CSF 407 – ARTIFICIAL INTELLIGENCE

Optimizing Feature Selection in Medical Classification

USING

Metaheuristic Search Algorithms

TEAM MEMBERS ID	NAMES
2021A7PS3116H	SIDDHARTH YAYAVARAM
2021A7PS3117H	ARNAV YAYAVARAM

GROUP ID - 37

Introduction

In the realm of real-world problem-solving, a common challenge emerges: the management of vast datasets with numerous attributes or features. Negotiating this data labyrinth can be an intricate task. Within these datasets lie attributes, some of which may be redundant, or irrelevant, all of which can significantly impede the performance of machine learning models.

Dimensionality too poses issues, potentially leading to overfitting and prolonged learning times. The crux of the matter is to reduce the dimensionality of these datasets while retaining their predictive accuracy. This is where the concept of feature reduction comes into play.

Feature reduction, branches into two key facets: feature construction and feature selection. Feature construction entails the creation of novel features derived from the original dataset, while feature selection(FS) involves choosing best inputs. We mainly focus on feature selection.

When dealing with a x features, 2^x subsets become potentially relevant. As x grows larger, assessing how good the model is for all the small partitions is very impractical. Various methodologies have been created to deal with this. However, many of these methods suffer from issues such as premature convergence, unnecessary complexity, and high computational costs. Consequently, the spotlight has shifted towards metaheuristic algorithms as a means of addressing these challenges.

These algorithms offer efficiency and effectiveness, capable of finding optimal feature subsets while preserving accuracy. Pruning features via these algorithms mitigates overfitting, reduces vulnerability to missing data, and facilitates better model interpretation and generalization.

For instance, in image analysis, feature selection helps discern the salient visual components, such as pixels and colours.

Mathematically, the FS task can be seen as :

Given: input data 'K' containing 'A' features.

Goal: to select the most relevant features,

Consider K = {p1, p2, …, pA}, the objective is to extract a subset Y = {p1, p2, ..., px},

where x < A, and p1, p2, p3, …, px represent the properties of the input.

Datasets can come in all shapes and sizes, we explore the task for datasets comprising of large number of images.

Feature extraction for images encompasses the derivation of higher-level features from raw data, such as edges, corners, or image segments. Some lower-level primary features include size, colour values, intensity.

Yet, the inherent challenge in image datasets lies in their vastness, as they often comprise thousands of features due to the multitude of pixels. The quality of an image is directly correlated with the number of pixels, which, in turn, amplifies computational complexity.

Feature selection for images can be achieved through wrapper methods, filters, and embeddings. Wrapper methods evaluate feature subsets by training models, making them suitable when labelled data is available. Filters, on the other hand, select features based on statistical criteria or domain knowledge, making them computationally efficient.

Embeddings transform high-dimensional image data into lower dimensions, often using techniques like PCA (Principal Component Analysis) or other embeddings. The choice depends on the task, dataset size, and computational resources, with a combination of methods often yielding the best results. Pre-trained networks are commonly used for extracting image embeddings, offering a powerful feature selection approach.

Given the size of the feature space, we can see FS as a problem we must try efficiently solve, thereby making it an ideal candidate for the application of metaheuristic-based optimizers. These algorithms have demonstrated their prowess in solving complex optimization challenges.

Specifically, we look at nature inspired algorithms, where the behaviour of the animal kingdom is emulated and serves as a guide to create algorithms. We borrow from the millions of years of evolutionary strategies which have been passed on from generation to generation, adapting to the changes in the environment every step of the way.

A growing area of research pertains to the use of image classification for disease detection, particularly through X-ray and CT-scan images. Over the past few years, metaheuristic algorithms have gained traction for optimizing deep neural network architectures. Their simplicity, flexibility, and problem-agnostic nature render them ideal candidates for tackling intricate refinements.

Metaheuristics excel at finding the best or close to the best answers for a whole slew of issues. They are known for their lack of reliance on derivatives and their stochastic nature. They initiate their optimization process by generating random solutions, obviating the need for derivative calculations typical of gradient-based techniques which are usually seen in other deep learning algorithms. What sets metaheuristic algorithms apart is their simplicity and flexibility, making them open to customization for specific problems. An amazing ability of these algorithms is to avoid converging to a solution unnecessarily quickly, a pitfall often associated with optimization. Their stochastic nature treats the problem as a black box, enabling them to iterate through all the search states thoroughly and evade local optima. Metaheuristics strike a delicate balance between exploration and exploitation, investigating promising search spaces before delving into local searches in the regions identified during exploration.

To delve into the metaheuristic algorithms applied in feature selection, we research into the realm of binary vector representations. In this paradigm, each solution vector is represented as a sequence of binary values, such as (1010101001...). Here, a '1' indicates the selection of a particular feature, while '0' signifies its exclusion from the subset. This binary representation forms the foundation for metaheuristic algorithms striving to identify optimal feature subsets.

The algorithms are either cooperative where collective intelligence is utilized to find a solution or competitive where only the best solutions “survive” each iteration.

We briefly explain a high-level overview of the common structure and iterative nature of many of the algorithms followed in the research papers we analyse.

The algorithms typically start by initializing a population of potential solutions. In each iteration, a subset of solutions is selected based on their fitness, modified to explore and exploit the solution space, and then evaluated for their fitness.

The best solutions are retained, and the process repeats until termination criteria are met. This is analogous to the principle of “Survival of the Fittest” proposed by Charles Darwin. Only the animals best suited to fully exploit the environment are allowed to survive.

These algorithms strike a balance between exploration and exploitation to efficiently search for optimal or near-optimal solutions, drawing inspiration from animal behaviours like foraging, flocking, or genetic evolution.

Although specific mechanisms may vary among algorithms, this iterative blueprint is a fundamental characteristic they share.

Another topic very important in the field of artificial intelligence is how the ethical considerations in feature selection are paramount in ensuring the responsible and fair use of machine learning models. These considerations encompass the mitigation of bias and the promotion of fairness in feature selection, transparency in the decision-making process, and the protection of privacy and data security.

Ethical and explainable artificial intelligence is a field whose importance grows daily with the advent of AI in our daily lives.

Moreover, ethical feature selection is an ongoing process that necessitates continuous monitoring and adaptation to evolving ethical concerns and societal norms. Ultimately, integrating these ethical principles into the feature selection process is fundamental to building models that are both effective and morally sound.

Literature Survey

1. Y Chen et Al. [1] introduce an innovative approach for feature selection called CCFS (Confidence-based and Cost-effective Feature Selection). CCFS leverages (Binary PSO) to enhance the efficiency of feature selection while minimizing costs. To achieve this, CCFS introduces a novel update mechanism that explicitly considers the confidence of each feature. This confidence calculation considers the similarity amid each feature and the categories of interest, as well as historical frequency for selection for each feature. Furthermore, the design of the fitness function in CCFS is comprehensive, taking into account multiple criteria .It considers not only classification accuracy but also the reduction ratio of features and the associated feature costs. To view the effectiveness of CCFS, experiments were conducted using the UCI cancer classification dataset, specifically focusing on lung cancer classification. Findings from this study demonstrate efficacy of the projected solution in improving feature selection for this critical task. With 30 individuals in the population and a termination after 100 iterations, CCFS achieved an impressive 84.375% classification accuracy at the 87th iteration. This outperformed ACC_BPSO by 3.125% and reached this peak accuracy four iterations earlier, all while maintaining a similar total feature cost. Incorporating feature confidence values during position updates not only improved classification accuracy but also accelerated convergence to the global optimum, avoiding local maxima. By considering both feature count and cost in the fitness function, CCFS enhanced feature selection efficiency, resulting in accurate and cost-effective feature subsets.

2. Riaz. M et Al. [2] utilize 23 publicly available separate datasets are used to analyse how metaheuristics can be applied at different steps in the classification of medical images for COVID-19. It uses chest images with various preprocessing methods such as Normalization techniques including Minmax scaling, image resizing, and Gray scaling, while augmentation methods involve rotation, translation, SMOTE. Mamoona et Al. compare accuracy of classification when features are selected using versions of nature inspired metaheuristic search algorithms such as Cuckoo Search algorithm (CSA), Whale Optimization Algorithm (WOA), Salp swarm algorithm (SSA), Marine predators algorithm (MPA). The paper concludes that many of these algorithms increase classification accuracy for these datasets.

WOA -- 99.22%

CSA -- 96.72%

SSA -- 95.91%,

FO-MPA -- 99.80%.

The limitations encompass challenges related to dataset size, class imbalance, noisy data, the time complexity of metaheuristics. The authors interestingly also noted the need for interdisciplinary collaboration to make significant progress.

3. V. Agrawal et Al. [3] perform a study employing the Artificial Bee Colony (ABC) metaheuristic procedure to conduct FS on CT Scan images of cancer. In the dataset 163 are positive and 108 are negative instances. The primary objective is to discern whether the input data indicates the presence of cancer. The process begins with image segmentation, achieved through the implementation of the Active Contour Segmentation (ACM) algorithm. A semi-automated system is devised to extract the Region of Interest from the segmented images. Subsequently, features such as the texture are extracted from the ROI. The algorithm comes from clever hunting actions of bees and is structured around 3 mechanisms: the bees which work, those who don't, and energy sources. This model based on group behaviour of bees can quickly find a decent solution. One distinctive feature of ABC is its ability to address both local and global search aspects. Few handle limited search, while searching bees manage search in the entire state space, which balances exploration and exploitation tasks. For classification, a hybrid approach combines ABC with (KNN) and (SVM) algorithms. Notably, the results indicate that the fusion of ABC with the 2nd classification method outperforms all other combinations. In the case of k-NN,

Accuracy – 97% with noisy data

Accuracy – 100% with the normal dataset

Consequently, SVM emerged as the optimal choice for classification, although dataset bias significantly impacted results.

4. Elaziz et Al. [4] introduce a novel method of classifying diseases from pictures, Artificial Hummingbird Algorithm with Aquila Optimization (AHA-AO). This algorithm's unique approach draws inspiration from the precision and efficiency of hummingbirds in selecting nectar from flowers. AHA-AO demonstrates a remarkable ability to sift through complex image data and select only the critical features essential for accurate diagnosis. It works by iteratively selecting, combining, and evaluating subsets of features, it uses a fitness function to measure the contribution of each feature subset to the performance of a deep learning model for medical diagnosis. AHA-AO continuously refines these subsets through selection, crossover, and mutation processes until it identifies the most effective feature subset. Its main benefit is the major improvement in efficiency while not compromising in accuracy. The approach tested on four inputs and related with other selection algorithms. This new algorithm had very impressive results reporting percentages of 87.30%, 97.50%, 86.90%, and 88.60% for accuracy for the respective datasets. Importantly, AHA-AO also demonstrated faster feature selection compared to other methods, successfully improving both the performance and efficiency of deep learning models for medical image diagnosis.

5. R Khurma et Al. [5] introduce and compare some interesting variations of Binary Moth Flame Optimization Algorithm (BMFO) incorporating chaotic maps within a framework that performs wrapper-based efforts. These maps primarily improve the initialization for various solutions, aiding the optimizer in escaping local minima and converging towards global optima. These approaches are only recently being used for FS, which are crucial for mitigating dimensionality issues impacting learning processes, including data overfitting and prolonged learning times. Among these, the Moth Flame Optimization (MFO) algorithm stands out as an effective choice for solving various optimization problems across diverse use cases. The proposed approaches are rigorously trained and evaluated on many image datasets achieving competitive performance when benchmarked against other futuristic metaheuristic algorithms. These results are compared with three well known algorithm strategies which include BGWO, BCS, BBA. Results showed that chaotic operators improved BMFO's classification accuracy. BBA was the top-performing method in 78% of datasets, followed by BCS at 22%. BMFO variants didn't surpass these methods. BBA and BCS were strong performers, while BMFO variants showed effectiveness across various datasets.

6. S. Fong et Al. [6] discuss the Swarm Search(SS) based methods. This framework is a versatile strategy compatible with various other processes. It serves as an encapsulating model, using accuracy as the objective function for the evaluation of feature subset candidates for a given classifier. The workflow begins with an initial random feature subset selection and refines the classification model's accuracy through stochastic searches, aiming to converge between the selected feature subset and the classifier. The classifier acts as a black-box evaluator, assessing feature subsets based on accuracy. The optimization function supports different metaheuristics, focusing on finding the best subpart for maximum classification accuracy. Standard brute-force tactic tests all possible subsets, but this is computationally intensive, especially with many features. To address this, we use stochastic search strategies like swarm-based metaheuristics with parallel search agents, seeking the optimal feature subset without exhaustive exploration. However, finding the absolute global best solution may be computationally infeasible, as Xin-She Yang pointed out. In their experiments, they combined three metaheuristics (PSO, BAT, WSA) with three popular classification algorithms (PN, DT, NB) to create diverse SS-FS methodologies. SS-FS methods consistently outperformed other methods in our evaluations, except when paired with Naive Bayes. For instance, on the Arrhythmia dataset, FS-Cfs reduced the error rate from 40% to 30%, while non-FS methods had around 10% error. FS-SS methods achieved high accuracy but had longer computation times. In contrast, FS-Cfs was computationally efficient, taking less than 0.5 seconds in most cases. FS-PSO-PN and FS-BAT-PN had 100% accuracy with FS-SS, while their peers had accuracy between 70-90 %, FS-SS methods retained 50% of all the initial properties for dimensionality reduction, making them valuable for high-dimensional biomedical datasets.

7. Majdi M. et Al. [7] discuss recent developments in optimization and memetic algorithms have sparked significant interest in hybrid metaheuristics. This paper explores two hybridization models that leverage the Whale Optimization Algorithm (WOA) techniques. In the initial attempt, Simulated Annealing (SA) is integrated, while in the next attempt, SA continually parses to find the optima discovered after each step of the metaheuristic algorithm. Exploring only the hopeful solution space is the main goal of incorporating SA. For understanding where this algorithm stands with respect to the competition, the writers have done experimentation with some yardstick data from online sources. These approaches are then compared against a few other methods. The findings conclusively demonstrate the efficacy of the highlighted techniques in enhancing performance compared to other encapsulating algorithms. These findings show how the suggested model can efficiently parse through the search states and help in increasing the accuracy of any categorization tasks performed utilising it. The proposed algorithm, implemented in MATLAB, used a (KNN) classifier with K=5 and Euclidean distance. Cross-validation (K-fold) was conducted with 100 iterations and a population size of 10. Hybrid algorithms outperformed the native one in classification accuracy and feature selection. Tournament Selection (TS)-based approaches, WOASAT-1 and WOASAT-2, enhanced exploration and produced better results than other methods. WOASAT-2 achieved the best approach, offering high accuracy and minimal feature subsets. The WOA algorithm demonstrated robustness and balanced exploration and exploitation, making it a reliable choice for feature selection.

8. L. Meenachi et Al. [8] present innovative methods for predicting cancer using data which can present information about the underlying structure of genomics through the application of metaheuristics for FS. While global feature selection methods such as ant colony optimization and genetic algorithms excel at identifying optimal features, challenges arise in selecting nearby features. To overcome these challenges, two new feature selection algorithms are introduced: the ACTFRO Algorithm, and the GATFRO Algorithm. These algorithms are evaluated using a fuzzy rough nearest neighbour classifier with on the relevant input data. Accuracy for categorization, time for problem solving, harmonic mean of the precision and recall, are among the metrics of evaluation. Results are obtained by combining these algorithms with the FRNN classifier. The hybrid approach of global, local feature selection yields good results across different training/testing ratios. Tenfold cross-validation confirms their robustness, with ACTFRO achieving an average accuracy of 88.54%, and GATFRO reaching 90.53%.These algorithms consistently outperform existing in the above discussed metrics of evaluation shorter computation times, making them promising for practical applications in cancer diagnosis and other fields.

9. Gehad Ismail Sayed et Al. [9] explore the Crow Search Algorithm (CSA), inspired by how crows hide their food, was introduced by Askarzadeh in 2016. However, CSA, like many optimization algorithms, faces challenges like slow convergence and local optima trapping. To overcome these limitations, they propose a new method named the Chaotic Crow Search Algorithm (CCSA). In their study, they apply CCSA to perform the tasks of FS for around a dozen yardstick input sources, utilizing just less than a dozen different enhancers which were originally employed in CSA. Their experiments demonstrate CCSA's exceptional ability to identify the best subpart, enhancing grouping performance while chopping down on the properties that actually had an impact on the solution. We observed that their experiments highlight the sine chaotic map as a particularly effective choice for significantly improving CSA's performance. They introduce CCSA, where chaotic variables replace random ones for crow position updates. Ten chaotic maps, such as sine and logistic, are used to enhance CSA's performance. Binary CCSA is proposed for feature selection, converting solutions into binary form. It focuses on improving classification accuracy, reducing feature subset length, and lowering computational costs. Parameters are initialized, and the fitness function combines accuracy and feature count using a weight factor. Position updates follow specific equations, and the optimization process ends after a set number of iterations. In summary, CCSA integrates chaos into feature selection, with the goal of boosting accuracy, reducing features, and saving computation.

10. Hany M. Harb et Al. [10] explore the vital role of classification analysis in healthcare applications, aiding medical diagnostic decisions and enhancing patient care quality. When working with vast medical databases, selecting a relevant subset of data for training is crucial. If irrelevant features are included in the training dataset, the task of grouping may become more prone to errors and harder to explain and understand. Our task which is to choose the most vital properties in the input data, an important step, addresses this issue. In this study, they introduce two feature selection methods, both integrated with PSO. To evaluate their performance, they compare them with algorithm using GA approaches. Applying 2 algorithms to a few distinct medical datasets. In their work, researchers propose a filter approach with Correlation-based FS. They also apply a wrapper approach using PSO to improve classifier predictions. Experiments were conducted on three medical datasets: Breast Cancer, Heart Statlog, and Dermatology. PSO with CFS was used for feature selection. The findings are: the proposed wrapper model reduced how many inputs we will be considering but still ensuring that the methodology performs as well as possible without becoming error prone for Naïve bayes and RBF classifiers. The Bayesian classifier showed the same accuracy as GA_CFS despite reducing six features out of nine. In the case of the Heart Statlog dataset, the KNN classifier's accuracy remained consistent with all inputs and improved slightly with the proposed filter PSO - CFS. However, GA_CFS slightly outperformed the proposed wrapper model by 2%.For the Dermatology dataset, feature subset selection improved all evaluation metrics, except for RBF classifier, which did not improve accuracy with the proposed models compared to GA_CFS. In summary, these experiments highlight the importance of feature selection, demonstrating improved classification accuracy for most classifiers while reducing irrelevant attributes.

11. Qiang Li et Al. [11] introduces a novel predictive framework, IGWO-KELM, integrating Improved Grey Wolf Optimization (IGWO) with (KELM). IGWO is a popular method utilized. GA is first used to create diverse starting states. Then, their optimizer is applied to rearrange these states accordingly to obtain the best possible subparts. The authors perform a competition between this new model with older more popular algorithms and test out the ability of the model through many standards. They note the clear dominance of this algorithm as a result of the competition. Utilizing available information, IGWO-KELM consistently outperforms the competing models in all measures. For the 1st dataset, IGWO-KELM achieves 97.45% accuracy with fewer features, and in the WDBC dataset, it attains 95.78% accuracy. Convergence analysis shows IGWO-KELM's quicker convergence and better solution quality, emphasizing key features like Fo, RPDE, D2, DFA, and PPE for medical diagnosis. Parameter tuning determines that a population size of 8 and 100 iterations yield optimal results for IGWO-KELM. In summary, IGWO-KELM offers efficient feature selection and high accuracy in medical diagnosis.

12. This paper [12] introduces an improved version of the Whale Optimization Algorithm called E-WOA. E-WOA incorporates strategies related to the behaviour of various species of aquatic mammals. The paper evaluates performance and compares it with established variants for solving issues requiring a complete run through of the search space. The results demonstrate that the new model surpasses all other competitors. Once the efficacy was established, it was adapted to create a binary version for FS for disease detection. BE-WOA's performance was assessed using medical disease datasets and compared against the best of the best in terms of real-world performance and also against a variety of standards. Comparatively this model definitely out-performs many others. WOA operates with a population of whales, and it uses probability rates and coefficient vectors to balance between these methods, promoting both exploration and exploitation. However, the original WOA suffers from issues like premature convergence and low diversity. To address these limitations, various WOA variants have been proposed, such as (LWOA), a version using maps, and a more stable version each introducing unique enhancements to the algorithm. These variants offer improvements over the original WOA, making them valuable tools in optimization tasks. The results of the paper are very interesting and show how wildlife inspired algorithms can play a huge part in our task.

13. Zenab Mohamed Elgamal et Al. [13] study the Harris Hawks Optimization (HHO) algorithm, inspired by the bird of preys impressive strategy while looking for food which has shown promise but faces some nagging issues. To address these challenges and adapt HHO for our task, they propose the Chaotic Harris Hawks Optimization (CHHO) algorithm. CHHO introduces a few important enhancements. Firstly, erratic functions are used during initialization to improve population multiplicity. Secondly, they incorporate some steepest ascent-based methods for maximized exploitation. They evaluate CHHO on 14 medical datasets and compare it with many other notable algorithms such as ALO, BOA, etc. CHHO demonstrates superior performance across most datasets, making it a promising approach for optimization tasks. In the exploration phase, HHO generates potential solutions using mathematical equations that simulate the hawks' behaviour. Chaotic maps are introduced to enhance solution diversity, while Simulated Annealing (SA) guides the entire search based on the ”preys” characteristics. The exploitation phase applies different strategies, such as soft and hard besiege, depending on prey energy and escape probability, aiming to improve solution quality. The Chaotic Harris Hawks Optimization (CHHO) algorithm improves HHO by using chaotic maps for initialization and integrating SA as a local search operator. CHHO consistently outperforms HHO and other optimization algorithms, particularly in feature selection tasks. CHHO offers a promising approach to complex optimization problems, although it may require some computational time.

14. Hoda Zamani et Al. [14] introduce FSWOA, a method chosen based on the behaviour from Humpbacks in their pursuit of food, showcasing its ability to reduce dataset dimensionality while maintaining acceptable accuracy for disease diagnosis. FSWOA comprises three primary steps: Initial Placement: They begin by generating k humpback whales and scattering them randomly across the search space. Each whale's position is evaluated, and the best whales are selected. The remaining whales adjust their positions towards the best whale. Bubble-Net Attacks: In this step, humpback whales initiate attacks using bubble-net strategies, which include strategies like shrinking encirclement and spiral position updates. This phase resembles an exploitation phase where each whale suggests feature subsets. These subsets are evaluated based on classifier accuracy. Random Prey Search: The third step, known as the exploration phase, involves humpback whales randomly searching for prey using information from members. The algorithm's performance is assessed using online data repositories. These datasets cover tasks such as diabetes diagnosis, breast cancer detection, heart disease prediction, and hepatitis prognosis. We normalize the datasets to account for real-world noise and incompleteness before applying the FSWOA algorithm. The algorithm selects feature subsets, and their quality is evaluated using standard metrics like sensitivity, specificity, precision, NPV, and accuracy. Sensitivity and specificity measure correct classification in positive and negative classes, while precision and NPV are also calculated. They implement the algorithm in MATLAB on standard hardware and test it 15 times on each dataset using random 70-30 training-testing splits. The KNN algorithm (K=3) evaluates the selected feature subsets, a maximum iteration limit of 60, an initial population size of 30, and lower and upper bounds for feature selection. It effectively reduces feature dimensionality, leading to more efficient classifiers, without the need for additional tables and figures.

15. In the scholarly work led by Tuan M. et Al. [15], the pressing concern of diabetes classification was addressed through a rigorous examination of patient medical records. With a commitment to advancing predictive accuracy and computational efficiency, the researchers introduced an innovative machine learning model. This model was designed with a distinct emphasis on enhancing performance through the integration of a aforementioned encapsulation methodologies , harnessing the capabilities of APSO and GWO. They comprehensively assess the efficacy of this pioneering approach, and study conducted comparative analyses against a cohort of well-established machine learning algorithms. The outcomes yielded remarkable insights. Firstly, the newly proposed methodology exhibited a notable capacity for substantially reducing the requisite number of features for precise diabetes prediction. This achievement was particularly noteworthy, given the intricate nature of medical datasets. Secondly, and indeed more impressively, the model demonstrated a significant enhancement in prediction accuracy. Specifically, the GWO-MLP configuration achieved a commendable accuracy rate of 96%. However, it was the APGWO-MLP configuration that emerged as the standout performer, surpassing all expectations with an exceptional 97% accuracy rate.

16. This paper [16] introduces an utilizes a GA approach with a wrapper method. The wrapper method uses GAs to explore and evaluate subsets of features by employing a model to assess their performance. The GA operates by searching for individuals to optimize the task. These candidates are represented as genotypes or chromosomes, using binary strings, and are translated into numerical forms as candidate solutions. The GA is begun as always by creating a population and a comparative score function. Support Vector Machine (SVM) is employed to classify features, and these results contribute to the fitness function. To assess their performance, they compare GA-SVM with several established methods. The framework is implemented in MATLAB, utilizing a dataset sourced from the internet. The GA employs several genetic operations: Roulette wheel selection: Individuals' fitness values determine their portion of a roulette wheel, and the wheel is spun to select parents for the next generation. Crossover (recombination): Two parent solutions produce a child through a crossover process. This process reproduces significant strings without creating new ones. Mutation: After crossover, strings undergo mutation by flipping bits from 0 to 1 and vice versa. This process aims to maintain genetic diversity and explore the search space. Elitism: Survival of the fittest to help GA performance. In conclusion, the paper's GA–SVM approach outperforms other feature selection methods, achieving a higher accuracy of 88.34% in classifying heart disease with selected features, as opposed to 83.70% with the whole feature set. ROC analysis supports the effectiveness of the SVM classifier.

Conclusion

To find diseases, the proper selection of important attributes from vast datasets is crucial for building accurate and efficient predictive models. Traditional feature selection methods often fall short in handling complex and high-dimensional data. To address this challenge, this report discusses the application of metaheuristic search algorithms for optimizing feature selection in medical disease detection. Furthermore, we outline our plan to implement an ensemble model that can compare different metaheuristic algorithms using various imaging datasets and analyse the results.

Medical disease detection has greatly benefited from advancements in imaging technologies and machine learning. However, the quality and quantity of medical imaging data pose challenges for accurate diagnosis and prediction. Traditional methods are often inefficient when dealing with high-dimensional data like medical images. Metaheuristic search algorithms offer a promising alternative by efficiently exploring feature spaces to identify relevant attributes.

Our report enumerates examples of the applications of these algorithms through evident improvements on the existing methodologies beforehand.

Learnings

Reading and analysing research papers helps us understand the different stages in research, including data collection, preprocessing along with data analysis and exploration.

Formulating a plan and identifying how to create improvements in existing techniques and introduce novel methods is also an essential step. Conducting relevant experiments followed by the justification and analysis of performance are essential, along with presenting results in a meaningful and worthwhile manner.

Throughout the course of this literature survey, guided by the simultaneous instruction in the course which expanded on our understanding of the subject material, we were able to understand what necessitates the need for metaheuristic algorithms in problems that have real world applications in the extremely important domain of healthcare and medicine.

Through the experimental sections of the papers we read, we saw how to create simulations and apply optimised models to real world datasets to evaluate performance.

This topic appealed to us due to our shared a passion for nature and keen interest in the animal kingdom.

Research in this field is often very inter-disciplinary requiring collaboration between mathematicians, computer scientists and biologists. This assortment of different viewpoints and backgrounds fosters a spirit of innovation leading many of these algorithms to offer a significant improvement to previously used methods.

It provides those who dive into it, a deeper and more meaningful appreciation of the value present in even the simplest forms of life.

We believe that one of the most important applications of Algorithms and optimization is in healthcare and medicine.

Through this report, we were able to explore the intersection of animal behaviour in nature and how evolution has guided them towards these wondrous strategies and how human beings can utilise the same techniques to improve our daily lives.

Research Gaps and Objectives

Previous Research has shown the utility and optimization benefits of metaheuristics of specific methods in disease detection.

We plan to build an ensemble model utilizing some lesser-known algorithms/ hybrid implementations of popular algorithms.

We also plan to explore multi-objective optimization which includes extending known algorithms to handle multiple conflicting objectives, maximizing accuracy while minimizing feature count, handling class imbalances in popular datasets.

As mentioned, we also must keep in mind the topics of ethics and explainability in the models we design, which is not commonly seen in the papers which we have surveyed.

These papers often just apply the algorithms and present the accuracy without paying due diligence to explaining the steps followed by these encapsulating models in the spirit of upholding ethics.

Finally, we will attempt to analyse the results to determine the most effective metaheuristic algorithm for feature selection in different medical scenarios while also trying to bridge some of the lesser known/studied areas as mentioned above.

Scope

This study's scope involves creating an ensemble model for medical data analysis, with a primary focus on using metaheuristic algorithms for FS. We will assess various metaheuristic approaches' effectiveness in selecting optimal feature subsets from medical datasets and integrate these features into ensemble models to enhance disease diagnosis and prognosis accuracy. Ethical considerations, interpretability, and practical clinical applications will be addressed. Comparative analyses with traditional methods will be conducted for comprehensive evaluation.

We will also try to utilise some lesser known and novel algorithms to possibly expand the field of optimization.

Methodology

1. Selection of Metaheuristic Algorithms

We plan to explore and implement a selection of metaheuristic search algorithms from the above-mentioned literature survey.

2. Datasets

We will gather diverse medical imaging datasets encompassing various diseases and imaging modalities (e.g., X-ray, MRI, CT scans). These datasets will serve as the evaluation criteria.

3. Feature Representation

To apply metaheuristic algorithms, we will represent the features as a binary string, with each bit corresponding to the inclusion or exclusion of a specific feature. The algorithms will evolve and optimize these binary strings to select the most relevant features.

4. Ensemble Model

We will develop an ensemble model that incorporates the output of different metaheuristic algorithms. The ensemble will allow us to leverage the strengths of each algorithm and potentially mitigate their weaknesses. The final ensemble will be trained on the selected features to make predictions for disease detection.

Execution Timeline

Task 1:

- Data Collection: Gather medical imaging datasets from reputable sources.
- Algorithm Implementation: Code and implement the selected metaheuristic search algorithms for feature selection.
- Data Preprocessing: Prepare the collected data for feature selection and model training.

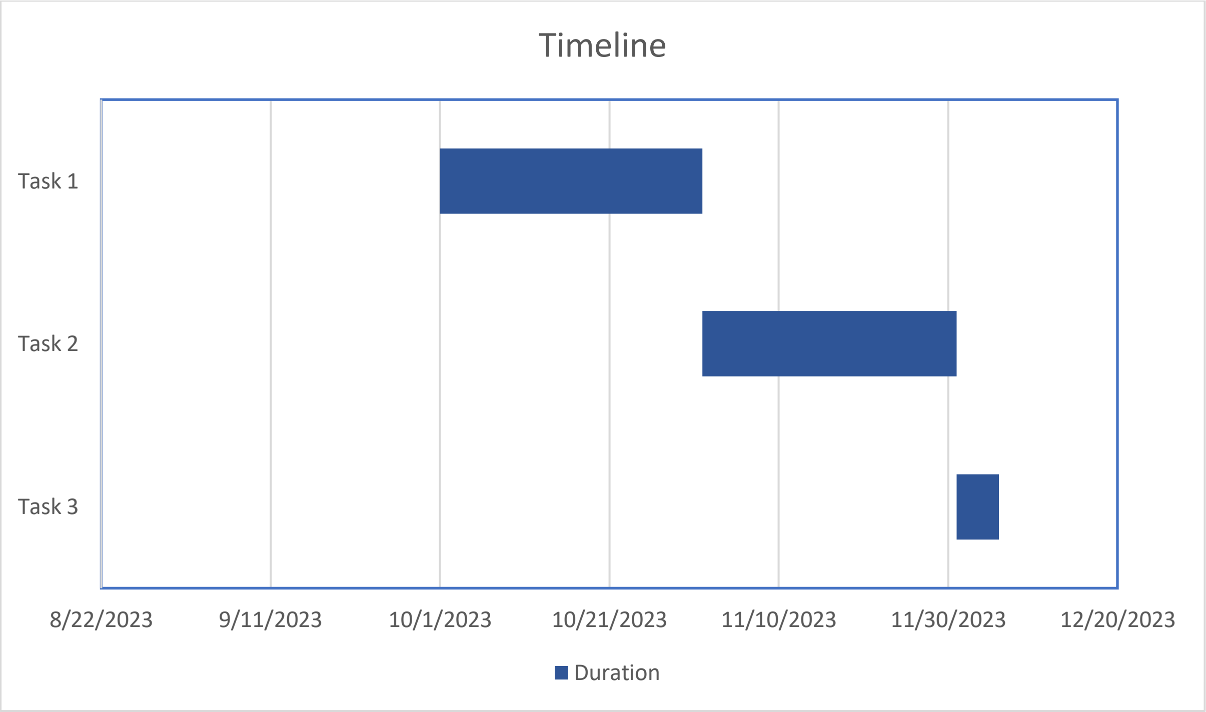
Task 2:

- Feature Selection: Apply metaheuristic algorithms to select relevant features for disease detection on multiple datasets.
- Ensemble Model Development: Design and build the ensemble model that combines the outputs of different algorithms.
- Model Training: Train the ensemble model on the selected features from the datasets.

Task 3:

- Performance Evaluation: Measure and compare the predictive abilities of the model using various metrics which have been enumerated in the literature survey portion of this undertaking.
- Results Analysis: Analyse the results to determine the effectiveness of different metaheuristic algorithms in different medical scenarios.
- Report Writing: Prepare the final report summarizing the methodology, results, and conclusions.

GANTT CHART



This report outlines our plan to use metaheuristic search algorithms for optimizing feature selection in medical disease detection. By implementing an ensemble model and evaluating different algorithms on diverse imaging datasets, we aim to identify the most effective approaches for FS in healthcare. This research has the potential to ultimately benefit both healthcare professionals and patients.

Bibliography

[1] Y. Chen, Y. Wang, L. Cao and Q. Jin, “An effective feature selection scheme for healthcare data classification using binary particle swarm optimization.,” in *9th international conference on information technology in medicine and education (ITME)*, 2018.

[2] M. Riaz, M. Bashir and I. Younas, “Metaheuristics based COVID-19 detection using medical images: A review.,” *Computers in Biology and Medicine, 144, 105344.,* 2022.

[3] V. Agrawal and S. Chandra, “Feature selection using Artificial Bee Colony algorithm for medical image classification.,” in *2015 eighth international conference on contemporary computing (IC3).*, 2015.

[4] M. A. Elaziz, A. Dahou, S. El-Sappagh, A. Mabrouk and M. M. Gaber, “AHA-AO: artificial hummingbird algorithm with Aquila optimization for efficient feature selection in medical image classification.,” *Applied Sciences, 12(19), 9710.,* 2022.

[5] R. A. Khurma, I. Aljarah and A. Sharieh, “An Efficient Moth Flame Optimization Algorithm using Chaotic Maps for Feature Selection in the Medical Applications.,” in *ICPRAM* , 2020.

[6] S. Fong, S. Deb, X. S. Yang and J. & Li, “Feature selection in life science classification: metaheuristic swarm search.,” *IT Professional, 16(4), 24-29.,* 2014.

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

[8] L. Meenachi and S. Ramakrishnan, “Metaheuristic search based feature selection methods for classification of cancer.,” *Pattern Recognition, 119, 108079.,* 2021.

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

[10] H. M. Harb and A. S. Desuky, “Feature selection on classification of medical datasets based on particle swarm optimization.,” *International Journal of Computer Applications.,* 2014.

[11] Q. Li, H. Chen, H. Z. X. Huang, Z. T. C. Cai and X. Tian, “An enhanced grey wolf optimization based feature selection wrapped kernel extreme learning machine for medical diagnosis.,” *Computational and mathematical methods in medicine,* 2017.

[12] M. H. Nadimi-Shahraki, H. Zamani and S. Mirjalili, “Enhanced whale optimization algorithm for medical feature selection: A COVID-19 case study.,” *Computers in biology and medicine, 148, 105858.,* 2022.

[13] 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, 8, 186638-186652.,* 2020.

[14] H. Zamani and M. H. Nadimi-Shahraki, “ Feature selection based on whale optimization algorithm for diseases diagnosis.,” *International Journal of Computer Science and Information Security, 14(9), 1243.,* 2016.

[15] T. M. Le, T. M. Vo, T. N. Pham and S. V. T. Dao, “A novel wrapper–based feature selection for early diabetes prediction enhanced with a metaheuristic.,” *IEEE Access, 9, 7869-7884.,* 2020.

[16] C. B. Gokulnath and S. P. Shantharajah, “An optimized feature selection based on genetic approach and support vector machine for heart disease.,” *Cluster Computing, 22, 14777-14787.,* 2019.