CSF 407 – ARTIFICIAL INTELLIGENCE

FEATURE SELECTION ON CANCER DATASET

USING

METAHEURICTIC ALGORITHMS

TEAM MEMBERS ID	NAMES
1 2021A7PS2084H	Chinni Vamshi Krushna
2 2021A7PS3111H	Pratik Patil
3 2021AAPS1971H	Rakul Chauhan
4 2021A4PS3195H	Saksham Sawhney

GROUP ID - 41

1. INTRODUCTION

The process of feature selection is crucial in the world of machine learning and decision making based on data because it has the power to greatly improve the effectiveness, understanding, and accuracy of predictive models. In order to reduce dimensions and concentrate on the most discriminative qualities, the process of selecting features entails selecting a subset of the most pertinent and useful features from the original dataset. A smart feature selection speeds up learning, minimizes the danger of overfitting, lowers computing costs, and makes models easier to comprehend.

The significance of feature selection cannot be understated as machine learning applications continue to spread across several industries including healthcare, finance, natural language processing, image identification, and more. A carefully designed feature selection method may increase the flexibility of algorithms, increase the generalizability of models, and contribute to the overall success of data-driven projects. As a result, several strategies and procedures have been investigated by both researchers and practitioners in order to overcome the difficulties associated with feature selection.

This study will examine several aspects of the issue, covering filter, wrapper, and embedding techniques as well as cutting-edge methods like metaheuristic-based feature selection. Feature selection is a multidimensional problem that has been handled from a variety of perspectives. We attempt to shed light on the advantages and disadvantages of each strategy, offering insights into their fit for various data kinds, issue domains, and aims through a critical analysis of the available literature.

The term "metaheuristic" can be broken down as follows:

1. "Meta" indicates that it operates at a higher level, guiding the optimization process rather than directly solving the problem.
2. "Heuristic" refers to a problem-solving approach that uses to make informed decisions.

Metaheuristics represent a category of powerful optimization algorithms designed to solve complex, combinatorial, and multi-dimensional problems. They are characterized by their ability to explore vast solution spaces, often without the need for explicit problem-specific knowledge. Metaheuristics draw inspiration from natural phenomena, such as genetic processes, swarm behaviour, and simulated annealing, to work out intelligent and efficient strategies for locating optimal solutions or approximations. Some well-known metaheuristics include Genetic Algorithms (GAs), Particle Swarm Optimization (PSO), Simulated Annealing (SA), and Ant Colony Optimization (ACO).

Even though conventional feature selection techniques have proven useful in a variety of situations, there is rising interest in using the strength of metaheuristic algorithms to address feature selection problems. Complicated feature selection issues have shown potential to be successfully handled by metaheuristic-based methods. By swiftly and effectively exploring huge search spaces, these algorithms—which include genetic algorithm, particle swarm optimization, simulated annealing and ant colony optimization, among others—offer a distinctive viewpoint and an approach to tackle the problem of feature selection.

In conclusion, feature selection continues to be a crucial preliminary processing stage in machine learning, having a significant impact on the effectiveness and readability of models. These metaheuristic methods are expected to open up new options in the selection of features and are likely to significantly advance the rapidly developing area of machine learning. They sit at an intersection of statistical optimization and machine learning.

In the upcoming sections of this literature survey, we will delve deeper into the realm of metaheuristic-based feature selection. We will explore the principles underlying these algorithms, their applications in feature selection, and their comparative advantages over traditional methods. We evaluate their performance relative to traditional methods and discuss benchmark metrics and datasets. Real-world applications and case studies across diverse domains follow, showcasing the practical implications.

After doing this research survey, our major objectives are to gain knowledge about the previous research and identify any areas that require further investigation. We can determine what is currently popular in the area we are researching in this way. We want to employ and research these unique algorithms on a specific issue using this as a starting point. This aids in bridging the knowledge gap between what we learn in the classroom and how we may use it in the real world. In other words, we aim to make sure that our research can be applied to tackle real-world issues rather of being purely theoretical.

2. LITERATURE SURVEY

This section contains a summary of research papers surveyed and reviewed.

Shukla et al. (2020) [1] used metaheuristic search algorithms for selecting features on gene databases to reduce the dimensions of microarray datasets and remove redundant or useless attributes. They implemented feature selection using genetic algorithm, ant colony optimization, particle swarm optimization, differential evaluation. Implementation was done using performance of 2 different classifiers, support vector machine and Kth – Nearest Neighbour (KNN), also a supervised ML algorithm and Lasso regression as objective function. Experiments were performed on five commonly used biomedical gene expression datasets. Accuracy, mean number of features selected and execution time were tracked for each of them. Highest accuracy achieved was by GA in DLBCL dataset and lowest was 76.91% by ACO on colon cancer. Results showed that GA gave the best performance out of all the four metaheuristic algorithms, with achieving 83% average accuracy. Drawbacks of metaheuristic approach for feature selection, such as computation cost and scalability were also highlighted.

Hijazi et al. (2021) [2] have worked on applying metaheuristics to ensemble learning, to get predictions that are more precise and improve robustness and generalization. Due to challenge posed by dimensionality curse, we face difficulties with performance and time. A simultaneous heterogenic feature selection using ensemble learning was proposed. This method uses grey wolf optimizer, GA and particle swarm optimization and has distribution, parallel ensemble feature selection, collection and combination and testing phases. This approach was implemented on three types of machines namely CPU, parallel many core CPU and P.GPU. A total of 21 different datasets were used. The experimental analysis centered around accuracy, increase in processing speed, the number of chosen features, G-mean and fitness scores. Results obtained with number of cores = 8 were as follows: P.GPU performed best to very close to best based on average accuracy in 71.4% of total datasets and similarly in other performance measures. CPU and P.CPU also produced similar results, giving best results in 86% dataset in terms of average features chosen.

Amini et al. (2021) [3] proposed a dual-layer method, a hybrid combining wrapper and embedded techniques, for feature (predictors) selection to shrink the figure of predictors. In initial layer, as a wrapper, i.e., as a means to evaluate quality and impact of predictor subsets on predictions, Genetic Algorithm (GA) is combined with linear regression and used to reduce the number of predictors and obtain a subset of predictors, thereby reducing search space, computation cost and error. As GA is not guaranteed to give the most optimal solution, it is made up for in second layer. The second layer uses Elastic Net (EN) regularization to remove any remaining redundant parameters. EN regularization is a technique that combines both Lasso (L1) and Ridge (L2) regularizations to balance the trade-offs between the two and performs better on high-dimensional data compared to other regressions along with reducing search space. Hyperparameter optimization was performed to fine tune the parameter of GA and EN. A real dataset of Mazie genetic data was used and outcome showed that the proposed method reduced dimensions of feature space by 80% without reducing accuracy. It is also to be noted that though GA has advantage in reducing search space, it has a high computation cost on larger set of data.

Yusta (2009) [4] in her work has compared the sequential forward floating selection, sequential backward floating selection and genetic algorithm with the following three strategies: memetic algorithm, tabu search and greedy randomized adaptive search procedure to solve the feature selection problem to reduce dimensionality and search space. GRASP works iteratively and combines with randomization the local search and greedy approach to find a good quality solution. TS works on the idea of exploring further ahead of local minimum and marking it 'Tabu' to avoid cycling. MA is an evolutionary algorithm using the local search procedures to further improve the global results. These are usually used to solve optimization problems. In the experiments performed over six databases, k-nearest neighbour algorithm was used as classification algorithm. GRASP performed the best followed by TS, MA, GA respectively and SFFA and SBFA coming in the last with equal performance. The accuracy measured by objective function was between 0.838 and 0.911 with standard deviation(σ) ϵ (0.0266, 0.0666). The results indicate the GRASP and TS perform better compared to other algorithms.

Diao et al. (2015) [5] researched on using nature-inspired meta-heuristics (NIMs) for feature selection and classified these algorithms according to the sources that influenced them, along with providing pseudo-codes to facilitate comparison. To evaluate the effectiveness of these strategies, they carried out systematic experimentation using three distinct feature subset-based evaluators. Notably, competitive size reduction capabilities were established by approaches like HS, GA, MA, and PSO, with the first three giving comparable results. However, TS encountered difficulties in optimizing dependency scores for particular datasets. Although TS managed to get best results for 6 of 12 datasets, it failed in optimizing FRFS dependency scores. Classification outcomes varied according to chosen feature subsets, highlighting the complexity of the solutions. Although the work admits the possibility of sub-optimal solutions in some contexts and gives the benefit of unified representation and algorithm comparison, it lacks a statistical evaluation of scores. In conclusion, the paper provides an extensive analysis of nature-inspired meta-heuristics for feature selection, suggesting future research directions in this field.

Mafarja et al. (2020) [6] in this paper, present three hybrid models, HSGW, RSGW, and ASGW, that merge the grey wolf optimization and whale optimization algorithm to handle feature selection issues. The effectiveness of these models in feature selection is carefully tested on several UCI datasets, and they are compared to both individual optimizers (GWO and WOA) and cutting-edge feature selection methods (GA, BPSO, BGSA, and BGOA). The authors study the effects of size of population and the number of iterations on the capability of the hybrid models, finally evaluating their classification accuracy across several datasets. The results demonstrate the HSGW hybrid model's effectiveness, outperforming other methods in the selection of informative features and enhancing algorithm performance. HSGW gives best results for nearly 44.44% (8) of total datasets, followed by RSGW and ASGW giving best results for 38.88% and 11.11% of total datasets. The research highlights the capacity of the hybrid models to find an equilibrium between utilization and discovery while acknowledging the unique strengths and potential limits of the GWO and WOA algorithms. The HSGW model, in particular, succeeds in this research's introduction of novel hybrid meta-heuristic methodologies for selection of features. The research advocates the ongoing adoption of feature selection problems approached from a multi-objective perspective and recommends that future investigations explore posteriori multi-objective optimization methods.

L. Wang et al. (2014) [7] aim to evaluate the capability of different algorithms by generating a set of benchmark problems that imitate various biomedical feature selection scenarios. For the goal of feature selection, this paper's methodology makes use of metaheuristic optimization methods, like HS, GA, DE, ACO, PSO and QEA, that are inspired by natural processes. The authors created a series of benchmark problems that covered several dimensionalities and included important, correlated, unnecessary, redundant, and deceptive features, replicating various biomedical scenarios in order to assess how well these algorithms performed. The results of various algorithms were compared in-depth by the authors in order to assess how well each algorithm was able to choose the best feature subsets. The work highlights the examination and comparison of the effectiveness of meta-heuristics in the context of feature selection, even though it does not go into specifics about the outcomes obtained.

Talbi et al. (2008) [8] compared genetic algorithm and particle swarm optimization, two population-based metaheuristic algorithms, to select features to classify of multidimensional data, cancer datasets in this case. A mixed algorithm, Geometric Swarm Particle Optimization (GSPO) is used. GA and GPSO are combined with SVM in order to increase classification accuracy. For statistical robustness, GPSO and GA algorithms were independently applied 10 times to each dataset. According to their findings, GPSO and GA both produce beneficial results in gene selection, leading to smaller subgroups with excellent classification accuracy. While GA often performs well in generating optimal solutions, GPSO exhibits greater accuracy and consistency on average across numerous runs. Both of these algorithms have a classification rate higher than 86% on many subsets. It is also to be noted that 10% accuracy is achieved with Colon tumour dataset. These suggested algorithms’ performance surpasses the current approaches in terms of categorization rates and the number of genes selected. To perform at their best, they demand processing power and parameter adjustment. In conclusion, this report highlights the effectiveness of GPSO and GA enhanced with SVM for microarray data classification feature selection, presenting strong findings across several cancer datasets and sketching out potential directions for further study and improvement in this area.

Mafarja et al. (2017) [9] investigated the Binary Ant Lion Optimization (BALO) method with multiple transfer functions as the central focus of this research study. The research attempts to overcome local minima issues and achieve higher classification accuracy to enhance the capability of the original Ant Lion Optimization (ALO) method. The approach uses a wrapper feature selection technique, a Euclidean distance matrix for assessment, and the K-nearest neighbours (KNN) and BALO classifiers. The suggested BALO techniques are extensively tested against cutting-edge algorithms like gravitational search algorithm, particle swarm optimization, and fundamental ALO algorithms on standard datasets. The findings show that BALO techniques, notably ALO-V3 fitted with v-shaped transfer functions, exceed PSO and GA in performance in most of datasets in terms of categorization precision. This dominance is due to ALO's skill in navigating the feature space and capacity to find equilibrium between exploitation and discovery. The simplicity and cheap computing cost of these algorithms, according to the article, make them suitable for real-world issue resolution.

El-Kenawy et al. (2022) [10] introduces the GSDTO algorithm as a distinctive approach for detecting transformer issues in power networks through the utilization of dissolved gas analysis (DGA). By choosing the best dataset characteristics and using the GSDTO algorithm along with LSTM for classification, it seeks to improve diagnosis accuracy. The study is divided into two main sections: attribute selection, where the effectiveness of the GSDTO is evaluated in comparison to alternative methods, and model optimization, where chosen attributes are applied to improve the LSTM model. The dataset has 460 samples, and the findings are quite convincing, with the model surpassing rivals and obtaining a stunning 98.26% accuracy. While acknowledging drawbacks like computational complexity and parameter fine-tuning, the research provides advantages like enhanced diagnostic accuracy and automated attribute selection. In conclusion, this work offers a novel method for identifying transformer faults, but more investigation is required to fill in knowledge gaps and evaluate the GSDTO+LSTM algorithm's scalability and runtime properties.

Guha et al. (2021) [11] addresses the issue of feature selection in machine learning in order to efficiently and rapidly process large datasets. The Great Deluge Algorithm (GDA) is integrated with the Genetic Algorithm (GA) and its version, Deluge-based Genetic Algorithm (DGA), in the authors' innovative framework to improve exploitational capabilities. Their strategy comprises using GA and DGA for feature selection, and tests using different classifiers on 15 UCI datasets shows that DGA performs better than other methods by having higher accuracy and processing efficiency. It has been demonstrated that population-based metaheuristics like GA and DGA offer better solutions than filter and wrapper techniques. Although combining GA and GDA is advantageous, there are still research gaps in understanding the trade-offs in feature selection techniques and investigating how well the framework can be used to different datasets and classifiers.

Bhattacharyya et al. (2020) [12] research study, which includes high-dimensional microarray datasets for cancer classification and 18 conventional UCI datasets, demonstrates the superior performance of the MA-HS algorithm.. It outperforms 12 cutting-edge feature selection algorithms, picking the fewest features in 61% of situations while obtaining the best accuracy in 72% of the datasets. This accomplishment is credited to the successful fusion of HS's exploitational skills and MA's inquisitive qualities, which overcame obstacles including early convergence in MA. To increase the potential of mixed algorithms for discovery and utilization, further study is still required to investigate the possibilities of various meta-heuristic algorithms. Their experiments, conducted on model datasets and multidimensional microarray datasets for cancer classification, verify the MA-HS algorithm's better performance. It outperforms 12 cutting-edge feature selection algorithms, achieving the best accuracy in 72% of the datasets while choosing the least attributes in 61% of cases. This success is attributed to the effective combination of MA's exploratory strengths and HS's exploitational prowess, overcoming challenges such as premature convergence in MA.

Sayed G. I. et al. (2019) [13] aim to develop a model for early prediction of drug toxicity during the initial stages of pharmacological development. The model employs a three-phase, organized methodology that includes pre-processing of the data, feature selection using a unique CDA technique, and classification using an SVM classifier. The 553 medicines in the under-consideration dataset each have 31 characteristics. The SMOTE is used in data pre-processing to alleviate class imbalance. A thorough comparison of several metaheuristic optimization techniques is required for the research. The outcomes illustrate the resilience and efficiency of the model, particularly emphasizing the superiority of the ROS algorithm. The research does not go into particular restrictions or drawbacks, although the approach has benefits including speed and efficiency in predicting medication toxicity. The importance of selecting features and metaheuristic optimization for early drug toxicity prediction science development is highlighted by this study's findings.

Sindhu R. et al. (2017) [14] endeavours to enhance the FS by constructing a procedure that It reduces the size of the feature set, improves the precision of classification, and streamlines the set of features. The study explores the landscape of FS techniques, grouping them into three crucial categories: filters, wrappers, and hybrid methods, each of which offers a distinctive strategy for tackling the complex problem of choosing the most important features. The research makes use of the capabilities of meta-heuristic search algorithms, like GA, which are useful tools for handling challenging optimization problems. The ISCA algorithm, a noteworthy development that utilizes the definitions of the functions of cosine and sine to improve feature optimization, is introduced in this study. This unique method navigates the feature space by combining the traits of these trigonometric functions, enabling a careful selection of the most useful qualities. The results of the study presented in this work confirm the effectiveness of the ISCA algorithm. It expertly cuts down on the number of feature subsets while establishing a remarkable balance between high precision in classification and the latest methods of feature reduction. Additionally, ISCA distinguishes itself by demonstrating the quickest processing time needed for this complex procedure, highlighting its effectiveness in real-world applications. This research creates new pathways for improved feature selection strategies by fusing mathematical inventiveness and the strength of meta-heuristic algorithms.

Kareem S. S. et al. (2022) [15] in this research have combined the Gorilla Troop Optimization, based on lifestyle of gorillas in forest and Bird Swarms algorithm, based on birds' nature, to improve selection of attributes for detection of security breaches int IoT. The difficulties presented by duplicate and pointless data in IoT applications are the focus of this hybrid approach. The tests performed on four datasets for IoT intrusion detection show that the GTO-BSA method beats a number of currently available metaheuristic methods. It achieves astounding accuracy rates of 95.5%, 98.7%, 81.5%, and 81.5% on tested datasets. This method shows how metaheuristic algorithms' dynamic search behaviour and global exploration ability can be used to enhance feature selection. An equilibrium between exploitation and discovery is important in the hybrid method. The study concludes by presenting a promising feature selection approach for IoT intrusion detection, with room for further investigation across diverse domains, multi-objective problem-solving, and hyperparameter optimization of machine learning.

Das, A. et al. (2020) [16] have conducted this study with the aim of designing an algorithm to classify languages in India, encompassing MFCC and LPC features, and harnessing the Late Acceptance Hill-Climbing and BBA algorithms. The main goal is to improve the effectiveness of conventional spoken language detection techniques, particularly in India's multilingual environment. The methodology entails preprocessing of the audio signal, feature extraction from the MFCC and LPC datasets, and implementation of the hybrid BBA-LAHC algorithm to enable intelligent feature selection from a set of about 1000 features. Languages were categorized using well known classifiers. The tests performed on two Indian language databases show that the suggested feature selection algorithm performs better than earlier techniques. It accurately distinguishes between Indian languages, with the best performance being equivalent to 92.3% and 91.5% for the specified set of features and unprocessed feature set, respectively, for the Random Forest classifier.

Dey et al. (2020) [17] in their research paper introduce a novel approach, GREO, for addressing the challenge of emotion recognition in speech. This method combines the optimization algorithms to enhance classification accuracy in SER. The approach of the study includes describing of dataset, prior processing, extraction of features from the dataset, GREO-based selection of features and categorization using the XGBoost classifier. Comparative experiments conducted on standard SER datasets reveal that GREO outperforms other feature selection algorithms in terms of precision of classification while selecting a reduced set of optimal features, achieving nearly 98% accuracy in one of the datasets and around 97% in other. However, the paper does not elaborate on the potential limitations of the GREO method. In conclusion, the study underscores the effectiveness of GREO in improving SER accuracy, achieving state-of-the-art results, and serving as a promising hybrid meta-heuristic FS approach for SER.

Rostami et al. (2021) [18] in their research paper address a notable gap in coverage and categorizing these methods. The authors conduct an extensive literature review, evaluating 85 papers with relevant keywords and utilizing filter and wrapper models for feature assessment. Through experiments comparing different swarm intelligence-based techniques using classifiers like SVM, NB, and AB, the study emphasizes varying performance outcomes across classifiers. The accuracy obtained by algorithms was less on NB compared to SVM and AB. ACO performed best on all classifiers, followed by PSO and others. While these methods excel in optimizing conflicting objectives and balancing exploration and exploitation, the challenge lies in parameter selection. Overall, the paper provides a thorough overview of selection of features using swarm intelligence, highlighting its pros, cons, and the significance of parameter optimization for optimal results.

El-Kenawy et al. (2020) [19] in their research paper introduces a novel hybrid optimization approach, GWO-PSO, designed to address the challenge of feature selection in the context of big data analysis. By leveraging the inherent strengths of both Gray Wolf and PSO, the study seeks to achieve harmony in the optimization process between exploitation and discovery. Through comprehensive experiments on diverse datasets, the authors demonstrate the usefulness of GWOPSO in enhancing feature selection. The selected feature subset leads to notable improvements in classification performance, particularly with the K-Nearest Neighbour (KNN) classifier. While this hybrid approach presents advantages in terms of feature selection efficacy, it does come with potential computational resource and time requirements. The research highlights the potential of GWOPSO as a useful tool for choosing features in the context of the analysis of big data, while also outlining potential directions for further investigation, such as testing on more complicated datasets and investigating parallel variations of the GWOPSO model.

Chen et al. (2012) [20] in their research paper present a methodological framework that amalgamates PS optimization and the 1-NN method for the objective of feature selection in domain of data mining, particular emphasis on its use in identifying obstructive sleep apnea (OSA). PSO, a stochastic optimization technique, is employed to iteratively fine-tune feature selection by simulating the social dynamics of particles. In conjunction, the 1-NN method, a nearest neighbour classification algorithm, is utilized to enhance the classification process. Through rigorous experimentation with life science datasets and comparisons against benchmark algorithms like BPNN, LR, SVM, and C4.5. The research illustrates the superior classification accuracy of the offered method. The paper concludes by highlighting the potential of this methodology for identifying critical factors and facilitating the diagnosis of medical conditions while acknowledging areas for further investigation in mitigating limitations associated with PSO.

Fong et al. (2014) [21] in their research paper examines the suitability of Swarm Search Feature Selection (SS-FS) techniques for biomedical data classification. The study employs a fusion of three metaheuristic algorithms and three established classification methodologies to form the SS-FS methods. Experimental evaluations, conducted on the Arrhythmia and MicroMass datasets reveal that the majority of SS-FS methods surpass feature selection with regard to similarity and non-feature selection approaches in terms of minimizing error rates. However, the performance of SS-FS methods is contingent on the specific classification algorithm employed, with Navies Bayes integration demonstrating comparatively less favourable outcomes. While the correlation-based feature selection methods offer marginal enhancements, non-feature selection methods exhibit elevated error rates. In conclusion, this research underscores the potential of SS-FS methods for effective high-dimensional biomedical data classification, with scope for further refinement through algorithm-specific optimization.

Sayed et al. (2018) [22] in their research paper introduces the CSSA, a novel hybridization approach aimed at augmenting the effectiveness of meta-heuristic algorithms, with a specific focus on the Salp Swarm Algorithm (SSA), through the integration of principles derived from chaos theory. Employing an extensive array of experimental investigations and evaluative criteria, CSSA exhibits noteworthy advantages over SSA and several established meta-heuristic algorithms. These advantages encompass higher mean fitness values, enhanced algorithmic stability, and an improved capacity for both exploration and exploitation within the optimization domain. Remarkably, CSSA showcases remarkable proficiency across diverse optimization challenges and feature selection tasks, effectively mitigating the pitfalls typically associated with local optima. However, it is worth acknowledging certain potential limitations, including the trade-off of slower convergence speed and susceptibility to local optima, which warrant further scholarly exploration and inquiry.

Nirmala Sreedharan et al. (2018) [23] in their research paper focuses on the enhancement of Facial Emotion Recognition (FER) systems through an optimized feature extraction and classification methodology. A novel bio-inspired algorithm, the GWO, is introduced to address the limitations of existing FER methods. The proposed methodology entails preprocessing facial images, optimized SIFT feature extraction, and GWO-based feature selection. Subsequently, emotions are recognized using the selected features and a GWO-Neural Network (NN)-based classification approach. Comparative evaluations against conventional FER methods demonstrate the superior performance of the proposed GWO-NN-based FER system, achieving commendable classification accuracies for different databases. While the GWO algorithm exhibits notable convergence speed and fine-tuning capabilities, certain potential limitations, such as inefficiencies in transverse orientation, merit further investigation. In conclusion, this research presents a promising avenue for advancing FER systems through the integration of GWO and optimized feature extraction techniques.

Gharehchopogh et al. (2022) [24] in their research paper is dedicated to the field of author identification, with a specific emphasis on stylometric analysis for discerning and attributing unique writing styles to different authors. The primary objective lies in the application of linguistic analysis to manuscripts, thereby facilitating the identification of authors and their distinctive textual patterns. The paper underscores the pivotal role of authorship identification in the prevention of plagiarism and the safeguarding of copyrighted content. The methodological framework comprises the computation of word frequencies and normalized frequencies across the entire corpus, accompanied by the integration of metaheuristic algorithms for feature selection. Furthermore, a customized associative classification approach is applied to tackle the authorship attribution challenge, with an innovative incorporation of chaotic maps to enhance exploration during optimization. Empirical validation, based on a substantial Arabic dataset encompassing texts from ten distinct authors, reveals the efficacy of the proposed method, achieving a commendable accuracy rate of 97.43%. Comparative analysis with other algorithms further underscores the superior performance of the devised approach, providing valuable insights into the domain of multilingual authorship discrimination and author identification methodologies.

M.M. Mafarja et al. (2017) [25] worked on analysing the effectiveness of various optimization techniques for feature selection. They evaluated the WOA algorithm's performance on various datasets against two modified versions, WOASAT-1 and WOASAT-2 (Whale Optimization Algorithm with Tournament Selection). WOASAT-2 regularly achieves 92.2% accuracy on average, which is better than the other methods. With an accuracy of 82.9%, the closest approach shows that WOASAT-2 is more capable of exploring and can search in high-performing areas of the feature space. Also, WOASAT-2 outperforms other optimizers for a greater number of characteristics. It reduces the search space by eliminating ones that are extra or unnecessary. WOASAT-2 uses the tournament selection process, lowering the likelihood of getting stuck in local maximum and allowing weak solutions to be selected. The findings demonstrate that WOASAT-2 is a superior method for feature selection tasks since it sets fewer characteristics and achieves high accuracy. The study shows the WOA algorithm's efficiency and ability to solve feature selection challenges by achieving balance between discovery and extraction.

Elgamal et al. (2022) [26] did a comparative analysis of several optimization algorithms for feature selection. The GOA, Genetic Algorithm (GA) and others were assessed. 14 sets of medical data were used for the assessment. As assessment measures for their analysis, they employed the quantity of features, precision of classification, fitness metrics, P-value, and convergence rate. The capacity of each method to choose the best feature subset and get high classification accuracy was evaluated. According to the results, the suggested algorithm was more accurate than other algorithms. CHHO is an enhanced version of the HHO (Harris Hawks Algorithm) algorithm using simulated annealing. On the medical datasets, it obtained an impressive average accuracy rate of 86.4%. The results showed how well the CHHO algorithm achieves high classification accuracy and how feature selection procedures in medical data analysis could benefit from its application.

In the study conducted by Guha et al. (2023) [27] several classification models were compared based on their accuracy. Various datasets, such as binary optimization problems and Microarray datasets, were used for evaluating the models. Seven publicly accessible Microarray datasets were taken into consideration. It was found that the hybrid version of the Equilibrium Optimizer method, DEOSA (Discrete Equilibrium Optimizer -Simulated Annealing), had outstanding outcomes. Six of the seven Microarray datasets had a classification accuracy of 100% achieved by DEOSA. In addition, compared to the total features in the datasets, DEOSA chose a relatively small quantity of features. The effectiveness of PBGSK (Binary Gaining Sharing Knowledge with Population reduction approach) and DEOSA were compared in the context of binary optimization problems. Five runs of each algorithm were conducted, and the mean accuracy and total counts of characteristics were noted. The outcomes demonstrated that DEOSA's use of low-dimensional feature vectors allowed it to achieve high accuracy. The study's findings showed how well DEOSA achieves high classification accuracy while using fewer feature dimensions.

Rostami et al. (2021) [28] discussed the optimization of parameter values in feature selection methods for classification tasks. To select the best values for the parameters, the authors suggested a method of parameter optimization based on Bayesian theory. Each set of parameter values was employed to measure the correctness of the model. The study included the results of tests done with KNN, SVM, and AdaBoost classifiers on various datasets. Standard deviations for the classification accuracy rates were also provided. In the majority of datasets, the CDGAFS approach continuously produced the best accuracy. The sensitivity analysis of the ω and θ parameters showed that adjusting them to specific values improves the classification accuracy. In comparison to previous feature selection techniques, the suggested method's computing complexity was evaluated, and it was found to be quicker and more effective. The correctness of the suggested method was 1.20% higher than ACO, and 1.57% higher than the ABC algorithm.

Taradeh et al. (2019) [29] studied the performance of the HGSA in comparison with other feature selection (FS) approaches. For a fair comparison, the algorithms were developed under identical settings and conditions. Fitness values, the number of chosen features, and categorization correctness were used to compare the various methods. The outcomes showed that, for the majority of datasets, HGSA performed more accurately than the other methods. HGSA outperformed GWO in 88.8% of the datasets using the KNN classifier, while GWO outperformed only 11.1%. Similarly, in 83.3% of the datasets, HGSA outperformed rest of the algorithms in the case of the Decision Tree classifier. HGSA's average accuracy varied based on the classifier and dataset and was between 0.770 to 1.0. The mean fitness values varied from 0.005 - 0.237, while the mean number of features selected varied from 0 - 43.03. The stability of HGSA's performance is indicated by its lower standard deviation numbers. HGSA's statistical significance was validated by the Wilcoxon test p-values, the majority of which were less than 0.05.

In their work, Ahmed et al. (2020) [30] proposed that the RTHS (Ring Theory based Harmony Search) algorithm, compared to other algorithms using the Naive Bayes classifier, performed better regarding classification accuracy. Moreover, for 72.22% of the collection, the RTHS algorithm chooses the fewest features. The study also covered the application of the RTEA meta-heuristic technique to improve the Harmony Search (HS) algorithm. This method attempts to satisfy the precision and time requirements for problem solving. By improving the exploration and exploitation processes, global optima are reached because there is a lower dependence on the starting values of HMCR. Three well-known classifiers are used to evaluate the RTHS algorithm: Kth-NN, Random Forest, and NB. The FS approaches are used for the trained data to identify the important features that make up the ideal feature subset. RHTS performed the best regarding classification precision, with 88.9% accuracy in 16 cases.

In their paper, Ghost et al. (2020) [31] provided information about the technical details and accuracy achieved by different models. Eighteen UCI datasets were evaluated to the proposed BSF and AβBSF. Various amounts of attributes and instances were present in both two-class and many-class datasets. The models proved their feature selection (FS) efficiency in 11 out of the 18 datasets, with a classification accuracy of > 90%. Various population sizes and maximum iteration counts were tested during the parameter tuning procedure. Through thorough experimentation, a population size of 20 was determined to be suitable for further experiments. Every iteration's fitness function value was also monitored. These methods use meta-heuristic algorithms, which are becoming increasingly popular because of their flexibility and ability to avoid local optima. These algorithms assess feature subsets and determine which feature subset is most suited for the given task.

Sayed et al. (2019) [32] discussed the CCSA (Chaos-based CSA) feature selection algorithm. The algorithm uses various fitness functions to measure the performance of algo. Using formulas that take into account the feature count in the original dataset, the best score thus far, and other factors, the fitness functions are defined mathematically. Additionally, the algorithm uses a fitness function that uses a weighted factor to combine the number of picked characteristics with classification accuracy. The optimal solution is found by maximizing the fitness function. Using 20 standard datasets, the CCSA feature selection technique was verified. The findings demonstrated that, in regards to best and average fitness values, the CCSA algorithm performs better than other meta-heuristic algorithms. Utilizing a small selection of characteristics significantly improved classification accuracy while cutting down on calculation time as well as memory use. Achieving 100% accuracy in some cases required less computational time. The CCSA feature selection algorithm's efficiency was demonstrated by the obtained findings.

Ali et al. (2017) [33] in their study have proposed the HPSOGA algorithm, a cutting-edge method to reduce molecular potential energy functions. It deals with the problem of locating global minima for complex compounds. Utilizing exploration and exploitation, dimensionality reduction, population partitioning, and genetic mutation, HPSOGA integrates both the GA and PSO. The algorithm's effectiveness and higher performance in comparison to benchmark and both the traditional algorithms are confirmed by numerical experiments on scalable molecular potential energy functions, which was then compared to that of other algorithms and the standard PSO. Results reveal that HPSOGA has potential for effectively resolving challenging global optimization issues. As shown by its higher performance when compared to traditional PSO and benchmark algorithms, HPSOGA is a strong hybrid strategy that balances exploitation and exploration in optimization problems.

3. CONCLUSION

The comprehensive analysis of the literature on metaheuristic-based feature selection has revealed a wide range of potential applications. This thorough analysis of research publications has expanded our understanding and highlighted important issues that require future study.

The surprising variety of metaheuristic algorithms used in feature selection is one of the main lessons learned from this survey. The diversity of algorithms highlights their adaptability, ranging from cutting-edge approaches like Grey Wolf Optimization (GWO) and Whale Optimization Algorithm (WOA) to more established techniques like Genetic Algorithms (GAs) and Particle Swarm Optimization (PSO). Due to their versatility, these algorithms stand out as effective tools in the feature selection toolbox and may be used for a variety of different problem domains. Further investigation into the relative efficacy of these algorithms in various situations may result in the creation of unique hybrid strategies that combine the advantages of several algorithms.

The multidisciplinary character of feature selection while using metaheuristic techniques has also been highlighted by this survey. These techniques have been successfully used by researchers from a variety of disciplines, including biology, power systems, intrusion detection, and speech-emotion recognition. This cross-domain adaptability highlights the adaptability of feature selection based on metaheuristics in dealing with a variety of real-world difficulties. Researchers can now investigate how these methods can be adapted to certain domains for more efficient feature selection.

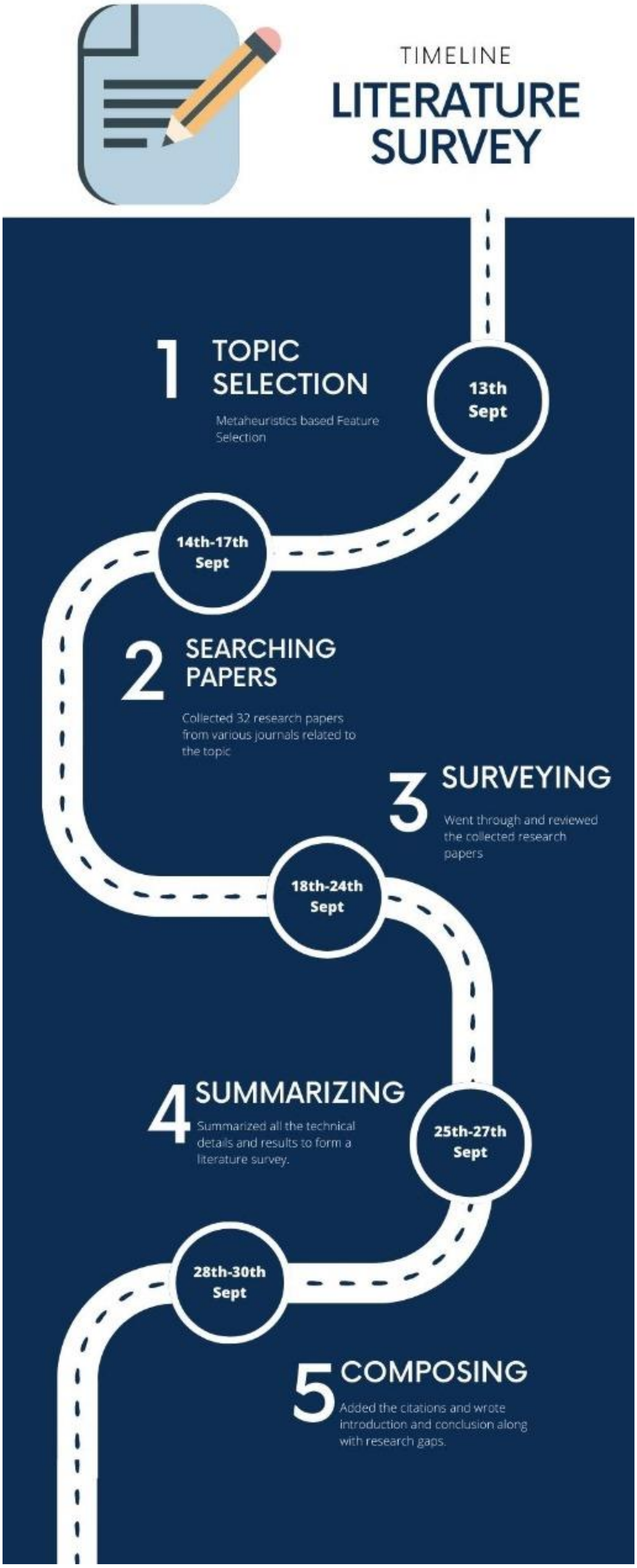
It is very noteworthy to make improvements to the scalability and effectiveness of feature selection metaheuristic algorithms. Much of the research that was surveyed was concerned with methods to increase the computational effectiveness of these approaches. Many strategies have been investigated, including parallelisation, hybridisation with complementary optimisation techniques, and algorithm parameter adjustment. Particularly when working with huge and complex datasets, these advancements are essential. Future studies should keep looking for novel ways to improve the effectiveness and usability of metaheuristic-based feature selection.

The report also highlights ongoing investigations into adaptive feature selection methods. These algorithms show promise in handling various and changing data sources since they can dynamically change their methods and parameters depending on the properties of datasets.

We have reached a critical choice as a result of our research survey journey: we will put the information we have learned in the field of cancer detection to use. Cancer is still a serious problem in today's society, and late-stage patients frequently don't respond well to the existing treatments. We are motivated to take action since there is a substantial research void in this area, especially because early-stage cancer is when therapies are most successful.

We have decided to use our discoveries about Genetic Algorithm (GA) and Particle Swarm Optimization (PSO) to tackle this problem. In the field of cancer diagnosis, these two potent metaheuristic algorithms are crucial. They are excellent for maximizing model parameters, ensembles, and feature selection. They are extremely important because of their capacity to manage complicated, high-dimensional data and adjust to changing trends. We seek to improve diagnostic precision and interpretability by utilizing GAs and PSO, making a substantial contribution to the continuing fight against cancer.

Our research journey, which began as a quest to understand the dynamics of feature selection through metaheuristic algorithms, has evolved into a commitment to the real-world application of this knowledge. We are dedicated to making a tangible impact on a critical issue that affects countless lives, further underscoring the practical significance of our research survey.



TIMELINE

LITERATURE SURVEY

1 TOPIC SELECTION

Metaheuristics based Feature Selection

13th Sept

14th-17th Sept

2 SEARCHING PAPERS

Collected 32 research papers from various journals related to the topic

3 SURVEYING

Went through and reviewed the collected research papers

18th-24th Sept

4 SUMMARIZING

Summarized all the technical details and results to form a literature survey.

25th-27th Sept

28th-30th Sept

5 COMPOSING

Added the citations and wrote introduction and conclusion along with research gaps.

Bibliography

[1] A. K. Shukla, D. Tripathi, B. R. Reddy and C. D., "A study on metaheuristics approaches for gene selection in microarray data: algorithms, applications and open challenges," *Evolutionary Intelligence,* vol. 13, pp. 309-329, 2020.

[2] N. M. Hijazi, H. Faris and I. Aljarah, "A parallel metaheuristic approach for ensemble feature selection based on multi-core architectures," *Expert Systems with Applications,* vol. 182, p. 115290, 2021.

[3] F. Amini and G. Hu, "A two-layer feature selection method using genetic algorithm and elastic net," *Expert Systems with Applications,* vol. 166, p. 114072, 2021.

[4] S. C. Yusta, "Different metaheuristic strategies to solve the feature selection problem," *Pattern Recognition Letters,* vol. 30, no. 5, pp. 525-534, 2009.

[5] R. Diao and Q. Shen, "Nature inspired feature selection meta-heuristics," *Artificial Intelligence Review,* vol. 44, pp. 311-340, 2015.

[6] M. Mafarja, A. Qasem, A. A. Heidari, I. Aljarah, H. Faris and S. Mirjalili, "Efficient hybrid nature-inspired binary optimizers for feature selection," *Cognitive Computation,* vol. 12, pp. 150-175, 2020.

[7] L. Wang, H. Ni, R. Yang, V. Pappu, M. B. Fenn and P. M. Pardalos, "Feature selection based on meta-heuristics for biomedicine," *Optimization Methods and Software,* vol. 29, no. 4, pp. 703-719, 2014.

[8] E. G. Talbi, L. Jourdan, J. Garcia-Nieto and E. Alba, "Comparison of population based metaheuristics for feature selection: Application to microarray data classification," in *IEEE/ACS International Conference on Computer Systems and Applications*, Doha, 2008.

[9] M. Mafarja, D. Eleyan, S. Abdullah and S. Mirjalili, "S-shaped vs. V-shaped transfer functions for ant lion optimization algorithm in feature selection problem," in *In Proceedings of the international conference on future networks and distributed systems*, 2017.

[10] E. S. M. El-kenawy, F. Albalawi, S. A. Ward, S. S. Ghoneim, M. M. Eid, A. A. Abdelhamid and A. ... Ibrahim, "Feature selection and classification of transformer faults based on novel meta-heuristic algorithm," *Mathematics,* vol. 10, no. 17, p. 3144, 2022.

[11] R. Guha, M. K. S. Ghosh, S. Shaw, S. Mutsuddi, V. Bhateja and R. & Sarkar, "Deluge based genetic algorithm for feature selection," *Evolutionary intelligence,* vol. 14, pp. 357-367, 2021.

[12] T. Bhattacharyya, B. Chatterjee, P. K. Singh, J. H. Yoon, Z. W. Geem and R. Sarkar, "Mayfly in harmony: A new hybrid meta-heuristic feature selection algorithm," *IEEE Access,* vol. 8, pp. 195929-195945, 2020.

[13] G. I. Sayed, A. Tharwat and A. E. Hassanien, "Chaotic dragonfly algorithm: an improved metaheuristic algorithm for feature selection," *Applied Intelligence,* vol. 49, pp. 188-205, 2019.

[14] R. Sindhu, R. Ngadiran, Y. M. Yacob, N. A. H. Zahri and M. Hariharan, "Sine–cosine algorithm for feature selection with elitism strategy and new updating mechanism," *Neural Computing and Applications,* vol. 28, pp. 2947-2958, 2017.

[15] S. S. Kareem, R. R. Mostafa, F. A. Hashim and H. M. El-Bakry, "An effective feature selection model using hybrid metaheuristic algorithms for iot intrusion detection," *Sensors,* vol. 22, no. 4, p. 1396, 2022.

[16] A. G. S. Das, P. K. Singh, A. Ahmadian, N. Senu and R. Sarkar, "A hybrid meta-heuristic feature selection method for identification of Indian spoken languages from audio signals," *IEEE Access,* vol. 8, pp. 181432-181449, 2020.

[17] A. Dey, S. Chattopadhyay, P. K. Singh, A. Ahmadian, M. Ferrara and R. Sarkar, "A hybrid meta-heuristic feature selection method using golden ratio and equilibrium optimization algorithms for speech emotion recognition," *IEEE Access,* vol. 8, pp. 200953-200970, 2020.

[18] M. Rostami, K. Berahmand, E. Nasiri and S. Forouzandeh, "Review of swarm intelligence-based feature selection methods," *Engineering Applications of Artificial Intelligence,* vol. 100, p. 104210, 2021.

[19] E. S. El-Kenawy and M. Eid, "Hybrid gray wolf and particle swarm optimization for feature selection," *Int. J. Innov. Comput. Inf. Control,* vol. 16, no. 3, pp. 831-844, 2020.

[20] L. F. Chen, C. T. Su, K. H. Chen and P. C. Wang, "Particle swarm optimization for feature selection with application in obstructive sleep apnea diagnosis," *Neural Computing and Applications,* vol. 21, pp. 2087-2096, 2012.

[21] S. Fong, S. Deb, X. S. Yang and J. Li, "Feature selection in life science classification: metaheuristic swarm search," *IT Professional,* vol. 16, no. 4, pp. 24-29, 2014.

[22] G. I. Sayed, G. Khoriba and M. H. Haggag, "A novel chaotic salp swarm algorithm for global optimization and feature selection," *Applied Intelligence,* vol. 48, pp. 3462-3481, 2018.

[23] N. P. Nirmala Sreedharan, B. Ganesan, R. Raveendran, P. Sarala, B. Dennis and R. Boothalingam R, "Grey wolf optimisation-based feature selection and classification for facial emotion recognition," *IET Biometrics,* vol. 7, no. 5, pp. 490-499, 2018.

[24] F. S. Gharehchopogh, I. Maleki and Z. A. Dizaji, "Chaotic vortex search algorithm: metaheuristic algorithm for feature selection," *Evolutionary Intelligence,* vol. 15, no. 3, pp. 1777-1808, 2022.

[25] M. M. Mafarja and S. Mirjalili, "Hybrid whale optimization algorithm with simulated annealing for feature selection," *Neurocomputing,* vol. 260, pp. 302-312, 2017.

[26] Z. M. Elgamal, N. B. M. Yasin, M. Tubishat, M. Alswaitti and S. Mirjalili, "An improved harris hawks optimization algorithm with simulated annealing for feature selection in the medical field," *IEEE access,* vol. 8, pp. 186638-186652, 2020.

[27] R. Guha, K. K. Ghosh, S. K. Bera, R. Sarkar and S. Mirjalili, "Discrete equilibrium optimizer combined with simulated annealing for feature selection," *Journal of Computational Science,* vol. 67, p. 101942, 2023.

[28] M. Rostami, K. Berahmand and S. Forouzandeh, "A novel community detection based genetic algorithm for feature selection," *Journal of Big Data,* vol. 8, no. 1, pp. 1-27, 2021.

[29] M. Taradeh, M. Mafarja, A. A. Heidari, H. Faris, I. Aljarah, S. Mirjalili and H. Fujita, "An evolutionary gravitational search-based feature selection," *Information Sciences,* vol. 497, pp. 219-239, 2019.

[30] S. Ahmed, K. K. Ghosh, P. K. Singh, Z. W. Geem and R. Sarkar, "Hybrid of harmony search algorithm and ring theory-based evolutionary algorithm for feature selection," *IEEE Access,* vol. 8, pp. 102629-102645, 2020.

[31] K. K. Ghosh, S. Ahmed, P. K. Singh, Z. W. Geem and R. Sarkar, "Improved binary sailfish optimizer based on adaptive β-hill climbing for feature selection," *IEEE access,* vol. 8, pp. 83548-83560, 2020.

[32] G. I. Sayed, A. E. Hassanien and A. T. Azar, "Feature selection via a novel chaotic crow search algorithm," *Neural computing and applications,* vol. 31, pp. 171-188, 2019.

[33] A. F. Ali and M. A. Tawhid, "A hybrid particle swarm optimization and genetic algorithm with population partitioning for large scale optimization problems," *Ain Shams Engineering Journal,* vol. 8, no. 2, pp. 191-206, 2017.