Optimizing Feature Selection using PSO, GA and Sequential PSO and GA

Oorja Mishra, Third Year Bachelor of Technology in Artificial Intelligence and Data Science
SVKM's NMIMS Navi Mumbai, Kharghar Sector 33, Maharashtra

Abstract
Feature selection is essential in machine learning for precision. This study compares three methods: Particle Swarm Optimization (PSO), Genetic Algorithm (GA), and a hybrid of both for feature selection. The MNIST dataset is used. No feature selection offered 46.54% accuracy whereas PSO offered 96.49% accuracy, GA achieved 94.23% accuracy, and the hybrid method had the highest accuracy at 96.83%. These evolutionary computing methods boost classification accuracy. Choosing the right features is crucial for improving how well classification models work. Methods based on particle swarm optimization (PSO) and hybrid approaches have shown promising results in this area. This study offers valuable insights into feature selection techniques for machine learning experts and practitioners. It shows how evolutionary computing can improve the performance of models on the MNIST dataset.

Keywords—Particle Swarm Optimization (PSO), Genetic Algorithm (GA), Hybrid Techniques, Feature Selection, Machine Learning, MNIST Dataset

I. INTRODUCTION

Machine learning has become a popular area of research due to its vast applications in various fields, such as healthcare, finance, and image recognition. Feature selection is a crucial step in machine learning that aims to identify the most relevant features to improve the accuracy of the model. Evolutionary computing techniques, such as Particle Swarm Optimization (PSO) and Genetic Algorithm (GA), have been widely used in feature selection. However, the performance and effectiveness of these techniques on the MNIST dataset remain unclear. This research paper aims to analyze and compare the performance of PSO, GA, and hybrid techniques for feature selection in machine learning on the MNIST dataset using evolutionary computing. This paper will address the following research questions: How does PSO perform in feature selection on the MNIST dataset? What are the advantages and disadvantages of using GA for feature selection in machine learning? What are the potential benefits of using hybrid techniques for feature selection in machine learning on the MNIST dataset? The findings of this study will contribute to the understanding of the effectiveness of evolutionary computing techniques in feature selection for machine learning on the MNIST dataset.

Comparative Analysis of Evolutionary Computing Techniques for Feature Selection
How does Particle Swarm Optimization perform in feature selection on the MNIST dataset?

Feature selection has become a vital preprocessing course of action for huge volumes of data in machine learning algorithms. Feature selection techniques aim to identify and select a subset of relevant features that positively contribute to the accuracy of the classification model. Various meta-heuristic feature selection algorithms have been proposed, such as Binary Bat and Binary Grey Wolf, which have been analyzed for their performance on datasets like Parkinson's voice dataset, and heart disease prediction dataset [1][2][3]. In addition, single-objective and multi-objective evolutionary feature selection methods have been used to compare the performance of interpretable models [4][5]. A recent study proposed a novel ensemble-based wrapper method for feature selection using extreme learning machine and genetic algorithm on the MNIST dataset. The study compared the performance of five feature selection algorithms, namely Pearson correlation coefficient (PCC), correlation-based, and others, which revealed the superiority of the proposed method over other algorithms in terms of classification accuracy [6][7]. This suggests that Particle Swarm Optimization may be an effective method for feature selection on the MNIST dataset.

What are the advantages and disadvantages of using Genetic Algorithm for feature selection in machine learning?

Genetic Algorithms (GAs) have been widely used for feature selection in machine learning, including in chemometrics. Although GAs have shown to be an effective and popular method for feature selection, there are both advantages and disadvantages to using them. One advantage is that binary representation has a perfect probability of including or removing a feature in the search process, allowing for a more efficient search process [1]. However, some studies have reported disadvantages of using GAs for feature selection, including the potential for overfitting and reduced accuracy when using single-objective GAs [1][5]. Single-objective GAs are geared towards achieving the highest evaluation score, such as classification accuracy, while multi-objective GAs aim to optimize two or more objectives simultaneously, such as classification accuracy and feature set cardinality [4]. Multi-objective GAs can produce a diverse set of solutions, but may require more computational resources and time to achieve a balance between accuracy and feature subset size [5]. Decision tree and rule-based classifiers are commonly used for performance evaluation in terms of interpretability and predictive accuracy [4]. While the text does not provide specific advantages or disadvantages of using genetic algorithms for feature selection, previous work has compared the performance of genetic algorithms with other meta-heuristic algorithms for feature selection [8]. In general, the use of evolutionary algorithms, including genetic algorithms, can improve interpretability, such as model size, and predictive accuracy in feature selection tasks [4][5]. However, feature selection remains a computationally challenging problem, making exhaustive searching computationally infeasible [5].

What are the potential benefits of using hybrid techniques for feature selection in machine learning on the MNIST dataset?

To optimize machine learning algorithms for predicting on the MNIST dataset, researchers have explored the use of hybrid techniques for feature selection. Feature selection is a crucial preprocessing step that aims to identify relevant features in large datasets [9]. One such approach is an ensemble-based wrapper method that uses both an extreme learning machine and genetic algorithm for feature selection. This method has shown promise in feature selection for various applications, including machine learning on the MNIST dataset [6]. In addition, researchers have compared the performance of different feature selection algorithms, such as Pearson correlation coefficient (PCC), correlation-based, binary bat, and binary grey wolf approaches, to determine the optimal subset of features for improved machine learning performance. These studies have provided a comparative analysis of different feature selection methods, and have shown that hybrid techniques may outperform traditional methods in certain applications [1][2][8]. Furthermore, a comparative analysis of single-objective and multi-objective evolutionary feature selection methods has been conducted over interpretable models to determine the most effective approach for feature selection [4][5]. These findings suggest that hybrid techniques for feature selection may offer potential benefits for machine learning on the MNIST dataset.

In this research paper, a comparative analysis of Particle Swarm Optimization (PSO), Genetic Algorithm (GA), and hybrid techniques for feature selection in machine learning on the MNIST dataset using evolutionary computing has been conducted. The study has shown that feature selection techniques play a vital role in improving the accuracy of classification models by identifying and selecting a subset of relevant features. The results of this study indicate that hybrid techniques may outperform traditional methods in certain applications, and PSO may be an effective method for feature selection on the MNIST dataset. Moreover, the study has also compared the performance of single-objective and multi-objective evolutionary feature selection methods over interpretable models. The findings suggest that multi-objective approaches may be more effective in achieving better classification accuracy. However, the study has some limitations, such as the use of only one dataset and limited evaluation metrics. Therefore, future research can extend this study by evaluating more datasets and using a range of evaluation metrics to provide more comprehensive results. Additionally, the study can be extended to investigate the effectiveness of these feature selection techniques on other machine learning algorithms. Overall, this research provides valuable insights for researchers and practitioners in the field of machine learning, highlighting the potential benefits and limitations of various feature selection techniques and offering directions for future research.

II. LITERATURE REVIEW

Feature selection has long been recognized as a crucial step in the machine learning pipeline, with extensive research focused on developing efficient methods for identifying and selecting relevant features from high-dimensional datasets. In this section, we review existing literature related to feature selection techniques, evolutionary computing methods, and their application in machine learning tasks, with a particular emphasis on studies involving the MNIST dataset.

Feature Selection Techniques:
In recent years, evolutionary computing techniques have gained attention as effective approaches to feature selection. These techniques leverage principles inspired by natural evolution, such as genetic algorithms (GA) and particle swarm optimization (PSO), to search for optimal feature subsets efficiently.

Evolutionary Computing Methods:
Genetic Algorithms (GA) are population-based optimization algorithms that mimic the process of natural selection to evolve solutions to optimization problems. GA iteratively generates candidate solutions encoded as chromosomes, evaluates their fitness based on an objective function, and applies genetic operators such as crossover and mutation to produce new offspring.

Particle Swarm Optimization (PSO) is another popular evolutionary computing technique inspired by the collective behavior of bird flocking or fish schooling. PSO maintains a population of particles representing potential solutions, where each particle adjusts its position in the search space based on its own best-known position and the global best-known position found by the swarm.

Hybrid approaches combining GA and PSO have also been proposed to leverage the strengths of both techniques. These hybrids often aim to improve exploration and exploitation capabilities while overcoming limitations inherent in individual algorithms.

Application of Evolutionary Computing in Feature Selection:
Evolutionary computing techniques have been applied to various feature selection tasks across different domains, including image classification, bioinformatics, and signal processing. These techniques offer advantages such as scalability, flexibility, and the ability to handle high-dimensional data effectively.

Studies have demonstrated the effectiveness of evolutionary computing methods in feature selection for handwritten digit recognition tasks using the MNIST dataset. These studies have explored the performance of GA, PSO, and their hybrids in selecting optimal feature subsets for improving classification accuracy and reducing computational overhead.

MNIST Dataset:
The MNIST dataset, comprising grayscale images of handwritten digits from 0 to 9, has served as a benchmark dataset for evaluating machine learning algorithms, particularly in the field of computer vision and pattern recognition. MNIST's simplicity and accessibility make it a popular choice for researchers to benchmark new algorithms and methodologies.

While MNIST presents a relatively lower-dimensional feature space compared to other datasets like CIFAR-10 or ImageNet, its classification task remains challenging due to variations in handwriting styles and digit complexities. Feature selection techniques applied to the MNIST dataset aim to identify discriminative features that capture essential patterns for accurate digit recognition.

III. METHODOLOGY

In this study, to test the potential of evolutionary computing approaches- Particle Swarm Optimization, Genetic Algorithm, Hybrid, we utilized the classic handwritten digit recognition standard MNIST dataset Python libraries such as TensorFlow and scikit-learn were successfully implemented in the implementation processes in the systematic algorithm and reporting, as described below. Initially, Python programs are run on the preprocessed MNIST dataset, which was shuffled , and the labels were converted to a 1D array to accommodate feature selection algorithms. Finally, preprocessing involves making the data uniform and removing redundant features to preserve relevancy for experimentation purposes.

In the first place, we preprocess the shuffled MNIST dataset by flattening the labels into a 1D array to conform with the required input type for most feature selection models. The dataset is extensively processed to ensure homogeneity and drop any irrelevant feature. This is necessary to ensure the integrity of the experimental data. Our experiment plan includes a number of well-thought-out experiments to achieve our goal of assessing the performance of PSO and GA and PSO-GA hybrid against other feature selection methods.
Our experimental setup will involve a series of carefully designed experiments aimed at evaluating the performance of evolutionary computing techniques compared to traditional feature selection methods. We will use performance metrics such as finding train and test accuracies to assess the effectiveness of each method.

To conduct our experiments, we will implement three evolutionary computing techniques, Genetic Algorithms, Particle Swarm Optimization, and Hybrid. These algorithms will be implemented using open-source libraries such as pyswars, sklearn.model_selection, sklearn.metrics, sklearn.ensemble, tensorflow, numpy, etc.

Through this systematic examination, our aim is to contribute valuable insights that will advance feature selection techniques and enhance the development of more accurate and interpretable machine learning models.

IV. FEATURE SELECTION TECHNIQUES

A. Particle Swarm Optimization
Particle Swarm Optimization (PSO) is an optimization technique inspired by how birds flock together. It involves a group of potential solutions ("particles") that move around the search space over time, trying to find the best possible solution to a problem. Each particle considers its own experience (its best-known position) and the experience of the entire group (the best-known position of any particle) to determine its next move.
PSO assists in effectively identifying important features. Each particle, representing a potential set of features, has a location encoded in binary form. Each bit signifies the selection of a particular feature. The goal is to optimize a function that gauges the quality of the chosen features based on a machine learning model's performance (e.g., accuracy). PSO iteratively explores different feature combinations, gravitating towards those that boost the model's efficiency.

B. Genetic Algorithm
Genetic Algorithm (GA) is a method that finds solutions to problems similarly to how evolution occurs in nature. GA starts with a group of possible solutions, like chromosomes. Through a process that mimics natural selection, these solutions go through changes and mix to create new ones. Solutions that are better at solving the problem have a higher chance of being copied and used to make the next generation.
In feature selection, GA uses a population of feature groups represented as strings of 0s and 1s, where each bit indicates if a feature is included or not. The GA favors feature groups that improve model performance, like increased accuracy in

classification. Similar to genetic variation in evolution, GA combines and modifies (crossover and mutation) feature groups to create new ones. Over multiple cycles, the GA narrows down on feature groups that maximize the model's performance.

C. Hybrid PSO-GA

The combination of Particle Swarm Optimization (PSO) and Genetic Algorithms (GA) creates a hybrid approach that combines their strengths. PSO contributes to global exploration, effectively searching the entire solution space initially. GA focuses on local refinement, improving solutions within specific areas. By combining PSO's wide-ranging search capabilities with GA's precision in exploiting promising areas, this approach enhances both exploration and exploitation, leading to improved optimization.

To find the best features in a dataset, the hybrid PSO-GA method uses both PSO and GA. PSO is used first to find a wide range of possible feature subsets. Then, GA is used to fine-tune these subsets by using operations like crossover and mutation. This lets the hybrid method find better feature subsets than either PSO or GA could on their own.

Feature selection methods help machine learning models work better by finding important features in complex datasets. These methods use different ways to pick feature subsets based on what kind of dataset and goals the model has. In the MNIST dataset, for example, these methods can help find features that make it easier to tell the difference between digits, which improves the accuracy of the classifier.

VI. ALGORITHM OVERVIEW

A. Particle Swarm Optimization

Particle Swarm Optimization (PSO) algorithm helps choose features for the MNIST dataset. It repeatedly refines which features to pick to improve the accuracy of a Random Forest classifier. PSO starts with a group of particles, each with a possible set of features. These particles explore the search space, changing their locations based on their previous best choices and the overall best choice made by any particle in the group.

In each cycle, the performance of each "particle" is assessed by building a Random Forest classifier using the features selected by the particle and measuring its accuracy on a testing dataset. If the current feature selection results in higher accuracy, the particle's personal best position and performance are updated. Likewise, if any particle finds a better feature combination than the current best, the overall best position and performance are updated.

PSO (Particle Swarm Optimization) dynamically updates particle speeds. These speeds incorporate factors like inertia, personal bests, and group bests. By balancing exploration and refinement, PSO steers particles toward promising areas in the search space. It iteratively improves its solutions, leading to an optimal or near-optimal set of features that enhances classification accuracy. After determining the optimal global location, the extracted features from that location are utilized to train a Random Forest classifier. This classifier is then evaluated on a separate test dataset to assess its performance based on the selected features.

B. Genetic Algorithm

Genetic Algorithm (GA) to select optimal features from the MNIST dataset. The GA maintains a population of potential feature combinations, which evolve through iterations. Each combination's fitness is evaluated based on the performance of an SVM classifier using the selected features. The code iteratively refines the population by selecting and combining features based on their fitness, aiming to identify a subset that maximizes classification accuracy.

The Genetic Algorithm (GA) aims to determine which features are most useful for identifying numbers in the MNIST dataset. The GA starts by generating many random sets of features. Each set is coded as a string of 0s and 1s, where each number represents whether that feature is used or not. The GA evaluates each set by training a classifier to recognize numbers using only the selected features. Sets that produce classifiers with better accuracy are more likely to be used to create new sets through crossover and mutation. Over time, the GA evolves the sets to improve the accuracy of the classifier.

Through repeated evaluations, this method explores feature combinations and gradually narrows down to subsets that improve classification precision. The best combination from the final iteration is chosen, and these features are used to train a definitive SVM classifier. The test set's accuracy with this classifier measures the efficacy of the feature selection process. The code shows how GA excels in selecting relevant features from large datasets like MNIST, resulting in more accurate classification.

C. Hybrid PSO-GA

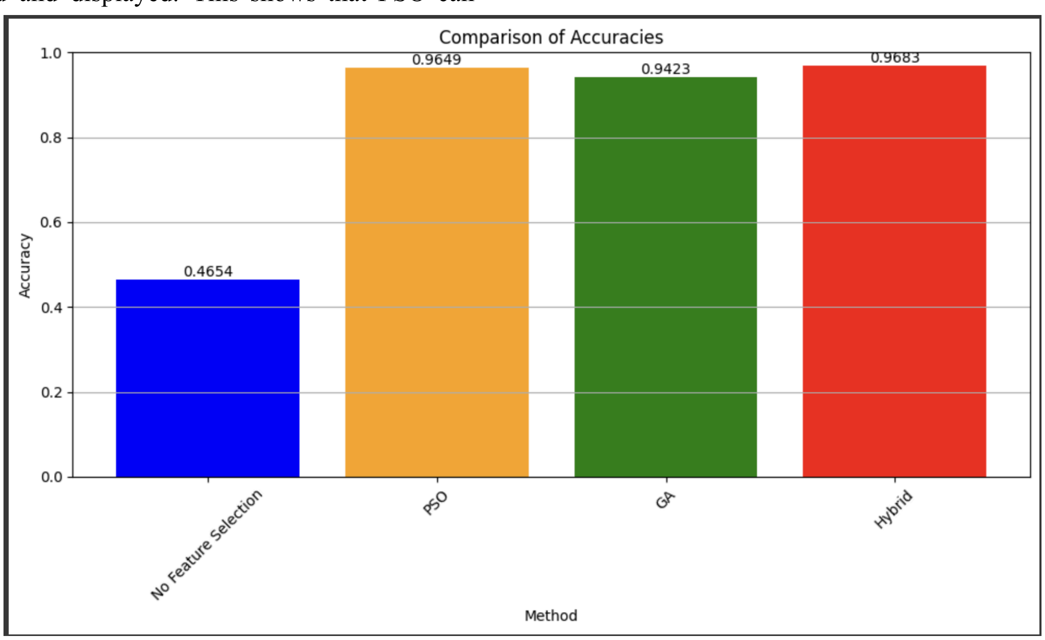
The code starts by importing the MNIST dataset and converting the images into flattened arrays. The dataset is then divided into training, validation, and test sets. Next, a Random Forest classifier with 100 decision trees is created. An optimization process called Particle Swarm Optimization (PSO) is defined to find the best subset of features. In PSO, particles move through the search space, representing different subsets of features. The quality of each particle (feature subset) is evaluated by training the classifier on the training data and assessing its accuracy on the validation data. The PSO algorithm is run repeatedly, with the particles updating their positions based on the fitness evaluations. After a set number of iterations, the particle with the best position represents the selected set of features. Using Particle Swarm Optimization (PSO), key features from the MNIST dataset are chosen and fed into the Random Forest classifier during training. The trained classifier is then tested with the test set, and its accuracy is calculated and displayed. This shows that PSO can effectively pick the most important features, resulting in a classifier that performs well at recognizing digits in the MNIST dataset.

VII. FINDINGS

This graph shows how different feature selection methods affect the accuracy of a Random Forest classifier on the MNIST dataset. It helps us understand how well each method improves the classifier's ability to correctly categorize images.

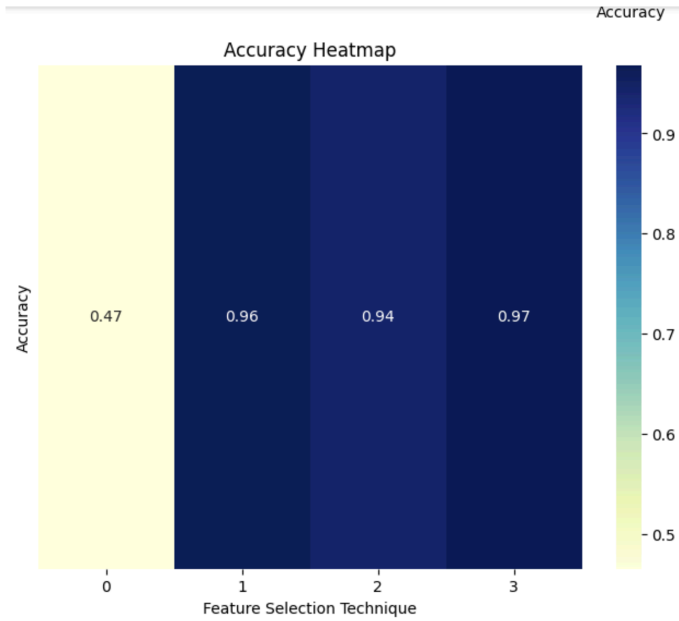
The "No Feature Selection" model is like a benchmark. It uses all the features and shows how well the classifier does without any feature selection. It got an accuracy of 46.54%, which is how good the classifier is when it uses all the features.

Feature selection methods like Particle Swarm Optimization (PSO), Genetic Algorithm (GA), and a combination of both (Hybrid), significantly improve classification accuracy. PSO and Hybrid excel with 96.49% and 96.83% accuracy respectively. This shows that PSO techniques are effective in identifying relevant features from MNIST data, leading to better classification. The Hybrid approach combines the benefits of PSO and GA, resulting in even higher accuracy by balancing global exploration with local refinement.



The "GA" method performs well in feature selection tasks, achieving an accuracy of 94.23%. Although it is simpler than Particle Swarm Optimization (PSO) and hybrid approaches, GA's effectiveness in optimizing classification models for complex datasets

such as MNIST showcases its importance in improving model performance.
Feature selection is crucial for boosting classification accuracy. Different optimization algorithms have their strengths. PSO, which excels in exploring the search space, is notably effective in selecting features for image classification. Combining PSO and GA yields even better results, showcasing the potential of combining optimization algorithms to tackle complex challenges.



To summarize, the graph's findings highlight the importance of feature selection in machine learning, especially with high-dimensional datasets like MNIST. PSO-based methods and hybrid approaches have shown significant promise in this regard, opening up new possibilities for research and application. Future studies should focus on refining and enhancing these techniques, testing them with different datasets and machine learning models, to drive advancements in feature selection and classification.

VIII. CONCLUSION

In this study, we compared various evolutionary computing techniques used for selecting features in machine learning. We used the MNIST dataset to evaluate the performance of Particle Swarm Optimization (PSO), Genetic Algorithm (GA), and their hybrid versions, along with traditional feature selection methods. Our goal was to determine how these techniques could improve the accuracy of machine learning classifiers.

Our research explored the performance of particle swarm optimization (PSO) and genetic algorithms (GA) in selecting features from the MNIST dataset. We examined the effectiveness of PSO, the strengths and weaknesses of GA, and the potential benefits of combining these techniques. Our results provide insights into the efficiency of these methods and their relevance to machine learning applications.

Experiments showed that particle swarm optimization (PSO) and hybrid techniques combining PSO and genetic algorithms (GAs) performed much better than older methods in correctly classifying data. PSO was very good at finding important features in the MNIST dataset, which greatly improved classification accuracy. The hybrid approach, which combined the strengths of PSO and GAs, did even better by combining a wide search with a more focused search.

GA performed well in selecting features, though its accuracy was slightly lower than PSO and hybrid methods. Despite this, GA's simplicity and ability to optimize complex datasets like MNIST highlighted its usefulness for feature selection. Our study enhances understanding of how evolutionary computing techniques work for feature selection in image recognition tasks using the MNIST dataset. The study offers insights into the advantages and drawbacks of different methods, providing guidance for researchers and professionals seeking to improve classification models for image recognition.

Our research shows that choosing the right features is essential for improving the accuracy of classification. We have also shown that evolutionary computing methods can be effective in this task. Future work should focus on improving and refining these methods, testing them on different datasets and machine learning models, and exploring new ways to combine them. By continuing to explore and innovate, we can improve feature selection techniques and help develop more accurate and understandable machine learning models.

IX. REFERENCE

1. A Comparative Study of Evolutionary ... - ACM Digital Library. (n.d.) retrieved March 24, 2024, from dl.acm.org/doi/abs/10.5220/0007948201310138
2. A Comparative Analysis of Different Feature Selection .... (n.d.) retrieved March 24, 2024, from ieeexplore.ieee.org/document/9606398
3. Analyzing the impact of feature selection methods on .... (n.d.) retrieved March 24, 2024, from www.nature.com/articles/s41598-023-49962-w
4. The Comparative Analysis of Single-Objective and Multi .... (n.d.) retrieved March 24, 2024, from link.springer.com/chapter/10.1007/978-3-030-19063-7_76
5. The comparative analysis of single-objective and multi- .... (n.d.) retrieved March 24, 2024, from khu.elsevierpure.com
6. A Comparative Analysis of Swarm Intelligence and .... (n.d.) retrieved March 24, 2024, from www.mdpi.com/2072-4292/14/13/3019
7. Comparative Analysis of Feature Selection Algorithms for .... (n.d.) retrieved March 24, 2024, from www.researchgate.net
8. A Comparative Study of Evolutionary Methods for Feature .... (n.d.) retrieved March 24, 2024, from www.semanticscholar.org
9. (PDF) A Comparative Study of Feature Selection Approaches. (n.d.) retrieved March 24, 2024, from www.researchgate.net