .....	57
Fig. 6.9: Classification accuracy results of five different MAs on the BreastEW dataset.....	57
Fig. 6.10: Classification accuracy results of five different MAs on the CongressEW dataset .....	57
Fig. 6.11: Classification accuracy results of five different MAs on the Exactly dataset.....	58
Fig. 6.12: Classification accuracy results of five different MAs on the Exactly2 dataset.....	58
Fig. 6.13: Classification accuracy results of five different MAs on the HeartEW dataset.....	58
Fig. 6.14: Classification accuracy results of five different MAs on the IonosphereEW dataset.....	59
Fig. 6.15: Classification accuracy results of five different MAs on the KrVsKpEW dataset.....	59
Fig. 6.16: Classification accuracy results of five different MAs on the Lymphography dataset.....	59
Fig. 6.17: Classification accuracy results of five different MAs on the M-of-n dataset.....	60
Fig. 6.18: Classification accuracy results of five different MAs on the PenglungEW dataset .....	60

Fig. 6.19: Classification accuracy results of five different MAs on the SonarEW dataset.....	60
Fig. 6.20: Classification accuracy results of five different MAs on the SpectEW dataset.....	61
Fig. 6.21: Classification accuracy results of five different MAs on the Tic-tac-toe dataset.....	61
Fig. 6.22: Classification accuracy results of five different MAs on the Vote dataset.....	61
Fig. 6.23: Classification accuracy results of five different MAs on the WaveformEW dataset .....	62
Fig. 6.24: Classification accuracy results of five different MAs on the WineEW dataset.....	62
Fig. 6.25: Classification accuracy results of five different MAs on the Zoo dataset.....	62
Fig. 6.26: Average features selected using five different MAs on the Breastcancer dataset .....	63
Fig. 6.27: Average features selected using five different MAs on the BreastEW dataset.....	63
Fig. 6.28: Average features selected using five different MAs on the CongressEW dataset .....	63
Fig. 6.29: Average features selected using five different MAs on the Exactly dataset.....	64
Fig. 6.30: Average features selected using five different MAs on the Exactly2 dataset.....	64
Fig. 6.31: Average features selected using five different MAs on the HeartEW dataset.....	64
Fig. 6.32: Average features selected using five different MAs on the IonosphereEW dataset.....	65
Fig. 6.33: Average features selected using five different MAs on the KrVsKpEW dataset.....	65
Fig. 6.34: Average features selected using five different MAs on the Lymphography dataset.....	65
Fig. 6.35: Average features selected using five different MAs on the .....	66
Fig. 6.36: Average features selected using five different MAs on the .....	66
Fig. 6.37: Average features selected using five different MAs on the SonarEW dataset.....	66
Fig. 6.38: Average features selected using five different MAs on the SpectEW dataset.....	67
Fig. 6.39: Average features selected using five different MAs on the .....	67
Fig. 6.40 Average features selected using five different MAs on the Vote dataset.....	67
Fig. 6.41 Average features selected using five different MAs on the WaveformEW dataset .....	68

Fig. 6.42: Average features selected using five different MAs on the WineEW dataset.....	68
Fig. 6.43: Average features selected using five different MAs on the Zoo dataset.....	68
Fig. 6.44: Average fitness obtained for the Breastcancer dataset using five different MAs .....	69
Fig. 6.45: Average fitness obtained for the BreastEW dataset using five different MAs .....	69
Fig. 6.46: Average fitness obtained for the CongressEW dataset using five different MAs .....	69
Fig. 6.47: Average fitness obtained for the Exactly dataset using five different MAs .....	70
Fig. 6.48: Average fitness obtained for the Exactly2 dataset using five different MAs .....	70
Fig. 6.49: Average fitness obtained for the HeartEW dataset using five different MAs .....	70
Fig. 6.50: Average fitness obtained for the IonosphereEW dataset using five different MAs .....	71
Fig. 6.51: Average fitness obtained for the KrVsKpEW dataset using five different MAs .....	71
Fig. 6.52: Average fitness obtained for the Lymphography dataset using five different MAs .....	71
Fig. 6.53: Average fitness obtained for the M-of-n dataset using five different MAs .....	72
Fig. 6.54: Average fitness obtained for the PenglungEW dataset using five different MAs .....	72
Fig. 6.55: Average fitness obtained for the SonarEW dataset using five different MAs .....	72
Fig. 6.56: Average fitness obtained for the SpectEW dataset using five different MAs .....	73
Fig. 6.57: Average fitness obtained for the Tic-tac-toe dataset using five different MAs .....	73
Fig. 6.58: Average fitness obtained for the Vote dataset using five different MAs .....	73
Fig. 6.59: Average fitness obtained for the WaveformEW dataset using five different MAs .....	74
Fig. 6.60: Average fitness obtained for the WineEW dataset using five different MAs .....	74
Fig. 6.61: Average fitness obtained for the Zoo dataset using five different MAs .....	74

## LIST OF TABLES

Table 1.1 Business benefits from Web 2.0 .....	13
Table 2.1 Metaheuristic methods inspired by insects and reptiles [65] .....	21
Table 2.2 Metaheuristic methods inspired by birds and sea creatures [65]....	21
Table 2.3 Metaheuristic methods inspired by plants, humans, and other animals [65] .....	22
Table 5.1: UCI datasets used [32].....	41
Table 6.1 Classification of Tweets by various lexicons .....	43
Table 6.2 Sample Tweets from the Ecobank Tweets dataset [47].....	44
Table 6.3 Accuracy scores per lexicon .....	45
Table 6.4 TF-IDF scores for some pre-processed text .....	46
Table 6.5 Model accuracy scores.....	47
Table 6.6 Parameter settings .....	49
Table 6.7 Summary results of proposed AMGWOPSO compared to other related state-of-the-art algorithms.....	51
Table 6.8. Average classification and features results of proposed AMGWPSO compared to other related state-of-the-art algorithms.....	51
Table 6.9 Mean fitness results of proposed AMGWPSO compared to other related state-of-the-art algorithms.....	52
Table 6.10 Best and Worst fitness results of proposed AMGWOPSO compared to other related state-of-the-art methods.....	53
Table 6.11: Wilcoxon signed-rank test for mean fitness evaluation metric...	54
Table 6.12 Execution time (in seconds) of proposed AMGWOPSO compared with the hybrid WOASAT-2.....	55

## LIST OF SYMBOLS, ACRONYMS, AND ABBREVIATIONS

ABC	Artificial Bee Colony
AI	Artificial Intelligence
ANN	Artificial Neural Network
AMPSO	Angle Modulated Particle Swarm Optimization
AMGWOPSO	Angle Modulated Grey Wolf Optimization Particle Swarm Optimization
BNB	Bernoulli Naïve Bayes
BGWOPSO	Binary Grey Wolf Optimization Particle Swarm Optimization
Bi-LSTM	Bidirectional Long Short-Term Memory
BPSO	Binary Particle Swarm Optimization
BWOA	Binary Whale Optimization Algorithm
CNB	Complement Naïve Bayes
CNN	Convolution Neural Network
DE	Differential Evolution
DEABC	Differential Evolution Artificial Bee Colony
DL	Deep Learning
DNN	Deep Neural Network
EC	Evolutionary Computation
FL	Fuzzy Logic
FS	Feature Selection
GNB	Gaussian Naïve Bayes
GA	Genetic Algorithm
GAPSO	Genetic Algorithm Particle Swarm Optimization
GSA	Gravitational Search Algorithm
GWO	Grey Wolf Optimization
HRH	High-level Relay Hybrid
k-NN	K-Nearest Neighbour
LSTM	Long Short-Term Memory
LTH	Low-level Teamwork Hybrid
MA	Metaheuristic Algorithm
ME	Maximum Entropy
ML	Machine Learning
MNB	Multinomial Naïve Bayes
NB	Naïve Bayes
NLP	Natural language Processing
NLTK	Natural language Toolkit
NN	Neural Network
OM	Opinion Mining
PR	Probabilistic Reasoning
PSO	Particle Swarm Optimization

PSOGWO	Particle Swarm Optimization Grey Wolf Optimization
RBF	Radial Basis Function
RNN	Recurrent Neural Network
SA	Sentiment Analysis
SC	Soft Computing
sLDA	Supervised Latent Dirichlet Allocation
SVM	Support Vector Machine
TF-IDF	Term Frequency Inverse Document Frequency
UCI	University of California, Irvine
UGC	User Generated Content
VADER	Valence Aware Dictionary and sEntiment Reasoner
WDF	Word Document Frequency
WOA	Whale Optimization Algorithm
WOASAT-2	Whale Optimization Algorithm Simulated Annealing Tournament Selection 2

# LIST OF PUBLICATIONS BY THE AUTHOR

## Journals with IF:

[1] **Botchway, R. K.**, Jibril, A. B., Oplatková, Z. K., Jasek, R., & Kwarteng, M. A. (2021). Decision science: a multi-criteria decision framework for enhancing an electoral voting system. *Systems Science & Control Engineering*, 9(1), 556-569.

[2] Jibril, A. B., Kwarteng, M. A., **Botchway, R. K.**, Bode, J., & Chovancova, M. (2020). The impact of online identity theft on customers' willingness to engage in e-banking transaction in Ghana: A technology threat avoidance theory. *Cogent Business & Management*, 7(1), 1832825.

[3] **Botchway, R. K.**, Jibril, A. B., Oplatková, Z. K., & Chovancová, M. (2020). Deductions from a Sub-Saharan African Bank's Tweets: A sentiment analysis approach. *Cogent Economics & Finance*, 8(1), 1776006.

## Journals (Scopus):

[1] Yadav, V., **Botchway, R. K.**, Senkerik, R., & Oplatková, Z. K. (2021, December). Robotic Automation of Software Testing From a Machine Learning Viewpoint. In *Mendel* (Vol. 27, No. 2, pp. 68-73).

## Conference:

[1] Yadav, V., **Botchway, R. K.**, Senkerik, R., & Komínková, Z. O. (2023, July). Robot Automation of Software Using Genetic Algorithm. In *2023 International Conference on Electrical, Computer, Communications and Mechatronics Engineering (ICECCME)* IEEE. Accepted.

[2] **Botchway, R. K.**, Yadav, V., Komínková, Z. O., & Senkerik, R. (2022, July). Text-based feature selection using binary particle swarm optimization for sentiment analysis. In *2022 International Conference on Electrical, Computer and Energy Technologies (ICECET)* (pp. 1-4). IEEE.

[3] Yadav, V., **Botchway, R. K.**, Senkerik, R., & Komínková, Z. O. (2021, December). Robot Testing from a machine learning perspective. In *2021 International Conference on Electrical, Computer and Energy Technologies (ICECET)* (pp. 1-4). IEEE.

[4] Odei, M. A., Amoah, J., Jibril, A. B., **Botchway, R. K.**, Naatu, F., & Korantwi, I-Barimah, S. (2021, September). A Review of Barriers Facing Social Media Usage Among Firms in Less Digitalized Economies. In *European Conference on Innovation and Entrepreneurship* (pp. 677-XVII). Academic Conferences International Limited.

[5] Kwarteng, M. A., Ntsiful, A., **Botchway, R. K.**, Pilik, M., & Oplatková, Z. K. (2020, December). Consumer Insight on Driverless Automobile Technology Adoption via Twitter Data: A Sentiment Analytic Approach. In *International Working Conference on Transfer and Diffusion of IT* (pp. 463-473). Springer, Cham.

[6] Kwarteng, M. A., Jibril, A. B., **Botchway, R. K.**, Kwarteng, O. V., Pilik, M., & Chovancova, M. (2019, September). Assessing pre-purchase risk attributes towards used-products: evidence from e-shoppers in the Czech Republic. In *Proceedings of the 3rd International Conference on Business and Information Management* (pp. 15-20).




[7] **Botchway, R. K.**, Jibril, A. B., Kwarteng, M. A., Chovancova, M., & Oplatková, Z. K. (2019, September). A review of social media posts from UniCredit bank in Europe: a sentiment analysis approach. In *Proceedings of the 3rd International Conference on Business and Information Management* (pp. 74-79).



# AUTHOR'S PROFESSIONAL CURRICULUM VITAE

## PERSONAL INFORMATION

Raphael Kwaku Botchway

-  TGM 3050,76001 , Zlin, Czech Republic
-  +420 774948141
-  botchway@utb.cz /ralph.botchway@gmail.com

Sex Male | Date of birth 20/06/1979 | Nationality Ghanaian

## WORK EXPERIENCE

---

2004- JULY 2006

Electricity Company of Ghana  
▪ Oracle Database Administrator

2006- 2010

National Investment Bank, Accra - Ghana  
▪ Senior I.T Officer

2012- SEPT 2016

May-Awurade Adom Ventures, Accra - Ghana  
▪ Operations Manager

APRIL 2017- SEPT 2017

Exxon Mobil Kft, Hungary  
▪ Internship

## EDUCATION AND TRAINING

---

April 2019-Date  
(Expected graduation: March 2023)

### PhD Candidate

Faculty of Applied Informatics, Tomas Bata University in Zlin, Czech Republic

Thesis Title: Soft computing Techniques for Sentiment Analysis and Feature Selection

- Participated in faculty Internal Grant Agency (IGA) research projects

2010-2012

### MSc Informatics & Systems Engineering

Czech University of Life Sciences in Prague, Prague-Czechia

1998-2002

BSc (Hons) Computer Science

Kwame Nkrumah University of Ghana & Technology, Kumasi - Ghana

## STUDY INTERNSHIP

---

April 15, 2022 - July 15, 2022

Stellenbosch University, South Africa

Traineeship title : Angle modulated PSO for feature selection

Mother tongue(s)

English & Akan

Other language(s)	UNDERSTANDING		SPEAKING		WRITING
	Listening	Reading	Spoken interaction	Spoken production	
English	C1	C1	C1	C1	C1

Levels: A1/2: Basic user - B1/2: Independent user - C1/2 Proficient user  
Common European Framework of Reference for Languages

**Communication skills**

Good communication and listening skills  
Team-work oriented

**Organisational /  
managerial skills  
Computer skills**

Effectively led a team that deployed and managed Bancassurance product in NIB Ltd nationwide.  
MATLAB, Python, SQL, Oracle PL-SQL, Oracle Database Administration

**Honours and Awards**

Czech Government/Government of Ghana Scholarship 2010

**Soft computingové techniky pro analýzu sentimentu a výběr  
příznaků**

**Soft Computing Techniques for Sentiment Analysis and Feature Selection**

Doctoral Dissertation

Typesetting by: Raphael Kwaku Botchway

Publication year: 2023