Journal of Advanced Sciences and Nanotechnology

Journal homepage: https://joasnt.com

Optimization-based feature selection and hybrid machine learning algorithms for big data classification: An investigation

Farah Sabah Khalaf [1,*], Essa Ibrahim Essa [1]

[1] College of Computer Science and Information Technology, University of Kirkuk, Iraq.

Article information

Article history:
Received: March, 20, 2023
Accepted: July, 11, 2023
Available online: Sept., 16, 2023

Keywords:
Big Data,
Data Classification,
Hybrid Optimization Algorithm,
Bear Smell Algorithm,
Xerus Algorithm.

Corresponding Author:
Farah Sabah Khalaf
stcha005@uokirkuk.edu.iq

Citation : Khalaf , F. S. ., & Essa, E. I. . (2023). Optimization-based feature selection and hybrid machine learning algorithms for big data classification: An investigation. *Journal of Advanced Sciences and Nanotechnology*, 2(2), 226–237.

Abstract

For the big data field, optimization algorithms aid in exploring and categorizing the vast data that cannot be dealt with since it is unstructured data. Therefore, extracting (meaningful) data is valuable for various applications. For this goal, many studies were proposed. However, this article presents a strategy for fast classifying massive data with less processing time. The main aim is to obtain data with higher accuracy with fewer mistakes. An appropriate optimization algorithm is selected to classify the data, extract the relevant features, and exclude the unwanted. These steps are conducted according to a study of the data state and the located environment. Additionally, we utilized the smell bear algorithm and Xerus algorithm as a hybrid optimization algorithm based on (hybrid machine learning feature selection), which helps to make the classification process successful and get high-level results. The verification with some of our counterparts demonstrates the superiority of our proposal.

DOI: https://doi.org/10.55945/joasnt.2023.2.2.226-237,
ISSN: 2791-0903/

1. Introduction

Because of the current technological development (such as the Internet of things, multimedia, educational universities, and electronic commerce) and other factors have prompted the increase in the amount of data in a threatening and frightening manner and at a high-speed rate. This increase in the number of data led to a new term known as (big data). Therefore, big data is unorganized, and unstructured data exceeds the typical capacity of storage. Therefore, this led to revealing four challenges: Data speed, data volume, data diversity and data change. Accordingly, managing or dealing with them is crucial. In many disciplines and applications, the goal here is how to process extensive data to reduce its size via extracting useful data and neglecting the [1-4]. The processing process is carried out by performing a data classification, which is a new statistical method for predicting user behavior in many areas of life. This classification process is carried out using an appropriate optimization algorithm based on (hybrid learning feature selection). Feature selection, which serves as the primary processing stage, selects subsets of characteristics from a big data collection and eliminates unnecessary or undesirable data. Exclusion of undesired elements (filter, embed, and cover) and inclusion of important features (as in [3] and [4]). Eventually, the principle of

feature selection with hybrid machine learning should achieve the following two goals:
1. Reduces the amount of data
2. Raises the performance and increases the accuracy of the outputs.
Consequently, we achieve structured classified data with high accuracy and less error.

Through our study, we can identify some points that contribute to the classification of big data, which are as follows:

a) Finding an appropriate optimization algorithm for classifying big data to be structured and organized (more useful).

b) Increasing the speed and reducing the processing time of extensive data and thus raising the performance level of the optimization algorithm for data classification.

c) Obtaining valuable data for developing new statistical models, techniques and methods can be utilized in different disciplines.

In the end, we can say that the aim of our study is thisprovide a service in the era of data inflation and information and data explosion. To achieve our aim, we explore and extract relevant data. Then, exclude unwanted ones by classifying data via an appropriate optimization algorithm to raise its performance.

2. Literature Review

In the literature, many researchers used hybrid optimization algorithms to classify data and raise the level of performance, as follows:

Sathyaraj et al. [5] utilized the Chicken swarm foraging algorithm for big data classification using the deep belief network classifier. The process classification of big data was performed using the dimensionality reduction method, which works on a new optimization algorithm named as chicken-based bacterial search algorithm (CBF). The algorithm comprises two stages: Training and examination. The algorithm shortens the work methods and reduces the time by using the technique of lowering dimensions and initial processing.

Awais et al. [6] proposed multilevel data processing using parallel algorithms for analyzing big data in high-performance prompting. The authors used the proposed algorithm with the improved dimensionality reduction method because it is further efficient than the Java iterator method. The experiments showed that it turned out that the algorithm used the HPC system to processing about 1 MB of data within half a second. Nevertheless, Java iteration steps spent significant periods processing the same data amount. Consequently, the data size has a significant role in any system. Still, the HPC system gives good results because, with the proposed algorithm, it takes a few seconds to process gigabytes of data.

Fateme et al. [7] suggested a new wrapper–filter methodology based on Hybrid genetic particle feature selection (GA-PSO) for feature subset selection. The combination of GA-PSO procedures tried to balance screening and consumption, yielding the best outcome. The authors named the algorithm as Smart HGP-FS Ave-Acc dataset. They proved that Smart HGP-FS outperformed its counterparts (wrapper approaches filtering techniques). As a dependable and active measure, the procedure helped explain advantage selection issues. The reason is that the hybrid approaches have shown the high efficiency of the presented algorithm in getting rid of irrelevant features. Additionally, improve the classification accuracy for the tested data. Consequently, the stability of the developed approach has been demonstrated.

Saba et al. [8] proposed a novel feature selection Method for classifying medical data using filters, wrappers, and embedded approaches. This study's primary goal was to investigate and identify a compelling feature selection. The classification was then carried out on the individual characteristics of the data using the support vector machine (SVM) classifier. The new framework employed several techniques for choosing features from filters, wrappers, and built-in algorithms. It is preserved from the university of California (UCI) data repository and uses two widely accessible standard datasets: the

Microarray and Cleveland datasets. The work utilized each technology's intersection of features with common characteristics to obtain outstanding performance, and SVM is then performed on the standard/shared attributes to provide the performance worker that further assesses the suggested model.

Masurah et al. [9] proposed an improved big data feature selection using a combination correlation-based feature selection. The paper introduced a different data mining technique, the correlation-based feature selection and dominance-based rough set approach (CFS-DRSA) hybrid approach. CFS-DRSA hybrid approach combines the CFS scheme with the BFS procedure and DRSA. The outcomes demonstrated that the hybrid CFS-DRSA approach is applicable to large data mining. Only the optimal attribute would be preserved and used in the decision-making since CFS-DRSA could identify unrelated and irrelevant characteristics throughout the result investigation process. Any researcher, or a decision worker, can cope with big data sets using the combination of CFS, DRSA, and the complete procedure, especially when there are hardware and software constraints. The findings revealed the necessity of creating and selecting an appropriate strategy for efficient decision-making.

Jianjiang et al. [10] proposed a map-balance-reduce and an improved parallel programming model for load balancing of MapReduce. It is a type of parallel programming that balances the load of the method of reducing dimensions and loses sight of experimental results when comparing this type of programming with the dimension reduction programming method. The possibility of a significant improvement in efficiency if programming was used Map-balance reduce (MBR) type with distributed standard data and abnormal.

Mahboubeh et al. [11] suggested an improved cost-sensitive representation of data for solving the big imbalanced data classification. In the study, a hybrid algorithm was proposed for the following goals. Reducing the data volume combines two methods, namely, identifying features. Additionally, taking related features to determine a solution to optimization problems without performing data processing. The results showed the superiority of the proposed hybrid algorithm method compared to other methods.

Laouni et al. [12] proposed analogous handling of the significant education base with the option of improving the final result of the expectation. From the findings, it turns out that our proposed method solved the obstacles to the prediction of large data, including the number and speed of extensive data. However, our process works negatively when the number of data is enormous, reaching thousands of totals, and this means illumination Learning that our strategy is working at total capacity. Therefore this makes it challenging to perform data processing through the computer.

Waad et al. [13] presented distributed evolutionary feature selection for big data processing. The authors used a new method for feature selection that depends on a genetic algorithm through parallel processing. The approach in this research works on dividing the data sets into sub-parts called islands that they work on using the dimensionality reduction method and the possibility of scaling the method GA with Amazon.

Loris et al. [14] proposed programming big data analysis: principles and solutions. They presented an analysis and match of the most used programming types for significant data analysis and the attributes of the leading program frames that work with them. In fact, these systems were matched based on more than a scale about three things: their characteristics, method of spread, advantages and disadvantages. Every system's analysis process is determined by investigating the type of programming and instructions to determine the advantages and disadvantages of each system.

Miguel et al. [15] investigated some important approaches in parallel programming in bioinformatics. Bioinformatics enables many types of parallel programming and activities around parallel programming. In computer and biological information, the authors found many problems that are often affected by the processing of extensive data and the fullness of memory and capacity storage due to the large numbers of biological data and the many related problems. Biological information is one of the most exciting types of research in which parallel programming can apply all its properties. For example, green computing, assembly computing, and privileged computing, where the computing has more than one core. In this study, change vector analysis was performed through the area of the spectral angle graph and the black threshold to provide an excellent method for such a type of software In addition to active applications

for the departments that process graphics. Therefore, we note that it offers a set of solutions that facilitate the programming process, which takes advantage of the outstanding performance of the new parallel computing.

Tassadaq et al. [16] employed the hybrid parallel processing/message transit interface parallel programming paradigms). On the Nord-III supercomputer, the authors created the serialized IPUM C program. On multiple distributed nodes of the Nord-III supercomputer at the Barcelona supercomputing facility, they tested the approach. They used comparable hybrid IPUM on a variety of processor cores to do that, making the performance and expandability of the strategy very flexible to a variety of required speeds. The authors plan to use a heterogeneous computing architecture with GPU and FPGA accelerators in the future to put IPUM into practice. The C co-estimation in the IPUM implementation process demands the greatest calculation and time.

Marowka et al. [17] presented a special section on parallel programming. They demonstrated the use of parallel programming to resolve practical computing issues and described a novel method to stencil on a subscriber the synchronization of data flow in parallel accounts.

Marouane et al. [18] studied the perspective of programming languages, MapReduce-depend on high level query language (HLQL) and offered insights like the ease of programming and customization to relate to Hadoop. They created a baseline for comparison to evaluate the expressiveness of MapReduce-based HLQL and assess whether the performance cost of providing more abstract languages is worth it. The best way to integrate native Structured query language (SQL) query processing at the top of MR has actually been found to be Big SQL. Pig Latin, on the other hand, showed his expressive limitations. Pig and JAQL performed the same in most benchmarks when the input size was increased. Despite having the highest proportion of source lines for code size among the others, Hive provides more comparable performance to Big SQL.

Majdi et al. [19] proposed three hybrid techniques in this research that perform noticeably better on the datasets than Grey wolf optimizer (GWO) and whale optimization algorithm (WOA). They explained that the hybrid approaches could demonstrate a higher modified rapprochement than the essential procedures in many datasets. The serialization approach and adaptive switching can improve classification accuracy on the primary and randomized exchanging methods.

Mushahhid et al. [20] presented the automation of global optimization and analog circuit design using a combination of the Whale optimization algorithm (WOA) and the metaheuristics grey wolf optimizer (mGWO) algorithms. They combined the strengths of the whale optimization algorithm and the modified gray wolf optimization algorithm, which yields the hybrid WOA-mGWO algorithm, intended to boost the algorithm's exploratory capability. To boost confidence and show its deep participation in the current state of the art, the best, worst, and worst solutions to the analog IC scaling challenge are offered.

A feature selection based on an improved artificial hummingbird algorithm using random opposition-based learning for Solving waste classification problems was proposed by Mona et al., [21]. The results obtained from this study proved the superiority of the two algorithms over other evaluation algorithms. Using random learning, the waste classification algorithms AHA-ROBL and AHA-OBL provided the necessary and ideal number of selected characteristics with the highest accuracy [21].

Using the enhanced grey wolf optimizer and SVM algorithms for breast cancer detection based on feature selection were proposed by Sunil et al. [22]. It is encouraged to utilize machine learning techniques to aid in the early detection of breast cancer tumors since early detection of breast cancer can minimize mortality. The employment of multiple techniques has further increased the precision of categorization. To achieve better classification results with less noise, this work introduces the improved GWO-SVM method, which utilizes the pros of the enhanced GWO algorithm to select a subset of the features [22].

Improved fruit fly optimization algorithm incorporating tabu search for optimizing the selection of elements in trusses presented by Yancang et al. [23]. The drosophila optimization algorithm was applied as a more efficient optimization technique. Although the drosophila optimization algorithm outperforms more widely used algorithms, it has a critical flaw for which we suggested a superior solution depends on dynamic search. The tube-type carrier structure is optimized using the Drosophila optimization algorithm

discussed above. The Drosophila optimization algorithm's stability and efficiency have increased according to the optimization and comparison results with other evaluation algorithms [23].

This article presents the Sandpiper Optimization Algorithm (SOA), a novel bio-inspired algorithm. This method has been primarily inspired by the sand bird's migration and attack patterns. It combines another method, a decision tree machine learning algorithm, with this approach to address practical applications. The testing outcomes demonstrated that the suggested method outperforms the most recent optimization algorithms and can solve challenging local optimization issues [24].

Beetle swarm optimization and an adaptive neuro-fuzzy inference system for analyzing large datasets for several diseases have been proposed by Parminder et al. [25]. They employed a strategy combining beetle swarm optimization with an adaptive ambiguous neural inference system to comprehensively study the interplay between many cardiovascular diseases and health care costs. BSO-ANFIS is compared to HODBNN, HRFLM, v2 DNN, and ICA to determine which method offers the best results (MH). Their findings show that BSO-ANFIS is superior to other approaches already in use [25].

Hongwei et al. [26] proposed a novel hybrid algorithm based on Particle swarm optimization PSO and FOA for target searching in unknown environments. A method called the hybrid multi-swarm FOA-PSO (MFPSO) was submitted in this paper to search for the target by the robot. To confirm MFPSO's optimal performance, several experiments are carried out in four sections. With several swarms of robots and expansive areas, MFPSO performs noticeably better than previous techniques (A-RPSO and RPSO). In most situations, the suggested method is more effective, more successful, and requires les s repetition. Future research on issues with numerous goals is something the authors try to investigate.

Malik et al. [27] studied the capuchin search algorithm as a novel meta-heuristic search algorithm for solving optimization problems. The study discussed a new meta-search strategy for resolving well-known optimization issues inspired by nature. The suggested method drew a significant amount of inspiration from the foraging habits of capuchin monkeys. The authors confirmed that CapSA has numerous benefits over other optimization algorithms and is best used to address other common and actual issues with complicated search spaces based on simulation results, results, analysis, and statistical testing. The suggested optimization technique offers a fundamental foundation for rather low-dimensional optimization problems, which may be expanded to address significant local optimization situations.

Hybrid annealing krill heerd and quantum-behaved particle swarm optimization (QPSO) suggested by Cheng et al. [28]. Since QPSO performs better in exploitation and AKH performs better in exploration, AKQPSO suggested improving population diversity based on these findings, and it performs better in exploitation and exploration. The best local value allows the algorithm to be improved [28].

Chitrakant et al. [29] presented an analysis of Bayesian optimization procedures for big data classification based on the map-reduced framework. In this work, a Correlative naïve bayes (CNB) classifier was employed as the foundational method, supplemented with optimization techniques like cuckoo search and gray wolf optimization. Compared to CNB and CGCNB classifiers, adopting fuzzy theory with a CNB classifier that incorporates an organic degree of qualities into the data set leads to performance advancements. The authors stated that in the future, classifier performance would be evaluated using log loss and training loss utilizing models like the FCNB, HCNB, and CGCNB.

Siyu et al. [30] presented Multi-kernel optimized relevance vector machine for probabilistic prediction of concrete dam displacement. The authors used main service life metrics obtained from a massive-height concrete arch dam in this research. The sample is integrated with Relevance vector machine (RVM), multi-core technology, hydrostatic season time (HST), and parallel java algorithm (PJA) statistical models. It was established that the suggested parameter optimization approach helps RVM provide reliable and high-accuracy forecasts. Developing ORVM and testing the impact of several kernel functions on predictive performance, such as Gaussian kernel, Laplace kernel, Polynomial kernel, and multi-kernel for their weighted blend. The proposed ORVM model has good performance in both training and test sets, making it ideal for predicting the non-stationary and nonlinear displacement of concrete embankments, regarding the model's endurance and predictive power [30].

Fei et al. [31] proposed hyperspectral image classification based on multiple reduced kernel

extreme learning machines. This work presented a technique for categorizing hyperspectral pictures depending on a scaled-down multi-core extreme learning machine. The proposed MRKELM has the benefit of both mini-kernel severe machine learning and multi-core learning compared to the present reduced kernel powerful learning technique machine, which achieves the multi-kernel stage when multi-kernel learning is added. The intricate link between the threatened groups and the hyperspectral picture is modeled using MRKELM. The outcomes show how successful the suggested MRKLM-based hyperspectral image categorization approach [31].

Shankar et al. [32] suggested an optimal feature-based multi-kernel SVM approach for thyroid disease classification. Any person's therapeutic science must include the classification of the thyroid gland. Rely on IGWO approach features have been chosen, built-in feature selection and classification approaches have been superiorly demonstrated in the thyroid data set. A data classification process was used to improve medicine, decision making and disease detecting. This study used an improved uti123 technique based on multicore SVM features to classify perturbations, 98.65% is MKSVM's accuracy compared to another classifier. The performance evaluation showed higher specificity, greater accuracy, and other metrics. Compared to previous works, the MKSVM classifier's use for data sets on thyroid disorders produces the most significant results [32].

Wasiat et al. [33] presented stock market prediction using machine learning classifiers and social media news. The model presented in this study used news and social media as external inputs to forecast future stock market patterns. Additionally, the authors demonstrated that after the third day, most classifiers' overall accuracy improved while the highest accuracy declined due to a mix of financial news and social media mood. They provided many data kinds and forecasting techniques in this paper. When they looked at how feature selection and spam reduction impacted the performance of the algorithms used for prediction, the authors discovered that they positively impacted the majority of the classifiers we chose. Additionally, they found that RF consistently produces accurate results, making it a good choice for forecasting stock trends.

Yao et al. [34] proposed an efficient hybrid Kernel extreme learning machine approach for early diagnosis of Parkinson's Disease. To solve the issue of PD diagnosis, the authors developed a practical mRMR-KELM hybrid approach in this work. A KELM classifier with feature selection technology, specifically an mRMR filter that boosts the performance of a KELM classifier with much fewer features, is the main component of the suggested strategy. The positive results from the PD data set have shown that the proposed hybrid technique may successfully identify the difference between people with Parkinson's and healthy individuals. It has been demonstrated that mRMR-KELM can classify objects with an accuracy of 96.47%. Empirical research has shown that the suggested diagnostic approach can help doctors make correct diagnostic decisions.

Multiple composite kernel extreme learning machine for hyperspectral images were studied by Ugur et al. [35]. This study utilized a novel approach utilizing an ELM-based MKL algorithm to combine CKs and HKs. Here, the authors contrasted the technique with some contemporary CK and MKL methods. The findings show that MCKELM is uniquely situated concerning the others. Contrary to traditional CK techniques, it does not necessitate extensive optimization activities, making it a more affordable option. Additionally, the spatial and spectral components do not require manual arrangement because the two domains are connected automatically by the MCK-ELM [35].

Yu et al. [36] investigated multi-kernel extreme machine learning for EEG classification in brain-computer interfaces. In this research, the authors categorized EEG in BCIs using an ELM-depend multi-core technique. The double non-linear feature spaces are integrated into two kernel variations with multi-core learning to produce a more accurate EEG classification. The MKELM approach is a leading choice for creating an upgraded MI-based BCI due to its higher performance, as evidenced by testing data. Additionally, our suggested technique may be easily adapted to various applications and is not just restricted to BCI [36].

Deterministic Multi-kernel based extreme ML for pattern classification proposed by Bhawna et al. [37]. Through the development of a multi-core learning dependent on ELM with multimodal feature

extraction utilizing GLCM, this paper explores current time pattern recognition issues. Analytically, the input weights and biases were determined to solve this problem. The authors employed multimedia feature extraction to improve the precision of evaluating the suggested technique. It is evident from the data analysis that the kernel version of ELM is more effective than ELM. OMKELM produces findings that are more precise than those of single kernel methods. The given DMK variations are superior to ELM, KELM, and OMKELM, according to our analysis of the experimental findings. The proposed algorithms aim to get beyond ELM's non-deterministic character and increase its accuracy and efficiency by working with upgraded kernels [37].

Farnood et al. [38] proposed Xerus Optimization Algorithm (XOA) as a novel nature-inspired metaheuristic algorithm for solving global optimization problems. The findings demonstrated the XOA algorithm's incredible performance, flexibility, and competitiveness on this broad range of criteria. The work can be expanded to address issues in optimization's discrete and synthesis areas. Also, the XOA technique can be applied to address the problems in those areas.

Ibrahim et al. [39] introduced an effective parallel reptile search procedure and snake optimizer method for feature selection. The proposed filtering solutions showed that increased walking in RSA is more efficient than SO. Additionally, the latest 25 iterations show that the RSA fishing support approach fully utilizes candidate solutions more than SO. The new RSA-SO had the lowest value for the OFS specified For nine out of 12 sets of data sets, making belly walking exploration and exploitation utilizing the RSA hunting style ineffective. This attested to the effectiveness of the suggested RSA-SO in getting rid of an undesired characteristic.

A survey on parallel clustering algorithms for Big Data was conducted by Zineb et al. [40]. This research showed that most methods discussed are related to aggregation algorithms, including K-Means, DBSCAN, and OPTICS. In fact, these algorithms are suitable for parallelism. Large data platforms have decreased their number in aggregation, such as one-to-one networks, due to advances in parallelism and computing [40].

Advances in significant data programming, system software and HPC convergence were investigated by Ching et al. [41]. This study explained the handling of technical problems in big data gates, information design, or proposing new strategies in data departments and HPC fields. It also led to a lot of related research. This study is a good learning example and reference for anyone interested in big data and HPC system [41].

Noha et al. [42] investigated feature selection in big data preprocessing based on a hybrid cloud-based model. The proposed method shows an excellent job compared to other concurrent methods. The proposed method combines the measurement of the unusual distance with the average measurable distance used for k, which Bohr has approximated. The results gave a new insight into how time and features are used compared to the nearest neighbors. It has been shown that there is a 12% improvement in the classification process compared to the other convergence algorithm because this method reduces the processing time.

Omid et al. [43] combined the optimization approach and machine learning. The authors suggested a workbook optimal feature selection for SAR image classification using biogeography-based optimization (BBO), artificial bee colony (ABC) and SVM. The article classified the land cover using the suggested HBBOSVM classifier. The best features were chosen using the ABC method in conjunction with the BBO relay factor, and the pictures were then classified using an SVM classifier. The findings of the HBBOSVM algorithm were then compared to those of the BBOSVM, ABCSVM, and PSOSVM algorithms and methods from earlier studies. HBBO performance was evaluated in light of 20 accepted norms. The findings demonstrated that the HBBOSVM is also superior to the competition in terms of class accuracy and kappa coefficient. HBBOSVM also displayed high stability and affinity. According to the findings, the HBBO is a practical university resource for measuring issues. The study showed that using both machine learning and optimization techniques together increases power.

A new hybrid PSO and parallel variable neighborhood search algorithm for flexible job shop scheduling with a get-together method were proposed by Parviz et al. [44]. This research provided a new

hybrid technique and mathematical model for the loose workshop scheduling issue with aggregation procedures. Each product is made by putting together a collection of various pieces. The new hybrid particle swarm has been enhanced, and the HPSOPVNS Neighbor Variable Parallel Search (HPSOPVNS) technique has been suggested because the task is not challenging. In this hybrid method, the (PVNS) procedure was utilized for local search close to solutions found in each iteration. Additionally, the PSO procedure is used for global search space exploration. The Taguchi technique was used to modify the metaheuristic algorithm parameters, PSO Algorithm, Variable Neighborhood Search Algorithm (HPSOVNS), Hybrid Genetic Algorithm, and Swarm Optimization.

Asen Toshev [45] proposed using PSO and Tabu search hybrid algorithm for flexible job shop scheduling problem-analysis of test results. The work used a hybrid PSO-TS method to resolve a flexible JSSP. The results of the first, a PSO, and the second, a TAPU investigation, indicated that the error rate is a promising 0.044%. The authors observed a little delay until the findings were visible for a few seconds. The proposed algorithm can be put to the test in more scenarios with more dimensions so that it can perform better.

Olatunji et al. [46] reviewed multiclass feature selection with metaheuristic optimization. To apply metaheuristic algorithms to overcome the multi-layer feature selection process issues, a shell-focused evaluation is done in this study. The authors searched for research on multi-layer feature selection challenges but were unable to locate any. There have been several methods proposed to enhance metaheuristic performance. Two or Larger binary strategies were found for addressing multi-layer issues from work performed on multi-layer feature selection. Specific metatheurism algorithms can answer high-level feature selection problems more efficiently than others.

Ali et al. [47] proposed using a novel nature-inspired metaheuristic procedure for optimization. The procedure was based bear smell search algorithm. The BSSA method was suggested for use in this investigation. Different types of standard careers and geometric problems are used to demonstrate and assess the effectiveness of the BSSA procedure. The BSSA algorithm's outcomes have been compared with those of other optimization techniques. The types of indicators analyzed, such as pairs test, Wilcoxon rank, and statistical analysis. The numerical outcomes demonstrated that the suggested BSSA gives competitive and superior results (when compared to previous optimization methods).

Sathish et al. [48] improved the black widow-bear smell search algorithm (IBWBSA) for optimal planning and operation of distributed generators in the distribution system. Submit this work with a mixed strategy for multi-objective planning and optimization of the operation of distributed generators (DGs) on a distribution system (DS). The IBWBSA technology mixes the Black Widow BWOA algorithm and the BSSA's exclusive bear scent search method that accelerates convergence. The findings show that with the best DS planning and operation, the multi-objective strategy reduces generating costs, energy loss, and voltage variation.

Rung et al. [49] compared choosing critical features for data classification based on ML approaches. In this paper, four feature selection Techniques-Random Forest (RF), (SVM), (KNN), and (LDA) were compared. Based on the performance of each feature selection technique, the optimum feature selection technique is then chosen. The importance of feature selection for categorizing data has been proven through experimentation. The experiment outcomes demonstrated that random forest is the most accurate classifier. RFE and RF techniques are highly effective for feature selection.

Nico et al. [50] proposed sleep stage classification using extreme learning machine and particle swarm optimization for big healthcare data. This study used three different types of algorithms, four different sets of categories, and two sets of characteristics to classify the different sleep phases based on the heart rate variability of the ECG data. The algorithm ELM, SVM, and ELM integration with PSO were utilized. The research demonstrated that ELM and PSO integration, followed by ELM and SVM integration, had the best accuracy. We may conclude that PSO integration enhances the ELM and SVM algorithm's accuracy.

Optimal placement and sizing of shunt capacitors in radial distribution system using polar bear optimization algorithm by Muhammad et al. [51]. Due to the rise in living standards, it is more difficult

for distribution system operators to minimize energy losses while preserving the voltage profile. The system may work better by resolving electricity quality problems and strengthening numerous technical and financial factors. In this study, the placement of capacitors and optimal scaling in a radial distribution system (RDS) under various loading situations were solved using the optimization method (PBOA). Actual power loss (PLoss) expenses and different SC costs were included in this paper. This technique was tested on several IEEE standard bus systems with various loads. While preserving allowable voltages on the system buses, the proposed PBOA lowers AOC and P RDS losses [51].

A hybrid optimization algorithm for water volume adjustment problem in district heating systems was proposed by Yi Han et al. [52]. In this research, the valve angle (VA) in each HES was improved using a hybrid polar bear optimization method in conjunction with chemical reaction optimization (HA-PBO-CRO). Comparing HAPBO-CRO and a non-dominant genetic screening algorithm (NSGAII) revealed that HAPBO-CRO is superior to NSGAII with better Pareto limits. Additionally, it offers management in a CHP plant an essential reference to help management decide whether to reduce the energy consumed to meet customer requirements [52].

Diaz et al. [53] analysis work a comparative analysis of meta-heuristic optimization algorithms for feature selection and feature weighting in neural networks. In this work, the use of identification methods for advantages selection and advantages weighting has been demonstrated, along with the advantages of these algorithms for enhancing FFNN accuracy and shortening computation times. In various datasets, experimental results show that FFNN with advantages selection and advantages weight has higher classification accuracy than existing algorithms. The meta-optimization algorithms used for comparison are found to be efficient in advantages selection and advantages weighting processes [53]. Inam and Idress [54] proposed new strategy adopted is to create a new dataset that results from combining the two databases. The test testing revealed that training using a mixture dataset outperformed using individual datasets in terms of metrics values, particularly when inter-dataset evaluation was used to solve the generalization problem.

3. Conclusion

This paper investigated finding a suitable optimization algorithm that helps explore and classify the big data we cannot deal with (i.e., unstructured). Consequently, extracting (meaningful) data that is useful from it that we can use in many applications. In this study, we were able to find many previous studies related to our study. This study aimed to illustrate a method for classifying large data with high speed and less processing time to obtain data with more accuracy and less error. Using a hybrid optimization algorithm based on feature selection and a hybrid learning function, we studied the state of the data and the environment in which it is located. Additionally, the investigation was based on choosing an appropriate optimization algorithm to classify the data, extract the relevant features and exclude the unwanted. Then, compare the results obtained from the chosen algorithm with other optimization methods available. We noted a significant improvement in the performance of the algorithm. In the end, we recommend in the future to research and conduct studies on hybrid optimization algorithms with the use of the best properties to obtain the best results.

Acknowledgement

This is an optional section.

Conflict of Interest

The authors declare that they have no conflict of interest.

References

[1] B. Tran, B. Xue and M. Zhang, "Variable-Length Particle Swarm Optimization for Feature Selection on High-Dimensional Classification," in IEEE Transactions on Evolutionary Computation, vol. 23, no. 3, pp. 473-487, June 2019, doi: 10.1109/TEVC.2018.2869405.

[2] A. Kaur , S. Jain & S. Goel , "Sandpiper optimization algorithm: a novel approach for solving real-life engineering problems," Appl Intell, vol. 50, no.2 , pp. 582–619 , August 2019, doi : 10.1007/s10489- 019-01507-3.

[3] A. Thakkar, R. Lohiya, "A survey on intrusion detection system: feature selection, odel, performance measures, application perspective, challenges, and future research directions" Artif Intell Rev , vol .55, no.1, pp. 453–563, July 2021, doi: org/10.1007/s10462-021-10037-9 .

[4] RC. Chen, C. Dewi, SW. Huang, "Selecting critical features for data classification based on machine learning methods", J Big Data , vol.7, no. 52 , July (2020) , doi.org/10.1186/s40537-020-00327-4.

[5] Sathyaraj R., Ramanathan L., Lavanya K., Balasubramanian V. & Saira Banu J. (2020). Chicken swarm foraging algorithm for big data classification using the deep belief network classifier." Data Technologies and Applications, vol. 55, no. 3, pp. 332-352, 2020, doi: 10.1108/dta-08-2019-0146.

[6] Awais Ahmad, Anand Paul, Sadia Din, M. Mazhar Rathore, Gyu Sang Choi & Gwanggil Jeon (2018). Multilevel Data Processing Using Parallel Algorithms for Analyzing Big Data in High-Performance Computing. Int. J. Parallel Program. 46, 3 (June 2018), 508–527. https://doi.org/10.1007/s10766-017-0498-x.

[7] Fateme Moslehi & Abdorrahman Haeri (2019). A novel hybrid wrapper–filter approach based on genetic algorithm, particle swarm optimization for feature subset selection." *Journal of Ambient Intelligence and Humanized Computing*, vol. 11, no. 3, pp. 1105-1127, 2019, doi: 10.1007/s12652-019-01364-5.

[8] Saba Bashir, Irfan Ulla Khattak, Aihab Khan, Farhan Hassan Khan, Abdullah Gani & Muhammed Shiraz (2022). A Novel Feature Selection Method for Classification of Medical Data Using Filters, Wrappers, and Embedded Approaches. *Complexity*, vol. 2022, pp. 1-12, 2022, doi: 10.1155/2022/8190814.

[9] Mohamad M., Selamat A., Krejcar O., Crespo RG, Herrera-Viedma E, & Fujita H. (2021). Enhancing Big Data Feature Selection Using a Hybrid Correlation-Based Feature Selection. *Electronics*. 2021; 10(23):2984. https://doi.org/10.3390/electronics10232984.

[10] Jianjiang Li, Yajun Liu, Jian Pan, Peng Zhang, Wei Chen & Lizhe Wang (2020). Map-Balance-Reduce: An improved parallel programming model for load balancing of MapReduce. Future Gener. Comput. Syst. 105, C (Apr 2020), 993–1001. https://doi.org/10.1016/j.future.2017.03.013.

[11] Mahbobeh Fattahi, Mohammed Hossein Moattar, & Yahya Forghani (2022). Improved cost-sensitive representation of data for solving the imbalanced big data classification problem. *Journal of Big Data*, vol. 9, no. 1, 2022, doi: 10.1186/s40537-022-00617-z.

[12] Laouni Djafri, Djamel Amar Bensaber, & Reda Adjoudj (2018). Big Data analytics for prediction: parallel processing of the big learning base with the possibility of improving the final result of the prediction. *Information Discovery and Delivery*, vol. 46, no. 3, pp. 147-160, 2018, doi: 10.1108/idd-02-2018-0002.

[13] Waad Bouaguel & Chiheb Eddine Ben NCir (2022). Distributed Evolutionary Feature Selection for Big Data Processing. *Vietnam Journal of Computer Science*, vol. 9, no. 3, pp. 313-332, 2022, doi: 0.1142/s2196888822500154.

[14] Loris Belcastro, Riccardo Cantini, Fabrizio Marozzo, Alessio Orsino, Domenico Talia & Paolo Trunfio (2022). Programming big data analysis: principles and solutions. *J Big Data* 9, 4 (2022). https://doi.org/10.1186/s40537-021-00555-2.

[15] Miguel A. Vega -Rodríguez & Jose. M. Granado-Criado (2018). Parallel Programming in Bioinformatics: Some Interesting Approaches. *International Journal of Parallel Programming*, vol.47, no.2, pp. 293-295,2018, doi: 10.1007/s10766-018-0605-7.

[16] Tassadaq Hussain, Saqib Amin, Usman Zabit & Eduard Ayguadé (2021). Implementation of a high-accuracy phase unwrapping algorithm using parallel-hybrid programming approach for displacement sensing using self-mixing interferometry. *The Journal of percomputing*, vol. 77, no. 9, pp. 9433-9453, 2021, doi: 0.1007/s11227-021-03634-6.

[17] Marowka, A. & Stpiczyński, P (2018). Special section on parallel programming. *J Supercomput* 74, 1419–1421 (2018). https://doi.org/10.1007/s11227-018-2278-9.

[18] Marouane Birjali, Abderrahim Beni-Hssane, & Mohammed Erritali (2018). Evaluation of high-level query languages based on MapReduce in Big Data. *Journal of Big Data*, vol. 5, no. 1, 2018, doi: 10.1186/s40537-018-0146-3.

[19] Majdi Mafarja, Asma Qasem, Ali Asghar Heidari, Ibrahim Aljarah, Hossam Faris & Seyedali Mirjalili (2019). Efficient Hybrid Nature-Inspired Binary Optimizers for Feature Selection. *Cognitive Computation*, vol. 12, no. 1, pp. 150-175, 2019, doi:10.1007/s12559-019-09668-6.

[20] Mushahhid A. Majeed. & Patri S. R. (2019). A hybrid of WOA and mGWO algorithms for global optimization and analog circuit design automation. *COMPEL - The international journal for computation and mathematics in electrical and electronic engineering*, vol. 38, no. 1, pp. 452-476, 2019, doi: 10.1108/compel-04-2018-0175.

[21] Ali, Mona A. S., Fathimathul Rajeena P. P., and Diaa Salama Abd Elminaam. 2022. "A Feature Selection Based on Improved Artificial Hummingbird Algorithm Using Random Opposition-Based Learning for Solving Waste Classification Problem" *Mathematics* 10, no. 15: 2675. https://doi.org/10.3390/math10152675.

[22] Sunil Kumar & Maninder Singh (2020). Breast Cancer Detection Based on Feature Selection Using Enhanced Grey Wolf Optimizer and Support Vector Machine Algorithms. *Vietnam Journal of Computer Science*, vol. 8, no. 2, pp. 177-197, 2020, doi: 10.1142/s219688882150007x.

[23] Yancang Li & Sida Lian (2018). Improved Fruit Fly Optimization Algorithm Incorporating Tabu Search for Optimizing the

Selection of Elements in Trusses. *KSCE Journal of Civil Engineering*, vol. 22, no. 12, pp. 4940-4954, 2018, doi: 10.1007/s12205-017-2000-0.

[24] Amandeep Kaur, Sushma Jain & Shivani Goel (2020). Sandpiper optimization algorithm: a novel approach for solving real-life engineering problems. Applied Intelligence 50, 2 (Feb 2020), 582–619. https://doi.org/10.1007/s10489-019-01507-3.

[25] Parminder Singh, A. Kaur, R. S. Batth, S. Kaur & G. Gianini (2021). Multi-disease big data analysis using beetle swarm optimization and an adaptive neuro-fuzzy inference system. *Neural Computing and Applications*, vol. 33, no. 16, pp. 10403-10414, 2021, doi: 10.1007/s00521-021-05798-x.

[26] Hongwei Tang, Wei Sun, Hongshan Yu, Anping Lin, Min Xue & Yuxue Song (2019). A novel hybrid algorithm based on PSO and FOA for target searching in unknown environments. Applied Intelligence 49, 7 (July 2019), 2603–2622. https://doi.org/10.1007/s10489-018-1390-0.

[27] Malik Braik, Alaa Sheta, & Heba Al-Hiary (2020). A novel meta-heuristic search algorithm for solving optimization problems: capuchin search algorithm. *Neural Computing and Applications*, vol. 33, no. 7, pp. 2515-2547, 2020, doi:10.1007/s00521-020-05145-6.

[28] Cheng Wei L.; Wang, G.-G. Hybrid Annealing Krill Herd and Quantum-Behaved Particle Swarm Optimization. *Mathematics* 2020, *8*, 1403. https://doi.org/10.3390/math8091403.

[29] Chitrakant Banchhor, & Srinivasu N (2020). Analysis of Bayesian Optimization Algorithms for Big Data Classification Based on Map Reduce Framework, 23 December 2020, PREPRINT (Version 1) available at Research Square [https://doi.org/10.21203/rs.3.rs-132776/v1.

[30] Siyu Chen, Chongshi Gu, Chaoning Lin, Kang Zhang & Yantao Zhu (2020). Multi-kernel optimized relevance vector machine for probabilistic prediction of concrete dam displacement. *Engineering with Computers*, 2020, doi: 10.1007/s00366-019-00924-9.

[31] Fei Lv, & Min Han (2019). Hyperspectral image classification based on multiple reduced kernel extreme learning machine." *International Journal of Machine Learning and Cybernetics*, vol. 10, no.12, pp. 3397-3405, 2019, doi: 10.1007/s13042-019-00926-5.

[32] Shankar L., S. K. Lakshmanaprabu, D. Gupta, A. Maseleno & V. H. C. De Albuquerque (2018). Optimal feature-based multi-kernel SVM approach for thyroid disease classification." *The Journal of Supercomputing*, vol. 76, no. 2, pp. 1128-1143, 2018, doi: 10.1007/s11227-018-2469-4.

[33] Wasiat Khan, Mustansar Ali Ghazanfar, Muhammad Awais Azam, Amin Karami, Khaled H. Alyoubi & Ahmed S. Alfakeeh (2020). Stock market prediction using machine learning classifiers and social media, news. *Journal of Ambient Intelligence and Humanized Computing*, vol. 13, no. 7, pp. 3433-3456, 2020, doi: 10.1007/s12652-020-01839-w.

[34] Yao Wei Fu, Chen, HL., Chen, SJ., Li, LJ., Huang, SS., Cai, ZN. (2014). A Hybrid Extreme Learning Machine Approach for Early Diagnosis of Parkinson's Disease. In: Tan, Y., Shi, Y., Coello, C.A.C. (eds) Advances in Swarm Intelligence. ICSI 2014. Lecture Notes in Computer Science, vol 8794. Springer, Cham. https://doi.org/10.1007/978-3-319-11857-4_39.

[35] Ugur Ergul & G. Bilgin (2017). Hyperspectral image classification with hybrid kernel extreme learning machine. *2017 25th Signal Processing and Communications Applications Conference (SIU)*, 2019, doi: 10.1109/siu.2017.7960244.

[36] Yu Zhang, Yu Wang, Guoxo Zhou, Jing Jin, Bei Wang, Xingyu Wang & Andrzej Cichocki (2017). Multi-kernel extreme learning machine for EEG classification in brain-computer interfaces. *Expert Systems with Applications*, vol. 96, pp. 302-310, 2018, doi : 10.1016/j.eswa.2017.12.015.

[37] Bhawna Ahuja & Virendra P. Vishwakarma (2021). Deterministic Multi-kernel based extreme learning machine for pattern classification. Expert Syst. Appl. 183, C (Nov 2021). https://doi.org/10.1016/j.eswa.2021.115308.

[38] Farnood Samie Yousefi, Noushin Karimian & Amin Ghodousian (2019). Xerus Optimization Algorithm (XOA): a novel nature-inspired metaheuristic algorithm for solving global optimization problems. *Algorithems and computation*, vol. 51, no. 2, pp. 111-126, 2019.

[39] Ibrahim Al-Shourbaji, Kachare Ph, Alshathri S, Duraibi S, Elnaim B & Abd Elaziz M. (2022). An Efficient Parallel Reptile Search Algorithm and Snake Optimizer Approach for Feature Selection. *Mathematics*. 2022; 10(13):2351. https://doi.org/10.3390/math10132351.

[40] Zineb Dafir, Yasmine Lamari & Said Chah Slaoui (2020). A survey on parallel clustering algorithms for Big Data. *Artificial Intelligence Review*, vol. 54, no. 4, pp. 2411-2443,2020, doi:10.1007/s10462-020-09918-2.

[41] Ching Hsien Hsu, Geoffrey Fox, Geyong Min, & Sugma Sharma (2019). Advances in big data programming, system software and HPC convergence. *J Supercomput* 75, 489–493 (2019). https://doi.org/10.1007/s11227-018-2706-x.

[42] Noha Shehab, Mahmoud Badawy & H. Arafat Ali (2021). Toward feature selection in big data preprocessing based on hybrid cloud-based model. *The Journal of Supercomputing*, vol. 78, no. 3, pp. 3226-3265, 2021, doi:10.1007/s11227-021-03970-7.

[43] Omid Rostami & Mehrdad Kaveh (2021). Optimal feature selection for SAR image classification using biogeography-based optimization (BBO), artificial bee colony (ABC) and support vector machine (SVM): a combined approach of optimization and machine learning. *Computational Geosciences*, vol. 25, no. 3, pp. 911-930, 2021, doi: 10.1007/s10596-020-10030-1.

[44] Parviz Fattahi, N. Bagheri Rad, F. Daneshamooz & S. Ahmadi (2020). A new hybrid particle swarm optimization and parallel variable neighborhood search algorithm for flexible job shop scheduling with assembly process. *Assembly Automation*, vol. 40, no. 3, pp. 419-432, 2020, doi: 10.1108/aa-11-2018-0178.

[45] Asen Toshev," Particle Swarm Optimization and Tabu Search Hybrid Algorithm for Flexible Job Shop Scheduling Problem

– Analysis of Test Results", Cybernetics And Information Technologies, Volume 19, No. 4, 2019, DOI: 10.2478/cait-2019-0034

[46] Olatunji O. Akinola, Akinola E. Ezugwu, Jeffrey O. Agushaka, Raed Abu Zitar & Laith Abualigah (2022). Multiclass feature selection with metaheuristic optimization algorithms: a review. *Neural Computing and Applications*, 2022, doi: 10.1007/s00521-022- 07705-4.

[47] Ali MAS, P. P. F, & Salama Abd Elminaam D. (2022). A Feature Selection Based on Improved Artificial Hummingbird Algorithm Using Random Opposition-Based Learning for Solving Waste Classification Problem. *Mathematics*. 2022; 10(15):2675. https://doi.org/10.3390/math10152675.

[48] Sathish K. R. & T. Ananthapadmanabha (2021). Improved black widow-bear smell search algorithm (IBWBSA) for optimal planning and operation of distributed generators in distribution system. *Journal of Engineering, Design and Technology*, 2021, doi: 10.1108/jedt-09-2020-0362.

[49] Rung -Ching Chen, Christine Dewi, Su-Wen. Huang & Rezzy Eko Caraka (2020). Selecting critical features for data classification based on machine learning methods. *Journal of Big Data*, vol. 7, no. 1, 2020, doi:10.1186/s40537-020-00327-4.

[50] Nico Surantha, Tri Fennia Lesmana & Sani Muhamad Isa (2021). Sleep stage classification using extreme learning machine and particle swarm optimization for healthcare big data. *Journal of Big Data*, vol. 8, no.1, 2021, doi: 10.1186/s40537-020-00406-6.

[51] Muhammad Waqar Saddique, Shaikh Saaqib Haroon, Salman Amin, Abdul Rauf Bhatti, Intisar Ali Sajjad & Rehan Liaqat (2020). Optimal Placement and Sizing of Shunt Capacitors in Radial Distribution System Using Polar Bear Optimization Algorithm. *Arabian Journal for Science and Engineering*, vol. 46, no. 2, pp. 873-899, 2020, doi: 10.1007/s13369-020-04747-5.

[52] Yi Han, Pengfei Pan, Hexin Lv & Guoyong Dai (2022). A Hybrid Optimization Algorithm for Water Volume Adjustment Problem in District Heating Systems. *International Journal of Computational Intelligence Systems*, vol. 15, no. 1, 2022, doi: 10.1007/s44196- 022-00091-8.

[53] Diaz, P.M., Jiju & M.J.E. (2022). A comparative analysis of meta-heuristic optimization algorithms for feature selection and feature weighting in neural networks. *Evol. Intel.* 15, 2631–2650 (2022), doi.org/10.1007/s12065-021-00634-6.

[54] Inam Abdullah Abdulmajeed & Idress Mohammed Husien (2022). MLIDS22- IDS Design By Applying Hybrid CNN-LSTM Model On Mixed-Datasets. Informatica, vol 46, no 8, 121-134, (2022),doi.org/10.31449/inf.v46i8.4348.