International Journal of
INTELLIGENT SYSTEMS AND APPLICATIONS IN ENGINEERING
ISSN:2147-6799 www.ijisae.org Original Research Paper

Performance Analysis of a Parameter Selection Model on a Big Data Set

Adusumalli Balaji[1]*, Ch. Indira Priyadarsini[2], Eluri Nageswara Rao[3], Kalluri Siva Krishna[4], Nangineni Srikanth[5], K Ravi kiran Yasaswi[6], Popuri Srinivasarao[7]

Submitted: 25/01/2024 **Revised**: 03/03/2024 **Accepted**: 11/03/2024

Abstract**:** Among the numerous phases within the data analysis process, the meticulous choice of parameters or attributes stands out as a pivotal stage. An erroneous choice in this regard can lead to suboptimal decisions. In the process of decision analysis, it proves advantageous for the decision-maker to have the capability to select and employ the most suitable model for identifying the optimally configured attribute set. In recent times, a substantial number of data scientists across various application domains have been attracted to the exploration of the advantages and disadvantages associated with big data. One prominent challenge arises when there is no appropriate model available to serve as a guiding framework, making the evaluation of extensive and diverse data in a big data environment particularly daunting for data scientists. Consequently, this study proposes an alternative parameterization approach capable of yielding an optimal attribute set while minimizing the associated learning, utilization, and maintenance costs. This model is constructed by integrating two complementary models with the soft set theory, best-first search algorithm, correlation-based feature selection, and rough set theory, all working synergistically as a parameter selection methodology. The proposed model has notably emerged as a strong contender in experiments focused on processing vast datasets.

Keywords: *soft set theory, best optimized attribute set, best- first search algorithm, correlation-based feature selection, rough set theory*

I. INTRODUCTION

In any industrial context, data analysis is a fundamental and pivotal task. This process encompasses a range of activities, such as data preprocessing, extraction, and selection, all aimed at facilitating optimal decision-making in addressing specific issues. The parameterization phase, guided by defined tools, mathematical formulations, or modeling methodologies [1-3], plays a crucial role in identifying the most suitable set of parameters. During this procedure, data undergoes a transformation from its raw and unstructured state to a cleaned, well-formatted, and optimized form. It is important to emphasize that an inefficient parameterization procedure can have a substantial impact on the process of making decisions, potentially resulting in suboptimal and incorrect outcomes. Several aspects can contribute to breakdown the parameterization process, with the quantity and characteristics of the data being prominent among them. Large datasets, particularly those marked by diverse criteria, imbalances, uncertainties, and inconsistent data values [4], pose significant challenges when it comes to analysis, particularly when inappropriate techniques and tools are employed. Across multiple application domains, including transport [7,8], healthcare [5], issues within the field of engineering [9], and finance [6], extensive research efforts have been undertaken to address these challenges. Wang et al. explored feature selection methods tailored to bioinformatics datasets [10], Pramanik et al. Explored the structuralframework and technological foundations in healthcare industry supporting the use of big data sets [11], and Shen et al. Suggested a combined method for evaluating the life insurance firms financial viability [12]. These studies underscored the profound impact of data complexity on decision-making processes among researchers.

In the decision making, researchers select the most appropriate tools for their specific needs. For example, they have the option to utilize rough sets to address non-linear problems and uncertainties [13,14]. They can also apply neural networks complex data structures analysis [11] and combine support vector machines (SVM) to manage high- dimensional data sets [15,16]. A variety of methodologies, models, and formulations have been put forward, each tailored to handle specific sets of challenges or problems. Some of these efforts introduced

[*1] *Computer Science and Engineering, Chalapathi institute of engineering and technology, Guntur, India.*
[2] *Deapartment of Mechanical Engineering, Chaitanya Bharati Institute of Technology, Hyderabad, India.*
[3] *Computer Science and Engineering(IOT), RVR & JC College of Engineering, Guntur, India.*
[4] *Computer Science and Engineering, Malineni Lakshmaiah Womens Engineering College, Guntur,India.*
[5] *Computer Science and Engineering, Brilliant Institute of Engineering and Technology,Hyderabad,India.*
[6] *Department of MBA, Lakkireddy Bali Reddy College of Engineering, Mylavaram, AP, India.*
[7] *Computer Science and Engineering(DS), RVR & JC College of Engineering, Guntur, India.*
India.Email:popurisrinivas333@gmail.com

innovative concepts and strategies in the process of decision making to enhance the use of software and hardware. Given the persistent and increasing complexity of data-related problems, these prior studies have significantly contributed to the highlighted field and are expected to continue doing so in the future.

Furthermore, there is a global trend among businesses to actively seek and propose solutions for addressing the challenges posed by big data. Many companies introduced a variety of technologies aimed at tackling issues of big data sets. In order to store, visualize and analyze vast volumes, numerous tools and methodologies have been developed. Oracle, for instance, provides a platform that facilitates the seamless integration, management, and analysis of big data. Google offers a range of services, all designed to assist users in working with and analyzing extensive datasets. Google has demonstrated that these services collectively enhance decision-makers' ability to make more informed choices. Additionally, several research studies explore the use of parallel processing techniques to effectively handle large volumes of data [17].

This research introduces a novel approach for selecting the most appropriate parameters within extensive datasets, utilizing a two-phase hybrid parameter selection model inspired by recent significant studies. In the first phase of parameter selection, the focus of this proposed model is on managing vast quantities of data, while in the second phase, it identifies and eliminates data characterized by uncertainty and inconsistency. Drawing from insights gleaned from prior experimental investigations [4,18], In the first phase of parameter selection, a hybrid approach combines best first search (BFS) and correlation feature-based selection (CFS) methods. In the subsequent stage, a fusion of soft set (SS) selection of parameter and rough set (RS) selection of parameter techniques are utilized to assess degree of uncertainty and consistency within the datasets. The main goal of this study is to offer the makers of decision an alternative, efficient, and cost-effective approach to facilitate the parameter selection process. This model operates efficiently without requiring a CPU performing high or extensive memory during the extraction, selection, or analysis of intricate datasets. Its intended use is in the preprocessing phase of data, where it generates an optimized dataset that can subsequently support the process of decision making. This proposed model offers valuable support for the data pre- processing task, making it suitable for implementation across various decision-making domains, including classification, clustering, and prediction.

The paper's structure is as given: In 1st Section, we provided a concise overview of the current problem associated with the proposed task. Section 2 incorporates relevant key studies pertaining to the proposed model. Section 3 delves into the methodology employed for implementing the proposed model. Section 4 substantiates the planned work by elucidating the data and experimental results. Finally, in 5th Section, we end the entire project and emphasize several key insights gleaned from the study.

II. ASSOCIATED WORKS

Subsequent sections delve into various areas of discussion relevant to the highlighted problem, such as big data, the choosing of parameters using correlation feature-based selection, as well as employing soft sets parameter selection and rough sets parameter choosing.

A. Big data

The advent of the big data era has indirectly influenced all components of information systems related to data, encompassing technology and processes. Big data is characterized as information abundant in quantity, speed, diversity, value, and accuracy, necessitating the application of suitable data processing techniques [19,20]. Big data velocity pertains to the speed at which data is processed, typically ranging from milliseconds to seconds during streaming, while big data volume encompasses data sizes spanning from terabytes to zettabytes. High variety signifies that the data exists in a wide array of formats, including text, numerical data, structured and unstructured data, multimedia, and more. Significant value is associated with large datasets, indicating that they encompass a wide array of information, spanning from easily accessible to highly valuable sources. Data veracity highlights the substantial levels of uncertainty and inconsistency inherent in extensive datasets. As per Information Management (IM), conventional relational databases lack the necessary capabilities to effectively manage extensive datasets. Big data originates from diverse sources, including online platforms, applications of transactions, logs, sensors, devices, video, and audio, and is generated continuously and at a substantial magnitude.

Big data has evolved into a remarkable and intricate challenge for professionals across various data-related domains, including database providers, data engineers, data analysts, and other affiliated communities [21–24]. The domain of big data encompasses four fundamental stages: generation of data, collection of data, storage, analysis [19]. The predominant focus of many endeavors has been on enhancing the efficiency and effectiveness of existing software, hardware, methods, or algorithms tailored for handling substantial datasets [25]. Prominent technologies associated with big data encompass the IoT, Hadoop, NoSQL, MapReduce and cloud computing.

Various architecturalstrategies, such as reasoning, extraction of information, and alignment of ontology, have been suggested to handle and deploy extensive datasets. Additionally, a range of models of big data like Cassandra, BigTable, MongoDB have been put forward for this purpose. Furthermore, a multitude of approaches, encompassing cloud computing, quantum computing was introduced to address these challenges. Numerous sectors, particularly finance, marketing, and retail, stand to gain significant benefits from harnessing big data. Big data offers a treasure trove of valuable insights that can assist these industries in innovating and developing new products and services. By employing appropriate analytical techniques and tools, companies can raise profits, enhance productivity, and optimize performance. According to [25], contemporary research primarily centers on the fields of big data processing and storage, with a particular emphasis on classification, clustering, and prediction strategies. However, there has been relatively limited research focused on advancing or introducing novel approaches in the field of large data pre- processing, leaving ample room for exploration and investigation within the realm of big data.

B. Correlation-based Feature Selection (CFS)

In 1999, Hall introduced a multivariate feature selection technique known as correlation-based feature selection (CFS), which falls within the category of filter-based feature selection methods. This method sifts through values of attribute by employing a heuristic function based on correlation. CFS evaluates attribute values by examining their correlation with the class as well as their correlation with other attributes. It then rates and selects attribute values accordingly, also removing those that lack significant links with either the class or the representative values [26,27]. CFS operates in two sequential phases, with the first phase involving the computation of correlation values among attributes and between classes. Meanwhile, in the second phase, it discerns the most pertinent attributes through exploration of the space of attribute, employing diverse techniques of heuristic search, including the best-first search method as described in reference [28]. Equation (1) [29] outlines the calculation used to determine the property in the dataset that exhibits the highest correlation.

$$cr_{zc} = \frac{f\overline{cr_{zi}}}{\sqrt{f + f(f-1)\overline{cr_{ii}}}} \tag{1}$$

In this context, let's define some key terms: cr_{zc} represents the value of heuristic assigned to an attribute of subset among total f attributes. cr_{zi} stands for average correlation between the class and the attributes, while cr_{ii} denotes average inter-correlation among pairs of attributes. In the process of reduction of data, the set of attributes with the maximum heuristic value is chosen as optimized attribute set, which is subsequently used in the subsequent analysis phase. One of the primary advantages of CFS lies in its computational efficiency, making it less complex in comparison to wrapper methods and other techniques. Nevertheless, in contrast to wrapper and embedded approaches, CFS does not demonstrate significant effectiveness in improving the performance of learning algorithm. Consequently, numerous researchers have endeavored to augment CFS's capabilities by combining it with other methods of selection of feature. CFS has been extensively applied across diverse domains, including tackling issues

associated with data of high dimension [30], applications of medical field [29], security concerns [26], and challenges in bio-computing [28]. Recent studies have illustrated how CFS has aided in elevating performance of decision making by optimizing already existing decision analytic techniques [30].

C. Soft set parameter selection

Another filtering technique employed for selecting the most appropriate attribute values within a dataset is known as soft set parameter selection. In this, probability is utilized to identify optimal attribute sets while eliminating attribute values characterized by ambiguity, uncertainty, or inconsistency [31]. Molodtsov introduced this concept in 1999 and has since undergone significant refinement by researchers to enhance its capacity to assist decision-makers in making informed judgments [32]. Certain scholars argue that, when it comes to identifying the best and less-than-ideal attribute values within the process of decision analysis, the SS parameter selection approach surpasses the RS parameter selection method. Additionally, some academics assert that the probability-based formulation of the SS parameter selection approach notably is more straightforward compared to that of the RS selection of parameter method. The SS selection of parameter approach has showcased its efficacy in numerous application domains, as evidenced by references [14, 33]. Within the domain of SS theory, a set of mapping from parameters to crisp subsets of universe is referred to as a soft set. The foundational concept underpinning Molodtsov's soft set theory is elucidated in the following definition [34]. For additional insights and examples related to soft set theory, please refer to references [34–36].

Definition 2.1: We start by establishing S as an initially defined collection of objects, ensuring that it is not empty. Following that, R is defined as a set of parameters which are in relation to the objects present in S. Let A is a subset of R, the power set of S is P(S). We introduce mapping, denoted as F: A → (S), where F associates elements from A with subsets of U. Consequently, the combination (F, A) is a SS over S. In essence, a SS over S can be described as a collection of parameterized subsets of S.

The SS approach demonstrated effectiveness when combined with various mathematic models and concepts. Some investigations were done to check the efficacy of the SS selection of parameter method. Nevertheless, when confronted with substantial volumes of data, this strategy proved inadequate in identifying the most and the sub optimal values of attribute. Soft set (SS) encountered challenges associated with high computational complexity and demanded substantial computer memory resources for the analysis process. Moreover, the selection of a SS parameter frequently resulted as the equal quantity of attribute values as initial characteristics within the chosen dataset [18,37,38].

D. Rough set parameter selection

Another method relying on mathematical principles is rough set (RS) parameter selection approach. This approach is based on the concept introduced by Pawlak in the year 1997 [39], advocates the use of probabilistic concepts to eliminate uncertain data. The primary objective of the RS approach is to mitigate issues associated with fuzziness, ambiguity, and inconsistency that can be inherent in various types of data [39]. Researchers have increasingly favored the application ofthe RS parameter selection approach, particularly when dealing with challenges posed by high-dimensional data [40,41]. For a more comprehensive understanding of the definition and development of rough set theory, interested individuals can refer to numerous research papers, including the one authored by Pawlak [39].

The effectiveness of the RS parameter selection approach was successfully demonstrated and is currently being applied across various sectors, including health science and finance, for tackling complex issues. It is employed in tasks such as classification, prediction, and optimization alongside other decision analytic methods [42,43]. Numerous scholars have undertaken efforts to enhance the RS incorporating and extending its actual concept as a novel framework, with a focus on improving its capabilities and advantages. Some researchers aim to amalgamate the RS framework with other theories, as seen in references [44,45], to further enhance its performance. In certain studies, the RS has been utilized as a complementary strategy to mitigate the limitations of alternative approaches. An illustrative instance of this is “Dominance-based Rough Set Approach (DRSA)” [46], introduced to maintain ordinal datasets, relationships with characteristics which are monotonic. This serves as an example of how scholars have expanded and enriched the RS theory through the introduction of novel formulations. The fundamental concept underlying rough set theory is as follows:

Definition 2.2: Let space of approximation be denoted as (S, σ), where S is a finite, non-empty universe set, and σ is a relation of equivalence on S, certain definitions and relationships hold. Specifically, if Y is considered as a subset of Universal set S, it might or not be expressible as a union of equivalence classes within S. When Y could be expressed as such, it is termed "definable"; otherwise, it is categorized as "indefinable." For instances where Y falls into the latter category (i.e., it is indefinable), can be made approximately into two distinct subsets, known as the lower and the upper nears of Y, as delineated in the subsequent discussion [47].

The function $\underline{b}(Y)$ is expressed as the combination of equivalence classes [y]σ, where each [y]σ is a subset of Y. Alternatively, b(Y) is represented as the combination of equivalence classes [y]σ, with the condition that their intersection with Y is not empty.

A rough set consists of two components, namely ($\underline{b}$(Y), b(Y)). The region of boundary is defined as the set $\underline{b}$(Y) minus b(Y). Consequently, if $\underline{b}$(Y) is equal to b(Y), then Y can be considered “definable”. Conversely, if $\underline{b}$(Y) minus b(Y) results in an empty, then Y itself is empty. In the context of a set Y, $\underline{b}$(Y) represents the largest set which is definable that is present within Y, and b(Y) represents the smallest set which is definable that encompasses Y.

III. METHODOLOGY

This offers a comprehensive elucidation of the theory and methodology utilized in the development of the hybrid parameter selection model. The initial four operations encompass preprocessing of data, deconstruction of data, selection of feature, and generation of result, with recommendation of data constituting the fifth step. The proposed model addresses two pivotal challenges: (i) the handling of data of high dimension and (ii) the management

of data marked by ambiguity and inconsistency. It is anticipated that this model holds the potential to mature into a comprehensive resource for facilitating informed decisions related to dataset selection for big data analysis and result generation. The full structure is represented in Figure 1.

The primary focus of the above described model centers on the phases of decomposition of data and selection of feature. The model initiates with preparatory section dedicated to data refinement. During this initial step, tasks such as data cleansing, formatting, randomization and normalization are performed. To make the data suitable for utilization in the subsequent procedures of data decomposition and feature selection, it must undergo cleaning and formatting procedures aligned with the specific requirements of relevant parameterization tools. The specific techniques employed for cleansing and formatting the selected data are contingent upon the composition and characteristics of that particular dataset. The data preparation phase encompasses the processes of data cleansing, formatting, standardization, and randomization, ultimately yielding a dataset that is both cleaned and formatted. This phase is often referred to as "Process 1."The phase two of this model is the decomposition stage of the data. In this, the objective is to partition the data into various groups or components. The need for data decomposition arises when the data size is too big to be analysed effectively using a single computational technique. So, this step is essential for determining the size characteristics and applying data reduction methods to each data set. There are two hypotheses under consideration: (1) that the data exceeds a threshold of 10,000, or (2) that it falls below 10,000, with this figure representing both the total number of instances and features. Previous research [4], inspired by the data decomposition technique which is speculative presented in [48], has examined and implemented this criterion. In the case of condition 1, the data undergoes a splitting process (S) if either the instance count or attribute count exceeds 10,000. The splitting technique is typically employed in parallel processing tasks to enhance processing speed and reduce execution times. The instance splitting procedure is initiated first, next is the process of splitting the attribute. The following is a formal definition of data decomposition:

Define A be the total groups and B be the total data points.

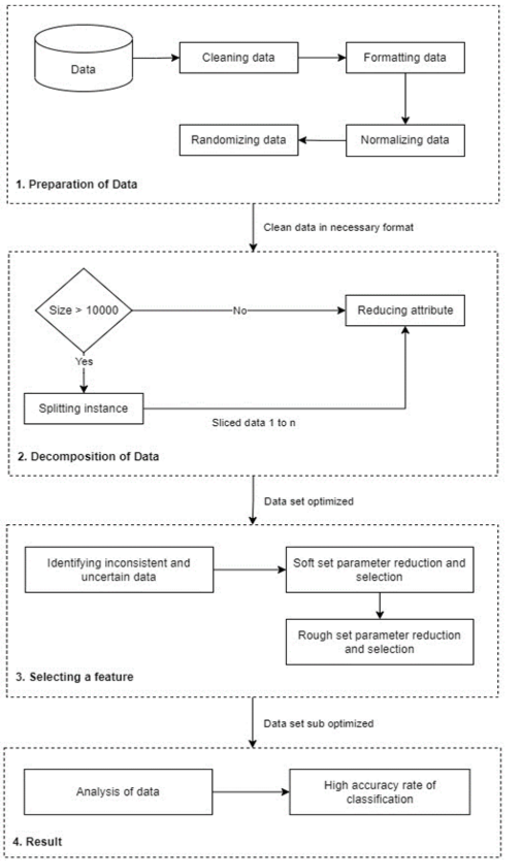

Fig. 1. Hybrid model architecture$A = (B/10,000)$ (2)

If A is not exactly divisible by 10000, then one will be added to the number of groups.

$A = A + 1$

(3)

As a result of the decomposition process, a set of datasets consisting of instances and attributes totalling < 10,000 will be named as SP(1) through SP(n). According to existing and prior research, it's common to find datasets used to evaluate the proposed model with more instances than attributes, typically having fewer than 10,000 attributes. Moreover, most parameterization algorithms and tools encounter difficulties when handling datasets larger than 10,000, particularly when dealing with less powerful computer systems for data analysis. To address this, the task divides the data into 10,000 instances using an optimistic and self-contained computational approach.

Each instance is considered independent of the others. If the data meets the second requirement, having both fewer than 10,000 instances and attributes, there is no need for the decomposition process. In such cases, this data should undergo the reduction of attribute step employing a reduction technique using hybrid.

The shown hybrid technique combines the best-first search (BFS) method for attribute search with the CFS for attribute evaluation. This hybrid approach identifies the most crucial attribute, designating it as the (OAS) most optimized attribute set. Subsequently, this undergoes attribute reduction utilizing CFS, BFS reduction approaches, as given in Figure 2. Before progressing to the next step, the outputs generated by the hybrid CFS process of reduction and BFS process of reduction for each group of SP are reviewed. This analysis aims to ascertain the total optimized attributes in every SP group and choose the SP group with the highest count of optimized characteristics. If multiple SP groups exhibit the same total highest optimized characteristics, therefore, first SP group is chosen. The procedure for identifying the SP group which is optimal or the most OAS is outlined in Algorithm 1. The data provided as input for this method is derived from the list of outputs, represented as SP1 through SPn, denoted as R1 through Rn. As previously mentioned, CFSBFS yields a

collection of attributes present in SP, culminating in the most OAS as the final outcome.

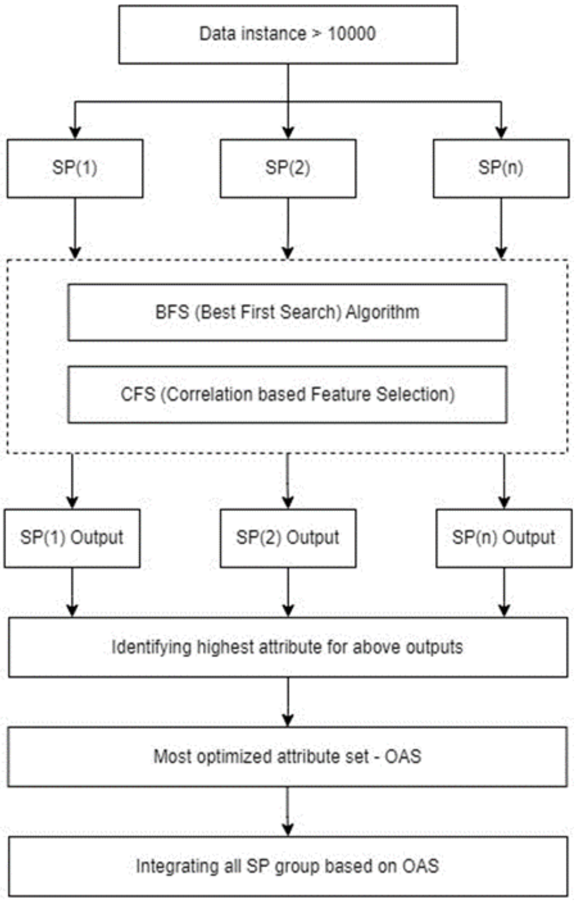
Data instance > 10000
SP(1)
SP(2)
SP(n)
BFS (Best First Search) Algorithm
CFS (Correlation based Feature Selection)
SP(1) Output
SP(2) Output
SP(n) Output
Identifying highest attribute for above outputs
Most optimized attribute set - OAS
Integrating all SP group based on OAS

Fig. 2. Phase of data decomposition

Algorithm 1: Algorithm to search Most Optimized Attribute Set (OAS)

Input: R1 to Rn are Optimized reduct sets. Output: Reduct set which is most optimal.

1) If the reduct set R contains multiple values, advance to step 2; otherwise, continue.

2) Choose the maximum attribute values, denoted as HR. If the previously attained HR possesses a distinct count of attribute values and has more than one value, proceed to step 3; otherwise, continue.

3) Choose the initial reduction set FR based on attribute values.

Above is the algorithm to search the most OAS.

Definition 3.1: Considering the sets R1, R2, ..., Rn within R, which is produced using CFSBFS process of attribute reduction and comprises a set of optimal reducts, we identify HR as the largest count of attribute values. HR is established when it exceeds Rn, with n representing the quantity of reducts in R. If HR is not equal to attribute values and value of HR is more than 1, it is then the value HR1.The result from 2nd phase will act as the income for 3rd phase, where the optimized dataset will undergo an additional parameterization step. In this stage, employing hybrid mathematical approaches such as the SS and RS reduction of parameter techniques, focus is on identifying values of uncertainty and consistency within the dataset. The objective is to create the most optimal dataset by eliminating these ambiguous values. The elimination process commences with a SS reduction of parameter and choosing phase, followed by a RS reduction of parameter and choosing phase. This method is referred to as the SSRS method, and is employed in Phase

3. Given the effectiveness of both methods in producing the optimal and sub-optimal datasets, a dual reduction and selection procedure is carried out.

Previous research and experimentation have shown that the SS reduction of parameter and choosing procedure cannot yield an optimal dataset. This approach tends to select every attribute present in the dataset as a result of the parameterization process, assuming the significance of each attribute for examination. This tendency raises concerns, particularly when dealing with extensive datasets, as it introduces a high degree of uncertainty. Consequently, a RS parameter selection approach is employed as a supplementary selection approach to address limitations of the SS selection process. The RS parameter selection approach also serve as validator to verify the accuracy of the outcomes produced by the SS method of selecting parameter. This selection of parameter technique reevaluates the dataset and determine uncertainty and consistency values, ultimately furnishing the most optimal dataset for use as input in the subsequent data analysis process. Algorithms 2 and 3 define the steps of 3rd phase, are shown in Figures 3 and 4.

The output of Phase 3 represents a refined dataset, devoid of uncertainties and inconsistencies. Moving on to Phase 4, which involves result creation, this refined output will serve as the input. Phase 4 entails a comprehensive examination of the data, employing techniques as categorization, regression and prediction. The efficiency of the described model of parameterization can be assessed by evaluating the accuracy of the outcomes, aiming for an accuracy rate of 100% or very close to it. These outcomes will demonstrate the model's performance, including its capability to handle extensive datasets, uncertain datasets, and inconsistent datasets. Phase 5, referred to as the data recommendation phase, signifies the culmination of all prior data analysis phases. It entails providing a summary of the dataset and suggesting its suitability for utilization in data analysis to the decision- maker. These recommendations are based on the accuracy and precision of the method's outcomes during the analytical phase. High-quality data is essential for decision-makers when selecting the optimal solution for any problem. The refined dataset inherently aids analytical techniques such as support vector machines in generating reliable decision-making results.

Definition 3.2: S is composed of the elements W, X, Y, and Z, S is formed by uniting all of its constituent elements. Each element is sequentially utilized, one after the other. Set S is suitable for various types of datasets and data analysis procedures, particularly those involving extensive datasets.

Example 1: Consider W to represent the initial phase of data preparation, X to signify the subsequent decomposition of data phase, Y to denote selection of feature, and Z to symbolize results production. When these sequential operations are executed to build a robust process to analyse data, the complete described model is identified as S.

Algorithm 2: Algorithm to perform soft set parameter reduction process

1) Represent the soft set in tabular form as (A, X). The soft set reduction set is denoted as (A, Y) for the soft set (A, X), where X is a subset of E, provided that Y is the reduction of X.

Input: Set X and a soft set (A, E). Output: Optimal solution.

2) Enter the selection parameters set X and identify all the reducts of (A, X).

3) Choose a single reduct set from (A, X) and regard it as (A, Y). Then, based on the predetermined weights, generate a weighted table for (A, Y).

4) Determine the value of k for which c_k is the maximum among all c_i values.

a) The optimal selection for the designated item is represented by hk. In case there are multiple values for k, any of the advantages can be selected.

b) ci represents the selected value of an object hi, where ci is calculated as the sum of hij values, with hij being the entries in the reduct soft set table.

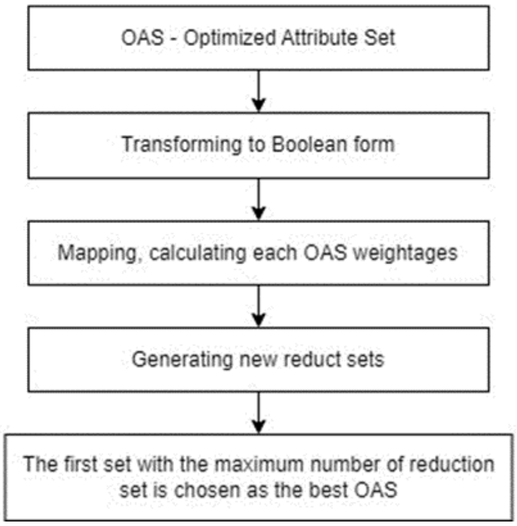
OAS - Optimized Attribute Set
Transforming to Boolean form
Mapping, calculating each OAS weightages
Generating new reduct sets
The first set with the maximum number of reduction set is chosen as the best OAS

Fig. 3. Process of soft set parameter selection

Algorithm 3: Algorithm to perform rough set parameter reduction process

Input: S denoted as (U, B, V, e), comprises a finite and non empty object set U, a finite and non empty attribute set B, a nonempty set of values V, and a function e that maps each object in U to a single value in V.

Output: Reduct sets in simplified format

1) Provide the information table S as input and perform data discretization.

2) Create a discernibility matrix of order n. The elements within the S table are characterized as e(p, q) = b ∈ B | e(p, b) ≠ e(q, b), where e(p, q) represents an attribute set differentiating between p and q. For every attribute b ∈ B, if e(x, y) = b1, b2,……., bk ≠∅.

3) Develop a discernibility function represented as ∑e(p, q) or the logical expression b1 ∨ b2... ∨ bk, as depicted by: E(B) = ∏(p, q)∈(U×U) ∑e(p, q).

4) If e(p, q) is empty, assign a value of 1 to the function of Boolean. Then, proceed with the process of reduction of attribute on basis of the Boolean function which is simplified, leading to the generation of a new and optimized reduct set.

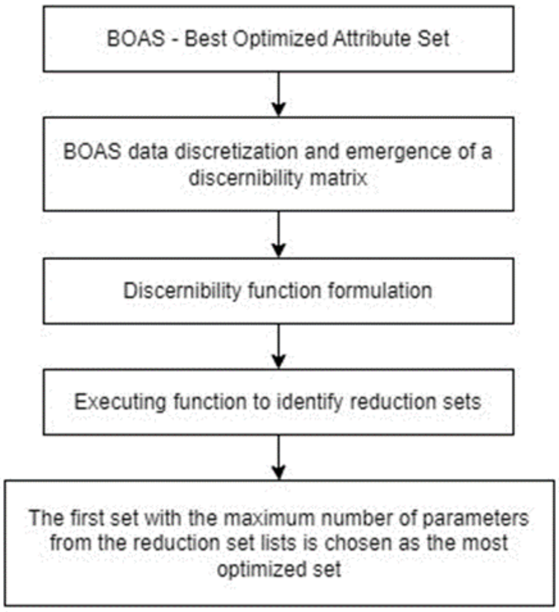
BOAS - Best Optimized Attribute Set
BOAS data discretization and emergence of a discernibility matrix
Discernibility function formulation
Executing function to identify reduction sets
The first set with the maximum number of parameters from the reduction set lists is chosen as the most optimized set

Fig. 4. Process of rough set parameter selection

IV. WORKS

The primary goal of this study is to calculate the effectiveness of the described hybrid model, which combines Phase 2 and Phase 3. The assessment will determine if the model can enhance the decision-making process by providing the best-optimized attribute collection. This model, referred to as M1, merges the

CFSBFS and SSRS parameterization techniques. The experimentation involved multiple tests conducted on various datasets within the categorization process. The experimental study utilized software tools such as WEKA, Matlab, and RSES. In the classification process, three popular classifiers, namely Neural Network Backpropagation (NNBP), Support Vector Machines (SVM), and Deep Learning (DL), known for their effectiveness in classifying a wide range of feature values, were employed. Additionally, two other hybrid techniques, CF-SGA (Correlation-based with Genetic Algorithm) and CFSGS (Correlation-based with Greedy Stepwise), referred to as M2

and M3, were implemented in Phase 2 to assess the hybrid model performance.

A. Desccription of data sets

The effectiveness of the proposed model was assessed using six meticulously selected datasets, which include Amazon-commerce-reviews (Amazon), Arcene, National Classification of Economic Activities (CNAE), Dota, Human Activity Recognition (HAR) and Poker. These datasets were taken from the Machine Learning Repository of UCI and Zenodo websites. The choice of these datasets aimed to evaluate the performance of Phases 2 (deconstruction of data phase), 3 (selection of feature phase) in finding the most relevant attributes for decision-making. Amazon, Arcene, and CNAE datasets were utilized to assess the selection of feature phase (Phase 3), while Dota, HAR and Poker datasets were employed to evaluate the decomposition of data process. Performance metrics such as accuracy rate, precision, F-measure, recall and Kappa statistic were employed to present the obtained results. A detailed description about the dataset characteristics is given in Table 1.

B. Standard models

The effectiveness of the described parameterization model, CFSBFS with SSRS, was assessed through a comparative analysis with two other benchmark models: CFSGA with SSRS and CFSGS with SSRS. CFSGA combines the CFS approach with a genetic algorithm, while CFSGS integrates the CFS method with a genetic search. This validation process aimed to identify the most frequently utilized model among the three models of parameterization developed. Furthermore, during the data analysis phase, three well-established classifiers were employed: the support vector machine (SVM), neural network backpropagation (NNBP), and deep learning (DL). Once again, neural network backpropagation demonstrated exceptional performance in data analysis tasks.

To assess the effectiveness of backpropagation of neural network, two other prominent classifiers are chosen. These three classifiers were compared to select the most suitable one for analysing the specified datasets during the classification process.

C. Benchmark on related works

The performance of the presented research has been verified through a comparative analysis with previous studies that utilized the same datasets. Table 7 illustrates the comparison between this research and well-known works. "Work 1 refers to the research by Wang et al., where they developed multivariate decision tree classifiers for large datasets partitioned randomly and using Principal Component Analysis (PCA)" [54]."Work 2 proposed by Garcia-Gil et al. combined Principal Component Analysis (PCA) and Random Discretization (RD) techniques for massive datasets" [55]. "Work 3 corresponds to the research conducted by Maillo et al., which enhanced the performance of k-Nearest Neighbors in large datasets through an iterative Spark-based architecture" [56].

In comparison to these three related endeavors, the presented research demonstrated commendable performance, as reflected in the results. The outcomes suggest that implementing the proposed approach is beneficial for analyzing extensive datasets. However, the final result was on par with the other outcomes, particularly in comparison to the high-performance distributed architecture computer system utilized in Work 3.

TABLE 1. DESCRIPTION OF DATA SETS

Data sets	**Number of instances**	**Number of attributes**	**Attribute characteristics**
Amazon	1500	10001	Real
Arcene	200	10001	Real
CANE	1080	857	Integer
Dota	92650	117	Integer

HAR	10229	562	Real
Poker	1025009	11	Integer, Real

V. RESULTS

The results from Phase 2 and 3 are evaluated in correlation with the achieved outcomes. Each stage is examined based on the quantity of selected attributes or parameters which are optimized. It is influenced by the quantity of attributes required for informed decision-making. An optimized attribute set in a dataset can significantly enhance the accuracy and meaningfulness of the decision analysis approach.

A. Parameterization results

The datasets underwent two distinct phases of parameterization. In the initial parameterization process of Phase 2, which utilized CFSBFS, CFSGA, or CFSGS, the objective was to decrease the total entries of attribute identifying relationships between individual attributes. In Phase 3, the second stage of parameterization was carried out to address attribute values that were ambiguous and inconsistent within dataset. Table 2 illustrates the decrease in the count of attributes between Phase 2 and Phase 3. 3rd phase yields the total quantity of the attribute set which is best optimized (BOAS) as the result of the whole process of parameterization. This BOAS plays a crucial role in the subsequent classification process by helping select the appropriate attribute set for decision analysis. Moreover, both parameterization methods contribute significantly to reducing processing time and memory usage, particularly when operating on non-high-performance machines.

TABLE 2. PARAMETERIZATION RESULTS

Data sets	**Attributes**	**Decomposed**	**M1**		**M2**		**M3**	
			CFSBF	***SSRS***	***CFSGA***	***SSRS***	***CFSGS***	***SSRS***
Amazon	10001	No	42	17	3643	10	42	10
Arcene	10001	No	74	6	4297	4	72	4
CNAE	857	No	27	27	307	97	27	5
Dota	117	Yes	21	21	55	55	21	22
Har	562	Yes	55	8	262	8	55	8
Poker	11	Yes	6	6	6	6	6	6

As indicated in Table 2, for all models (M1, M2, and M3), the count of BOAS decreases from larger to smaller dataset sizes, with all attributes undergoing significant reduction after the second parameterization process (SSRS). This observation highlights that many datasets contain values indicating ambiguity and consistency. The significance of the parameterization process can be further illustrated by examining its impact on the classification process.

TABLE 4. PRECISION, RECALL AND F-MEASURE FOR CFS-BFS

Data sets	**SVM**			**NNBP**			**DL**		
	P	***R***	***F-M***	***P***	***R***	***F-M***	***P***	***R***	***F-M***
Amazon	0.19	0.17	0.15	0.04	0.35	0.36	0	0.21	0
Arcene	0	0.46	0	0.58	0.58	0.58	0.57	0.55	0.55
CNAE	0.82	0.80	0.81	0.83	0.78	0.79	0.87	0.79	0.79
Dota	0.56	0.56	0.57	0.54	0.57	0.59	0.55	0.58	0.56
Har	0	0.85	0.85	0.76	0.73	0.71	0.83	0.82	0.81
Poker	0	0.60	0	0	0.56	0	0	0.49	0

B. *Classification results*

The classification performance of all hybrid models was evaluated by subjecting the most optimized attribute sets produced by these models to a classification test. The results, displayed in Table 3, indicate the accuracy rates achieved. Except for the Amazon dataset, all hybrid models achieved classification accuracies exceeding 50%. However, these models proved ineffective in helping classifiers to accurately separate the Amazon dataset, in which the accuracy ranged from a mere 9% to 32%. This poor performance might be attributed to the dataset itself, which potentially contains duplicates and has an incorrect data structure due to the extensive attribute set size.

Interestingly, both the proposed model and M3 (CFSGS) demonstrated strong performance when utilizing SVM and deep learning classifiers for the classification of the Har dataset. Both models achieved an accuracy rate of 84% with SVM and 82% with deep learning, while scoring 71% with NNBP. However, it's worth noting that the M3 model did not provide assistance in classifying the CNAE dataset, resulting in "NA" notation due to the inadequacy of the reduction set generated by M3 for the task of classification. Furthermore, the outputs describe that only M2 (CFSGA) performed well with NNBP and SVM classifiers, but deep learning classifier did not. These findings suggest that while the total quantity of optimal attribute sets is same across various models, the specific attributes chosen can impact the data analysis process.

In addition to comparing the described model with other well-known models, an experiment was conducted using all of the above considered datasets. The results revealed that most of the datasets posed challenges for the classifiers, particularly NNBP.

NNBP faced difficulties due to the presence of multiple network layers, which required substantial processing time and memory resources for the analysis process. In contrast to the other three models, SVM and DL classifiers were able to accurately classify the Har dataset without the need for any parameterization techniques. This highlights the importance of the parameterization process, which involves data breakdown and parameter selection, in reducing processing time and memory usage. This can be observed in the significant difference in processing times between the shown model and a model without any methods of parameterization. The presence of "NA" or "0" suggested that the classifier either struggled to process dataset or required an extensive amount of time to do so.

TABLE 5. CFS-GA F-MEASURE SCORE

Data sets	**M2**			**M3**		
	SVM	**NNBP**	**DL**	**SVM**	**NNBP**	**DL**
Amazon	0.10	0.07	0	0.14	0.29	0
Arcene	0.42	0.6	0.7	0.59	0.61	0.60
CNAE	0.9	0.69	0.8	0	0	0
Dota	0	0	0	0.58	0.58	0.57
Har	0.63	0.61	0.6	0.76	0.69	0.75
Poker	0	0	0	0	0	0

C. *Discussion*

The average performances of classification, both across all models and data sets, is depicted in Figure 5. The Har data set stands out with the accuracy rate at 71.8% and highest among the remaining data sets. Simultaneously, this model, when coupled with the DL (Deep Learning) classifier, surpassed its counterparts with a performance of 62.3%. The collective findings illustrate an uneven distribution of classification results across all data sets. Therefore, an alternative assessment metric is being employed to calculate the efficacy of this approach. In contrast to established benchmark models, in our evaluation, we ascertain that the model we have introduced possesses substantial utility as a parameterization model. Accordingly, we have examined potential challenges within the datasets by means of precision, F-measure and recall analyses. The F-measure score indicates a balanced relationship between recall and precision, with a score of 1 or higher signifying superior performance [49].

In Table 4, you can find the precision, F-measure and recall values for each and every classifier across all considered data sets as evaluated by the suggested model. When considering the F-measure scores, it's evident that data sets Amazon and Poker, exhibit subpar performance compared to the other data sets.

TABLE 3. CLASSIFICATION RESULTS

Data sets	**Without PM**			**M1**			**M2**			**M3**		
	SVM	***NNBP***	***DL***	***SVM***	***NNBP***	***DL***	***SVM***	***NNBP***	***DL***	***SVM***	***NNBP***	***DL***
Amazon	38	NA	56	15.9	26.7	29.9	9.9	7.3	9.2	15.8	28.1	31.4
Arcene	52	NA	77.8	56.8	67.8	66.1	57	69	71.9	66.5	63.8	64.2
CNAE	0	NA	NA	77.9	77	81.7	73.9	73	86.2	*NA*	*NA*	*NA*
Dota	73.8	71.7	NA	56	58.8	55.9	98.7	98.7	0.96	57.8	57.9	55.8
Har	94.8	NA	97.7	84.1	70.8	81.2	62.1	60	57.5	83.8	71.1	80.03
Poker	56.2	49.1	49.3	59.7	53.8	48.58	58.7	54.9	49.1	58.8	53.5	48.32

The fact is the suggested model achieved an F-measure score exceeding 0.5% across all classifiers for all data sets highlights its ability to aid in the classification of large data sets, except for the Poker and Amazon data sets. Furthermore, Table 4 illustrates that, excluding Poker and Amazon data sets, the proposed model was successful in identifying all related instances across all analyzed data sets, as indicated by the precision values. Notably, the suggested model, when employing the NNBP and Deep Learning combination, consistently achieved high precision values, accurately identifying the true data set. The F-measure scores for the other benchmarking models (M2 and M3) are presented in Table 5. As observed in both tables, neither of the benchmarking models could assist classifiers in categorizing the Amazon and Poker data sets. In the case of the Dota data set, M2 also faced challenges, while M3 struggled with the CNAE data set.

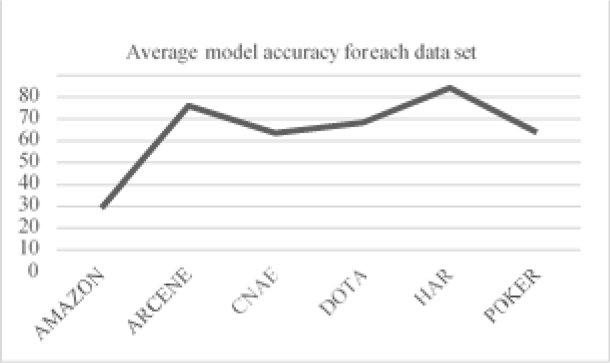

Fig. 5(a). Average model accuracy for each data set

D. Analysis of data sets used

As previously discussed, both the Amazon and Poker data sets yielded classification results that were less accurate for all parameterization models, with accuracy rates falling below 60%. Nevertheless, in contrast to some other research studies, such as [50], it's worth noting that the accuracy rate achieved for the Poker data set surpasses the typical classification accuracy level, which has led to the omission of results in many research studies [51]. Additionally, we assessed the correlation coefficient of the employed data sets using Kappa statistics, serving as another evaluation metric. The value's potential to reach 1 signifies the extent of the relationship between the attribute and the class [52,53].

Table 8 presents the Kappa statistic results for each model across all data sets. As mentioned earlier, it's worth noting that only the CANE and Har data sets demonstrate strong correlation between the attribute and the class. For M3 (CFS- GS), the parameterization phase failed to identify the most optimized characteristic, leading to an incorrect data interpretation during the classification process.

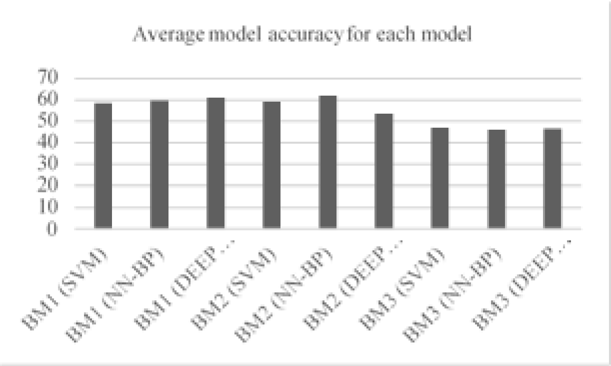

Fig. 5(b). Average model accuracy for each model

The conducted experiments yield three key insights. Firstly, it is crucial to consider the number of attributes and instances before commencing the data analysis process, as this can significantly impact processing time and memory usage. Secondly, it is essential to early identify the characteristics and values of the data set, especially when a weak association between class and attribute exists. Lastly, it is advisable to refrain from employing methods that may yield inaccurate or unsuitable results for the data intended for use in the decision- making process.

TABLE 7. COMPARISION OF VARIOUS WORKS

Data sets	**Proposed work**	**Work 1**	**Work 2**	**Work 3**
Poker	61.04	54.3	55.1	53.9

VI. CONCLUSION

The process of making the decision in big data entails a comprehensive workflow that encompasses data collection, processing, and generating optimal results. It necessitates significant investments in top-tier hardware, software, and skilled labor, incurring substantial costs. Researchers from diverse domains have conducted numerous studies to identify the most effective approaches, methods, and tools for managing big data. Some popular strategies for parameterization encompass machine learning algorithms and probabilistic theories. Big data processing can leverage a range of widely adopted technologies and methodologies, including Hadoop, Apache Cassandra, MongoDB, Apache Spark, Apache Storm, and R Programming.

TABLE 6. CFS-BFS KAPPA STATISTIC SCORE

Data sets	**M1**			**M2**			**M3**		
	SVM	**NNBP**	**DL**	**SVM**	**NNBP**	**DL**	**SVM**	**NNBP**	**DL**
Amazon	0.13	0.29	0.26	0.06	0.08	0.07	0.17	0.24	0.26
Arcene	0	0.35	0.29	0.02	0.39	0.45	0.26	0.23	0.23
CNAE	0.83	0.75	0.83	0.75	0.74	0.79	NA	NA	NA
Dota	0.11	0.12	0.09	0	0	0	0.19	0.19	0.15
Har	0.75	0.60	0.78	0.51	0.56	0.50	0.75	0.60	0.78
Poker	0.25	0.17	0.18	0.22	0.18	0.17	0.25	0.17	0.18

This research delves into the evaluation of a hybrid parameterization model, leveraging a variety of machine learning algorithms for the efficient management of extensive datasets, drawing inspiration from contemporary tools and technologies. The primary focus of this study revolves around the volume and diversity of the data. To select the most suitable machine learning technique for incorporation into the this model, many experimental procedures were done. As delineated in the above sections, the overall performance of the shown model surpasses that of established benchmark models. It has been illustrated that this proposed approach adeptly addresses the challenges posed by large datasets riddled with ambiguity and inconsistency. The model demonstrates its capability to partition vast datasets into multiple groups without undermining the integrity of the class-attribute relationship. Nonetheless, subpar outcomes were observed due to data imbalance and a lack of correlation among datasets. These suboptimal results can be attributed to two principal factors: attribute parameterization choices and the inherent characteristics of the dataset. To mitigate high error rates and low classification accuracy, future endeavours should involve the analysis of balanced datasets.

Acknowledgments

The successful culmination of any research endeavor relies on a harmonious interplay of invaluable components - sagacious counsel, unwavering guidance, and an enabling environment. These three elements serve as the bedrock of our research paper titled "Performance Analysis of a New Hybrid Parameter Selection Model on a Large Dataset".

We extend our heartfelt gratitude to Dr. Kolla Srinivas, Principal of R.V.R. & J.C. College of Engineering, Guntur, Andhra Pradesh, India, whose gracious consent allowed us to embark on this academic journey.

We extend our profound appreciation to Dr. M. V. P. Chandrasekhara Rao, Professor and Head of the Department of Computer Science and Business Systems (CSBS), as well as Computer Science and Engineering (Data Science). His relentless encouragement and unwavering support were instrumental in shaping our research and guiding us toward successful completion.

Lastly, our sincere appreciation goes out to the dedicated Teaching and Non-Teaching staff of the CSBS Department, whose collective efforts created an environment conducive to rigorous research and scholarly pursuits.

In each of these individuals and the nurturing environment of our institution, we found the pillars that supported our research endeavors, and for that, we are immensely grateful.

REFERENCES

[1] D. Kumar, R. Rengasamy, *Parameterization reduction using soft set theory for better decision making*, Pattern Recognition, Informatics and Mobile Engineering, 2013, pp. 3–5.

[2] N. Anitha, G. Keerthika, *A framework for medical image classification using soft set*, Curr. Trends Eng. Technol. (2014).

[3] M. Mohamad, A. Selamat, *Analysis on hybrid dominance-based rough set parameterization using private financial initiative unitary charges data*, LNAI Asian Conference on Intelligent Information and Database Systems, Springer, Cham, 2018, pp. 318–328.

[4] M. Mohamad, A. Selamat, *A two-tier hybrid parameterization framework for effective data classification*, New Trends in Intelligent Software Methodologies, Tools and Techniques, Vol. 303, IOS Press, 2018, pp. 321–331.

[5] Y. Liu, Y. Zhang, J. Ling, Z. Liu, *Secure and fine-grained access control on e-healthcare records in mobile cloud computing*, Future Gener. Comput. Syst. 78 (2018) 1020–1026.

[6] S.B.A. Kamaruddin, N.A.M. Ghani, N.M. Ramli, *Best forecasting models for private financial initiative unitary charges data of east coast and southern regions in peninsular Malaysia*, Int. J. Econ. Stat. 2 (2014) 119–127.

[7] A. Ahmad, M. Khan, A. Paul, S. Din, M.M. Rathore, G. Jeon, G.S. Choi, *Toward modeling and optimization of features selection in Big Data based social Internet of Things*, Future Gener. Comput. Syst. 82 (2017) 715–726.

[8] P. Sawicki, J. Żak, *The application of dominance-based rough sets theory for the evaluation of transportation systems*, Proc. Soc. Behav. Sci. 111 (2014) 1238–1248.

[9] M. Cecconello, S. Conroy, D. Marocco, F. Moro, B. Esposito, *Neural network implementation for ITER neutron emissivity profile recognition*, Fusion Eng. Des. 123 (2016) 637–640.

[10] L. Wang, Y. Wang, Q. Chang, *Feature selection methods for big data bioinformatics: A survey from the search perspective*, Methods 111 (2016) 21–31.

[11] M.I. Pramanik, R.Y. Lau, H. Demirkan, M.A.K. Azad, *Smart health: Big data enabled health paradigm within smart cities*, Expert Syst. Appl. 87 (2017) 370–383.

[12] K.Y. Shen, S.K. Hu, G.H. Tzeng, *Financial modeling and improvement planning for the life insurance industry by using a rough knowledge based hybrid MCDM model*, Inform. Sci. 375 (2017) 296–313.

[13] M. Esposito, A. Minutolo, R. Megna, M. Forastiere, M. Magliulo, G. De Pietro, *A smart mobile, self-configuring, context-aware architecture for personal health monitoring*, Eng. Appl. Artif. Intell. 67 (2018) 136–156.

[14] X. Ma, Q. Liu, J. Zhan, *A survey of decision making methods based on certain hybrid soft set models*, Artif. Intell. Rev. 47 (2017) 507–530.

[15] N. Allias, M.N. Megat, N. Megat, M.N. Ismail, *A hybrid gini PSO- SVM feature selection based on Taguchi method : An evaluation on email filtering*, Proceedings of the 8th International Conference on Ubiquitous Information Management and Communication, ACM, 2014, pp. 55–59, http://dx.doi.org/10.1145/2557977.2557999.

[16] Z. Masetic, A. Subasi, *Congestive heart failure detection using random forest classifier*, Comput. Methods Programs Biomed. 130 (2016) 54–64.

[17] B. Ait Hammou, A. Ait Lahcen, S. Mouline, *APRA: An approximate parallel recommendation algorithm for Big Data*, Knowl.-Based Syst. 157 (2018) 10–19.

[18] M. Mohamad, A. Selamat, *A new soft rough set parameter reduction method for an effective decision-making*, New Trends in Intelligent Software Methodologies, Tools and Techniques, Vol. 297, IOS Press, 2017, pp. 691–704.

[19] A. Hassani, S.A. Gahnouchi, *A framework for business process data management based on big data approach*, Procedia Comput. Sci. (2017).

[20] Y.-C. Ko, H. Fujita, *An evidential analytics for buried information in big data samples: Case study of semiconductor manufacturing*, Inform. Sci. 486 (2019) 190–203, http://dx.doi.org/10.1016/j.ins.2019.01.079, http://www.sciencedirect.com/science/article/pii/S00200255193005X.

[21] J. Luo, H. Fujita, Y. Yao, K. Qin, *On modeling similarity and three- way decision under incomplete information in rough set theory*, Knowledge-Based Syst. (2019) 105251, http://dx.doi.org/10.1016/j.knosys.2019.105251, http://www.sciencedirect.com/science/article/pii/S0950705119305635

[22] H. Fujita, A. Gaeta, V. Loia, F. Orciuoli, *Hypotheses analysis and assessment in counter-terrorism activities: a method based on OWA and fuzzy prob- abilistic rough sets*, IEEE Trans. Fuzzy Syst. (2019) 1, http://dx.doi.org/10. 1109/TFUZZ.2019.2955047. H. Fujita, A. Gaeta,

[23] V. Loia, F. Orciuoli, Improving awareness in early stages of security analysis: A zone partition method based on GrC, Appl. Intell. 49 (2018) 1063–1077.

[24] H. Fujita, A. Gaeta, V. Loia, F. Orciuoli, *Resilience analysis of critical infrastructures: A cognitive approach based on granular computing*, IEEE Trans. Cybern. 49 (5) (2019) 1835–1848, http://dx.doi.org/10.1109/TCYB. 2018.2815178.

[25] J. Akoka, I. Comyn-Wattiau, N. Laoufi, *Research on big data – A systematic mapping study*, Comput. Stand. Interfaces 54 (2017) 105–115.

[26] L. Koc, T.a. Mazzuchi, S. Sarkani, *A network intrusion detection system based on a Hidden Naïve Bayes multiclass classifier*, Expert Syst. Appl. 39 (18) (2012) 13492–13500.

[27] S. Chebrolua, S.G. Sanjeevi, *Attribute reduction in decision-theoretic rough set model using particle swarm optimization with the threshold*

param- eters determined using LMS training rule, Knowl.-Based Syst. 57 (2015) 527–536.

[28] O.S. Soliman, A. Rassem, *Correlation based feature selection using quantum bio inspired estimation of distribution algorithm*, Lecture Notes in Com- puter Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), LNAI, vol. 7694, 2012, pp. 318–329.

[29] N.F. Abubacker, A. Azman, S. Doraisamy, *Correlation-based feature selec- tion for association rule mining in semantic annotation of mammographic*, Pattern Recognit. Lett. 32 (2011) 482–493.

[30] S. Chormunge, S. Jena, *Correlation based feature selection with clustering for high dimensional data*, J. Electr. Syst. Inf. Technol. (2018) 4–11.

[31] D. Molodtsov, *Soft set theory-first results*, Comput. Math. Appl. 37 (4) (1999) 19–31.

[32] J. Chai, E.W.T. Ngai, J.N.K. Liu, *Dynamic tolerant skyline operation for decision making*, Expert Syst. Appl. 41 (15) (2014) 6890–6903.

[33] Y. Liu, K. Qin, L. Martínez, *Improving decision making approaches based on fuzzy soft sets and rough soft sets*, Appl. Soft Comput. J. 65 (2018) 320–332.

[34] X. Ma, N. Sulaiman, H. Qin, T. Herawan, J.M. Zain, *A new efficient normal parameter reduction algorithm of soft sets*, Comput. Math. Appl. 62 (2) (2011) 588–598.

[35] F. Feng, X. Liu, V. Leoreanu-Fotea, Y.B. Jun, *Soft sets and soft rough sets*, Inform. Sci. 181 (6) (2011) 1125–1137.

[36] M. Irfan Ali, *A note on soft sets, rough soft sets and fuzzy soft sets*, Appl. Soft Comput. J. 11 (4) (2011) 3329–3332.

[37] M. Mohamad, A. Selamat, *Recent study on the application of hybrid rough set and soft set*

theories in decision analysis process, Lecture Notes in Artificial Intelligent, LNAI, 9799, 2016, pp. 713–724.

[38] M. Mohamad, A. Selamat, *A new hybrid rough set and soft set parameter reduction method for spam e-mail classification task*, Lecture Notes in Artificial Intelligent, LNAI, 9806, 2016, pp. 18–30.

[39] Z. Pawlak, *Rough set approach to knowledge-based decision support*, European J. Oper. Res. 99 (1997) 48–57.

[40] Local rough set: *A solution to rough data analysis in big data*, Internat.

[41] J. Approx. Reason. 97 (2018) 38–63, http://www.sciencedirect.com/science/article/pii/S0888613X1730486.

[42] A. Oussous, F.Z. Benjelloun, A. Ait Lahcen, S. Belfkih, *Big data technologies: A survey*, J. King Saud Univ. Comput. Inf. Sci. 30 (2018) 431–448.

[43] T.K. Sheeja, A.S. Kuriakose, *A novel feature selection method using fuzzy rough sets*, Comput. Ind. 97 (2018) 111–121.

[44] J. Liu, Y. Lin, Y. Li, W. Weng, S. Wu, *Online multi-label streaming feature selection based on neighborhood rough set*, Comput. Ind. 84 (2018) 273–287.

[45] B. Huang, Y.L. Zhuang, H.X. Li, D.K. Wei, *A dominance intuitionistic fuzzy- rough set approach and its applications*, Appl. Math. Model. 37 (12–13) (2013) 7128–7141.

[46] W.S. Du, B.Q. Hu, *Dominance-based rough set approach to incomplete ordered information systems*, Inform. Sci. 346–347 (2016) 106–129.

[47] S. Greco, B. Matarazzo, R. Slowi, *Algebra and topology for dominance-based rough set approach*, Z.W. Ras, L.-S. Tsay (Eds.), Advances in Intelligent Information Systems, Springer, 2010, pp. 43– 78.

[48] M.I. Ali, B. Davvaz, M. Shabir, *Some properties of generalized rough sets*, Inform. Sci. 224 (2013) 170–179.

[49] A. Grama, A. Gupta, G. Karypis, V. Kumar, *Principles of parallel algorithm design*, Introduction to Parallel Computing, second ed., Addison Wesley, Harlow, 2003.

[50] H. Li, D. Li, Y. Zhai, S. Wang, J. Zhang, *A novel attribute reduction approach for multi-label data based on rough set theory*, Inform. Sci. 367–368 (2016) 827–847.

[51] I. Triguero, D. Peralta, J. Bacardit, S. García, F. Herrera, *MRPR: A MapReduce solution for prototype reduction in big data classification*, Neurocomputing 150 (2015) 331–345.

[52] A. Arnaiz-Gonzalez, J.F. Diez-Pastor, J.J. Rodriguez, C. Garcia- Osorio, *In- stance selection of linear complexity for big data*, Knowl.- Based Syst. 107 (2016) 83–95.

[53] S.K. Pal, S.K. Meher, S. Dutta, *Class-dependent rough-fuzzy granular space, dispersion index and classification*, Pattern Recognit. 45 (7) (2012) 2690–2707.

[54] G.R. Teixeira de Lima, S. Stephany, *A new classification approach for detecting severe weather patterns*, Comput. Geosci. 57 (2013) 158–

165.

[55] F. Wang, Q. Wang, F. Nie, W. Yu, R. Wang, *Efficient tree classifiers for large scale datasets*, Neurocomputing 284 (2018) 70–79.

[56] D. García-Gil, S. Ramírez-Gallego, S. García, F. Herrera, *Principal compo- nents analysis random discretization ensemble for big data*, Knowl.-Based Syst. 150 (2018) 166–174.

[57] J. Maillo, R. Sergio, I. Triguero, F. Herrera, *kNN-IS: An Iterative Spark-based design of the k-Nearest Neighbors classifier for big data*, Knowl.-Based Syst. 117 (2017) 3–15.

[58] Mohamad, Masurah, et al. "*An analysis on new hybrid parameter selection model performance over big data set*" Knowledge-Based Systems 192 (2020): 105441.

[59] Srinivasarao, Popuri, and Aravapalli Rama Satish. "*A Novel Hybrid Optimization Algorithm for Materialized View Selection from Data Warehouse Environments*" Computer Systems Science & Engineering 47.2 (2023).

[60] Srinivasarao, Popuri, and Aravapalli Rama Satish. "*A Hybrid Metaheuristic Framework for Materialized View Selection in Data Warehouse Environments*" International Journal of Cooperative Information Systems (2023): 2350021.