MULTISTAGE FEATURE SELECTION METHODS FOR DATA CLASSIFICATION

MASURAH BINTI MOHAMAD

UNIVERSITI TEKNOLOGI MALAYSIA

MULTISTAGE FEATURE SELECTION METHODS FOR DATA CLASSIFICATION

MASURAH BINTI MOHAMAD

A thesis submitted in fulfilment of the
requirements for the award of the degree of
Doctor of Philosophy

School of Computing
Faculty of Engineering
Universiti Teknologi Malaysia

AUGUST 2021

ACKNOWLEDGEMENT

In preparing this thesis, I was in contact with many people, researchers, academician, and practitioners. They have contributed towards my understanding and thought. In particular, I wish to express my sincere appreciation to my thesis supervisor Prof. Ts. Dr. Ali Selamat. Thank you for everything, your guidance, advice and motivation will always be remembered. To my examiners, Prof. Dr. Shahrul Azman Mohd Noah and Prof. Dr. Naomie Salim, thank you so much for evaluating my thesis and work. I was really encouraged by the constructive feedback to improve my work and make me a good researcher in the future. To Kementerian Pendidikan Malaysia (KPM), UiTM and especially UiTM Perak, thank you for the sponsorship and the given opportunity in pursuing my PhD study.

ABSTRACT

In data analysis process, a good decision can be made with the assistance of several sub-processes and methods. The most common processes are feature selection and classification processes. Various methods and processes have been proposed to solve many issues such as low classification accuracy, and long processing time faced by the decision-makers. The analysis process becomes more complicated especially when dealing with complex datasets that consist of large and problematic datasets. One of the solutions that can be used is by employing an effective feature selection method to reduce the data processing time, decrease the used memory space, and increase the accuracy of decisions. However, not all the existing methods are capable of dealing with these issues. The aim of this research was to assist the classifier in giving a better performance when dealing with problematic datasets by generating optimised attribute set. The proposed method comprised two stages of feature selection processes, that employed correlation-based feature selection method using a best first search algorithm (CFS-BFS) and as well as a soft set and rough set parameter selection method (SSRS). CFS-BFS is used to eliminate uncorrelated attributes in a dataset meanwhile SSRS was utilized to manage any problematic values such as uncertainty in a dataset. Several bench-marking feature selection methods such as classifier subset evaluation (CSE) and principle component analysis (PCA) and different classifiers such as support vector machine (SVM) and neural network (NN) were used to validate the obtained results. ANOVA and T-test were also conducted to verify the obtained results. The obtained averages for two experimental works have proven that the proposed method equally matched the performance of other benchmarking methods in terms of assisting the classifier in achieving high classification performance for complex datasets. The obtained average for another experimental work has shown that the proposed work has outperformed the other benchmarking methods. In conclusion, the proposed method is significant to be used as an alternative feature selection method and able to assist the classifiers in achieving better accuracy in the classification process especially when dealing with problematic datasets.

ABSTRAK

Dalam proses analisis data, keputusan yang baik dapat dibuat dengan bantuan beberapa sub proses dan kaedah. Proses yang paling biasa adalah pemilihan ciri dan proses pengelasan. Pelbagai kaedah dan proses telah dibangunkan untuk menyelesaikan banyak masalah seperti ketepatan pengelasan yang rendah, dan masa pemprosesan yang lama yang dihadapi oleh pembuat keputusan. Proses analisis menjadi lebih rumit terutama ketika berurusan dengan set data kompleks yang terdiri daripada set data yang besar dan bermasalah. Salah satu penyelesaian yang dapat digunakan adalah dengan menggunakan kaedah pemilihan ciri yang efektif untuk mengurangkan waktu pemprosesan data, mengurangkan ruang memori yang digunakan, dan meningkatkan ketepatan keputusan. Namun, tidak semua kaedah yang sedia ada mampu untuk menangani masalah ini. Tujuan penyelidikan ini adalah untuk membantu pengkelas dalam memberikan prestasi yang lebih baik ketika memproses set data yang bermasalah dengan penghasilan set atribut yang dioptimumkan. Kaedah yang dicadangkan terdiri daripada dua tahap proses pemilihan ciri yang menggunakan kaedah pemilihan ciri berdasarkan kolerasi bersama algoritma carian pertama terbaik (CFS-BFS) dan kaedah pemilihan ciri set lembut dan set kasar (SSRS). CFS-BFS digunakan untuk menghapuskan atribut yang tidak berkorelasi dalam set data sementara SSRS digunakan untuk mengendalikan setiap nilai yang bermasalah seperti ketidakpastian dalam set data. Beberapa kaedah pemilihan ciri penanda aras seperti penilaian subset pengkelas (CSE) dan analisis komponen utama (PCA) dan beberapa pengkelas berbeza seperti mesin vektor sokongan (SVM) dan rangkaian neural (NN) digunakan untuk mengesahkan hasil kajian yang diperoleh. ANOVA dan Ujian-T juga dijalankan untuk mengesahkan hasil yang diperoleh. Purata yang diperoleh daripada dua eksperimen menunjukkan bahawa kaedah yang dicadangkan mempunyai prestasi yang setara dengan kaedah penanda aras yang lain dari segi membantu pengkelas dalam mencapai prestasi pengelasan tinggi untuk data set yang kompleks. Purata yang diperoleh untuk eksperimen lain telah menunjukkan bahawa cadangan kerja adalah lebih baik berbanding kaedah penanda aras yang lain. Kesimpulannya, kaedah yang dicadangkan adalah signifikan untuk digunakan sebagai kaedah alternatif pemilihan ciri dan dapat membantu pengkelas dalam mencapai ketepatan yang lebih baik dalam proses pengelasan terutama ketika mengendalikan set data yang bermasalah.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF ABBREVIATIONS

ABC	-	Artificial Bee Colony
ACC	-	Accuracy
AI	-	Attribute identifier
ANOVA	-	Analysis of variance
AUC	-	Area Under the ROC curve
BFS	-	Best First Search
BM	-	Benchmark method
BOAS	-	Best optimised attribute set
CFS	-	Correlation-based Feature Selection
CFS-BFS	-	Correlation-based Feature Selection Best First Search
CFS-GA	-	Correlation-based Feature Selection Genetic Algorithm
CFS-GS	-	Correlation-based Feature Selection Greedy Stepwise
CSGS	-	classifier subset evaluation genetic search
CSES	-	classifier subset evaluation evolutionary search method
DL	-	Deep learning
DRSA	-	Dominance-based rough set approach
DT	-	Decision Tree
F-M	-	F-measure
FIFO	-	first in first out
FN	-	False negative
FP	-	False positive
KEEL	-	Knowledge Extraction based on Evolutionary Learning
M	-	method
MAE	-	Mean absolute error
NN	-	Neural Network
NNB	-	Neural network back propagation

NP	-	Non-deterministic polynomial time
NPV	-	Negative predictive value
OAS	-	Optimised attribute set
P	-	Precision
PCA-Ranker	-	Principle component approach with ranker
PPV	-	Positive predictive value
PR	-	Parameter reduction
R	-	Recall
RBFS	-	RecursiveBest-First Search algorithm
RF	-	Random Forest
ROC	-	receiver operating characteristic curve
RQ	-	Research question
RST	-	Rough set theory
SENS	-	Sensitivity
SPEC	-	Specificity
SSRS	-	Soft set rough set
SST	-	Soft set theory
SVM	-	Support Vector Machine
TN	-	True negative
TP	-	True positive
UTM	-	Universiti Teknologi Malaysia
WFS	-	Without feature selection
WRAP-RS	-	Wrapper subset evaluation and random search method

LIST OF SYMBOLS

$*$	-	Incomplete information
cr_{zc}	-	Heuristic value
cr_{zi}	-	Average value
f	-	Number of attribute
n	-	Node
U	-	Non-empty finite set
E	-	Set of parameters
F	-	Function
A	-	Subset
$P(U)$	-	Power set of U
R	-	Binary relation on U
$\underline{app}(X)$	-	Lower approximation
$\overline{app}(X)$	-	Upper approximation
$\sum$	-	Sum
σ	-	Sigma
$\subset$	-	Subset
G	-	Number of groups
D	-	Number of data
SP	-	Splitting process
$R_1 - R_n$	-	Optimal reduction set
$>$	-	Larger than
$\prod$	-	Product
$\neq$	-	Not equal
$\varnothing$	-	Var nothing
$\cup$	-	Union
S	-	Analysis process

LIST OF APPENDICES

CHAPTER 1

INTRODUCTION

1.1 Background Problem

Data analysis is the most crucial task in any application field. This process involves several tasks such as data pre-processing, feature extraction and feature selection that will assist the decision-maker in getting the best solution for a specified problem. An ineffective data analysis will affect the decision-making process and return wrong or inefficient solutions. There are several factors that might cause the data analysis process to become ineffective. The size and characteristics of the data are two main factors that will downgrade the efficiency of the data analysis process. A complex dataset might be comprised of large-sized data, which means that it has multiple types of criteria and also imbalanced, uncertain and inconsistent data values (Oussous et al., 2018). Complex datasets are difficult to analyse, especially when unsuitable methods and instruments are used. Some data analysis methods are unable to manage or analyse a large volume of data at one time. In addition, this sort of difficulty is also faced by certain hardware and software. Two key problems that are typically related to data analysis process are long processing times and high storage availability (Ait Hammou et al., 2018). Pre-processing data and selection of features are among the essential processes used to resolve these issues. Ineffective decisions might be generated if these two processes are wrongly conducted (Houari et al., 2016). In addition, over-fitting during the feature reduction process might also pose a challenge to the decision analyst (Qian et al., 2018).

Many application fields such as healthcare (Mursalin et al., 2017), finance (Dong et al., 2018), transportation (Ahmad et al., 2017), engineering (Kushal et al., 2020), bio-informatics (Wang et al., 2016) and security (Vijayanand et al., 2018) have conducted various research works that related to data analysis process. These studies have shown that, the data analysis process especially on feature selection do influence the decision-

making process. Feature selection also might help the decision analysis method such as the classifier to increase the learning accuracy, minimize processing time and eliminate unrelated and redundant data (Cai et al., 2018). In addition, problematic data, such as ambiguity, inconsistency, imbalanced and missing values, can be managed using feature selection methods (Zhou et al., 2017; Sheeja and Kuriakose, 2018; Hosseini and Moattar, 2019). According to the literature reviews that had been conducted, various feature selection methods that have been proposed either by proposing new method or enhancing the existing methods or integrating several single methods. For example, work done by Gao et al. named as Minimal Redundancy-Maximal New Classification Information (MR-MNCI) that integrated two feature selection criteria (class-dependent feature redundancy and class-independent feature redundancy) (Gao et al., 2018). Another example of proposing new feature selection method is done by Anaraki and Usefi in his work (Anaraki and Usefi, 2019) that identifies correlation between features based on perturbation theory.

From a literature study it can be concluded that different methods are needed to perform different data analysis problems. For examples, Fuzzy and rough sets can be used to handle uncertainty and nonlinear data problems (Esposito et al., 2018), while neural networks are suitable for use in analysing complex data (Choudhury and Pal, 2019) and support vector machines (SVMs) might be implemented to deal with high-dimension data when they are incorporated with other methods (Tao et al., 2019). Recently, different approaches, methods, frameworks or formulations were proposed, each of which took into consideration different kinds of problems or issues that need to be solved. Some of the works highlighted the performance of the proposed methods or models, some initiated new definitions, some considered the whole architecture of the proposed approach, and some investigated the capability of the hardware and software used in the decision-making process. All of these works have contributed to the focused area, and there will be no end to these works because data issues will always be emerging and becoming more complex.

1.2 Research Background

The feature selection process is a crucial process in data analysis. It is used to prepare the data into an optimised dataset that is ready to be analysed using any data analysis method. The feature selection process not only helps the data analyst to reduce or select the most relevant features but also helps to decrease the use of memory space, reduce the processing time and improve accuracy (Luan and Dong, 2018). The decision-making process becomes more effective when decision analysts make use of the feature selection process, especially when dealing with complex datasets. Researchers from many different areas such as science, engineering, medical, social science and economics have initiated either single feature selection methods, hybrid feature selection methods, or generalized or specific methods that are applicable for solving any decision-making problem.

However some of the existing approaches are incompatible to act as an efficient tool for selecting features. For example, a probability approach and fuzzy set approach cannot solve the problem of multidimensional characterisation properties (Hassan and Al-Qudah, 2019). The classical rough set theory and fuzzy set are a few approaches that have been mentioned as ineffective approaches for conducting the feature selection process, and they generate a low accuracy rate for the computation results (Singhal et al., 2018). Other traditional methods, such as support vector machines (SVM) and decision trees (DT), also have difficulties in handling certain types of datasets, in particular multivariate large datasets, where high computational costs such as time and space are required. SVM is suffer from instability problem where it requires exact factor or features in order to achieve high accuracy meanwhile DT require high volume of input data at the beginning of analysis process (Ghaddar and Naoum-Sawaya, 2018; Rao et al., 2019).

Inevitable presence of problematic values in the dataset is one of the data analysis issues. Some of the values are uncertainty, inconsistency, and imbalanced. The problem of uncertainty refers to the vague criteria that are to be evaluated during the decision-making process (Fahmideh and Beydoun, 2018). Uncertainty means incomplete information or it can also be defined as an attribute whose value is unknown.

According to (Durbach and Stewart, 2012), uncertainty can be divided into two types, namely, i) external and ii) internal uncertainty. Uncertainty which appears when the value of an attribute is derived from post events is categorized as external uncertainty. This type of uncertainty is caused by environmental situations and other related decision areas without the involvement of the decision-maker. Uncertainty that is caused by the decision-maker is categorized as internal uncertainty such as incomplete definitions of human preferences, incomplete human judgments and incomplete information. This situation will affect the decision-maker when deciding on a strategic resolution for problems that need to be solved. Imbalanced datasets also could affect the erroneous and could mislead the decision analysis results. The datasets are considered to be imbalanced when the ratio of the number of instances from the majority class to the number of instances of the minority class is higher or equal to 2 (Hosseini and Moattar, 2019). Imbalanced data might occur in real world problems such as medical diagnosis, banking fraud detection and bioinformatics (Liu and Zio, 2019). Inconsistency denotes a situation where a variable has more than one conflicting values whilst vagueness is defined as a property of sets or concepts which can be characterized into the limits of the set (Bello and Verdegay, 2012). Inconsistent data problems may occur when there is a situation or event which has been misinterpreted by the decision-maker. Invariably, incorrect information will create an inaccurate and inappropriate solution, especially when dealing with high dimensional data.

Moreover, single and classical decision-making algorithms and tools are unable to assist the decision-maker in solving the uncertainty problems successfully. The uncertainty issue will make the problem more complicated, especially when it involves a large dataset. Thus, the accuracy, completeness and cleanliness of the data are questionable (Maugis, 2018). Recently, these kinds of datasets have been generated tremendously from different applications, especially from social media applications such as Facebook, Twitter, Instagram and online shopping applications. These problems can degrade the performance of decision analysis methods such increasing the computational time and memory space and reducing the accuracy, especially when single or classical methods are being used (Qian et al., 2018). However, these data complexities cannot be analysed manually without the use of an efficient method as these data need to be pre-processed in order to reduce the difficulties (Chormunge and Jena, 2018; Luan and Dong, 2018). The complex values need to be eliminated, reduced

or properly analyse so that, the best solution can be made during the decision-making process.

Many decision analysis methods especially on feature selection have been initiated to deal with complex datasets. All such methods, whether single or hybrid, are aimed at increasing the decision analysis performance or at generating an optimal solution. Probability and fuzzy set theories are among the useful mathematical theories that have attracted the most attention from researchers in dealing with complex datasets and particularly on uncertainties (Ali et al., 2019; Esposito et al., 2018). Generally, these two theories have distinguished definitions of the imprecise concept. Among the popular methods in these theories are rough set and fuzzy set. The rough set theory proposes a theory of approximation (upper and lower); while the fuzzy set theory considers unclear boundaries in dealing with imprecise data. Instead of rough set and fuzzy set theories, there is another useful mathematical method that has emerged recently for handling uncertainty problems. This method, which was proposed by Molodstov and is known as the soft set theory, was developed to overcome the limitations of the classical rough set and fuzzy set theories that categorised under theory of probability and theory of fuzzy sets. It was initiated to enhance the capability of the rough set and fuzzy set theories, which have their own difficulties. Probability theories have trouble analysing non-stochastic data problems and are more suited for use in engineering and not in social science fields. Meanwhile the fuzzy set theory often has difficulties defining the membership function on a particular analysis problem (Molodtsov, 1999). While soft set theory stated that it could resolve the difficulties faced by probability and fuzzy set theories, some of the algorithms were unable to produce a sub-optimal set of attributes for complex data during the data analysis process due to computational complexity problems (Akram et al., 2019).

Based on the aforementioned problems, this study has selected the feature selection method as main research components, and complex data as its research domain area. Two key issues will be highlighted: firstly, the use of feature selection as a selection method to minimise the number of instances or attributes in a dataset by eliminating uncorrelated attributes between attributes and class. This is done by stage 1 of feature selection process by using correlation-based feature selection with best first

search method (CFS-BFS). Secondly, the feature selection method used to choose the most optimised attribute set within the complex data values in a dataset by eliminating uncertain values using a combination of soft set and rough set (SSRS) feature selection method. The aim of these multistage feature selection method is to assist the classifier in achieving better performance on accuracy by analysing problematic attribute values in the datasets.

1.3 Problem Statement

As discussed in the previous section, feature selection is one of the crucial tasks in the decision-making process, especially when dealing with big datasets. It is important for the researcher or decision-maker to select the appropriate method to be applied in the feature selection process. In order to select the most appropriate method, the description of the data such as their size and characteristics should be taken into consideration. This is because if the wrong method is selected it will decrease the decision-making performance or even increase the cost of the software and hardware used. Besides, some feature selection methods are incapable of generating the optimised set of attributes to be used in the decision analysis process. Some methods also have difficulty in analysing large-sized data sets, and will give a low analysis performance such as insufficient memory space, a long processing time, and non-deterministic polynomial-time (NP) hardness problem (Vijayanand et al., 2018; Faraway and Augustin, 2018). Therefore, it is good to really know the detailed process of making decisions so that the desired solution can be arrived at easily and will be cost effective. Of late, lots of software and hardware are available in the market. However, most of them are not so cost effective and require the decision-maker to spend a lot on the setup and maintenance such as the cost for the installation of the software and database, and also the cost of training the staff. In conclusion, two main problems that lead to other decision-making issues will be highlighted in this thesis. The problems are: i) the complexity of the data characteristics, and ii) the limitations of existing feature selection methods in handling the multiple conditions especially on uncertainty values of datasets.

Thus, this research proposed a multistage feature selection methods as an alternative in the data analysis process to deal with complex datasets by combining several feature selection methods. The main research question was: "How can an effective feature selection method be constructed by utilizing existing methods to improve the classification process?" In order to increase the understanding in constructing the proposed work, several research questions (RQ) were derived from the main research question as follows:

RQ1: What are the existing feature selection methods used for data analysis and how these methods being utilized?

RQ2: What kind of feature selection methods can be enhanced to solve data complexity issues?

RQ3: How does the proposed methods reduce the dimension and select the optimised attributes of a complex dataset?

RQ4: How does the performance of the proposed feature selection methods over the other single and hybrid methods?

RQ5: How does the performance of the proposed method with different kinds of datasets?

1.4 Research Aim

The aim of this research was to generate an optimal attribute set by reducing the irrelevant values in a dataset using the proposed multistage feature selection methods in the classification process.

1.5 Research Objectives

1. To combine correlation-based feature selection with best first search methods that can reduce the irrelevant large number of attribute.

2. To combine soft set and rough set feature selection methods that can eliminate problematic values in datasets.

1.6 Scope of the Study

This research focused mainly on problematic datasets, in particular on the uncertainty, imbalanced and inconsistency values that will be used in the analysis process to produce an effective solution. This research used different sizes of data, either instances or attributes, in order to test the performance of the proposed work. The selection of data is based on the multiple data characteristics such as multi-variate, uni-variate, time-series, sequential, missing values, numeric and text. In addition, imbalanced ratio of the datasets were also tested. This research also considered different types of feature selection methods such as single or combination of several feature selection methods as bench-marking methods to the proposed feature selection method. There are two combination approaches, i) using WEKA that implements feature selection method such as correlation-based feature selection and combined with best first search method as searching technique, ii) execute the feature selection methods in sequence such as soft set feature selection process and followed by rough set feature selection process. All the datasets were secondary datasets that were obtained from various sources such as the UC Irvine Machine Learning Repository (https://archive.ics.uci.edu/ml/index.php), Knowledge Extraction based on Evolutionary Learning (KEEL, https://sci2s.ugr.es/keel/datasets.php), Weka datasets (https://storm.cis.fordham.edu/gweiss/data-mining/datasets.html) and Kaggle datasets (https://www.kaggle.com/datasets). These datasets were selected based on the characteristics such as consist of multiple values, large size of attributes and instances and data that used for feature selection competition such as Arcene and Madelon. Finally, the software and hardware that were used in the experimental works were easily obtained and implemented, and did not require a high cost for installation and purchase. The specifications of the hardware used are as follows, Intel Core i5-8250U CPU at 1.60 GHz and 4 GB of memory using 64 bit Windows 10 operating system. The software that being used for the experimental works are Matlab 2014, Weka version 3.8.3, Rough set Exploration System (RSES), and Ms. Excel.

1.7 Significance of Findings

It is hoped that the process that was conducted and the outcome of the research can contribute either in general or to a specific target group in the data science community. Below are listed several significant outcomes of the research:

1. It is hoped that this research will provide an alternative feature selection method that can analyse data effectively with a combination of several single feature selection methods (correlation-based feature selection, best first search, soft set and rough set feature selection methods).
2. This research can provide proper guidelines to other researchers who may want to analyse data by using similar approach with provided flow charts and algorithms shown in Chapter 3.
3. This research helps the decision maker to eliminate the number of large irrelevant attribute by using proposed multistage feature selection processes that analysed uncorrelated and problematic attributes that might exist in the datasets. So that, the processing time and space of memory can be reduced while assisting the classifier in achieving high performance accuracy.
4. This research has also improved accuracy on several number of complex type of datasets such as for Arcene and Dota datasets as shown in Chapter 5 Experimental work 2.
5. The results of this research can be a guideline and a bench-marking work for other researchers in the selection of an appropriate method to be combined when dealing with imbalanced and problematic datasets.
6. The combination methods from this study can be a bench-marked method to other research works with the same process (classification).
7. The results of this research could be a comparison to other researchers that used the same datasets as being employed in the three experimental works.

1.8 Thesis Organization

The contents of this thesis have been structured into seven chapters as follows:

- Chapter 1: explains the overall view of the proposed research by discussing the research area and problem background, problem statement, the objectives, scope of the research, the research questions, significance of the research findings, and also the thesis organization.
- Chapter 2: discusses the important research areas from the related domains that have been reviewed during the research study. The important research areas include feature selection method, theoretical framework of the feature selection methods, uncertainty value of datasets, existing feature selection methods.
- Chapter 3: describes the methodology of the proposed research by presenting its conceptual framework. This chapter also defines the formulation that was used to conduct the proposed research phase by phase from the data collection until the generation of results. The detailed structures of two stages of feature selection phase that were constructed to conduct the data analysis process are also explained. The first phase, is a combination of correlation-based feature selection with the best first search named as CFS-BFS. This phase will acts as a feature reduction method that reduces the uncorrelated attribute that exist in a dataset. Meanwhile, the second phase was purposely constructed for the attribute selection process. This phase was carried out by two mathematical methods, namely the soft set and rough set feature selection methods, which were integrated together and named as the SSRS feature selection method.
- Chapter 4: presents the proposed first stage of the feature selection process known as CFS-BFS. The results and performance of the proposed method are also provided. This chapter also presents the whole process of the proposed method with several existing feature selection methods. These methods were used as a bench-marking procedure with the implementation of different classifiers to verify the performance of the proposed method.
- Chapter 5: presents two analysis works that involved evaluation of the second stage of the feature selection process. The results and its performance with

regard to imbalanced and large datasets also being reported. Several existing feature selection methods were used as a bench-marking procedure. Different classifiers were also employed to verify the performance of the proposed method.

- Chapter 6: concludes the outcomes of the study, the contributions that reflect the research objectives, the limitations and some recommendations that can be implemented in the future.

1.9 Summary

This chapter discussed the overview of the research work in relation to the domain area, problem background and statement, and the research questions that were used as a guideline to construct the research objectives and scope. Finally, this chapter gave an overall summary of the research by stating the research contributions with the thesis structure.

REFERENCES

Aberson, C. L. (2019), *Applied power analysis for the behavioral sciences*, Routledge.

Abu-Khzam, F. N., Li, S., Markarian, C., Meyer auf der Heide, F. and Podlipyan, P. (2018), 'Efficient parallel algorithms for parameterized problems', *Theoretical Computer Science* 76, 2–12.

Abubacker, N. F., Azman, A. and Doraisamy, S. (2011), 'Correlation-Based Feature Selection for Association Rule Mining in Semantic Annotation of Mammographic', *Pattern Recognition Letters* 32, 482–493.

Agarwal, M., Hanmandlu, M. and Biswas, K. K. (2011), 'Generalized intuitionistic fuzzy soft set and its application in practical medical diagnosis problem', *IEEE International Conference on Fuzzy Systems* 3, 2972–2978.

Ahmad, A., Khan, M., Paul, A., Din, S., Rathore, M. M., Jeon, G. and Choi, G. S. (2017), 'Toward modeling and optimization of features selection in Big Data based social Internet of Things', *Future Generation Computer Systems* 82, 715–726.

Ahmad, N. M. R., Herawan, T. and Deris, M. M. (2010), 'A Framework of Decision Making Based on Maximal Supported Sets', *International Symposium on Neural Networks* pp. 473–482.

Ait Hammou, B., Ait Lahcen, A. and Mouline, S. (2018), 'APRA: An approximate parallel recommendation algorithm for Big Data', *Neurocomputing* 157, 10–19.

Akram, M., Ali, G. and Alcantud, J. C. R. (2019), 'New decision-making hybrid model: intuitionistic fuzzy N-soft rough sets', *Soft Computing* 23, 9853–9868.

Alcantud, J. C. R., Varela, G., Santos-Buitrago, B., Santos-García, G. and Jiménez, M. F. (2019), 'Analysis of survival for lung cancer resections cases with fuzzy and soft set theory in surgical decision making', *PLoS ONE* 14, 1–17.

Alexopoulos, A., Georgios, D., Andreas, K., Phivos, M. and Gerasimos, V. (2020), 'Two-step classification with SVD preprocessing of distributed massive datasets in apache spark', *Algorithms* 13, 1–24.

Ali, A., Ali, M. I. and Rehman, N. (2019), 'Soft dominance based rough sets with applications in information systems', *International Journal of Approximate Reasoning* 113, 171–195.

Ali, M. I., Davvaz, B. and Shabir, M. (2013), 'Some properties of generalized rough sets', *Information Sciences* 224, 170–179.

Ali, N. and Abbas, N. A. (2018), 'Support vector machine with Dirichlet feature mapping', *Neural Networks* 98, 87–101.

Ali, R., Siddiqi, M. H. and Lee, S. (2015), 'Rough set-based approaches for discretization: a compact review', *Artificial Intelligence Review* 44(2), 235–263.

Almustafa, K. M. (2020), 'Prediction of heart disease and classifiers' sensitivity analysis', *BMC Bioinformatics* 21, 1–18.

Alomary, R. Y. and Khan, S. A. (2014), 'Fuzzy logic based multi-criteria decision-making using Dubois and Prade's operator for distributed denial of service attacks in wireless sensor networks', *2014 5th International Conference on Information and Communication Systems, ICICS 2014* pp. 1–6.

Anaraki, J. R. and Usefi, H. (2019), 'A Feature Selection based on perturbation theory', *Expert Systems with Applications* 127, 1–8.

Anowar, F., Sadaoui, S. and Selim, B. (2021), 'Conceptual and empirical comparison of dimensionality reduction algorithms (PCA, KPCA, LDA, MDS, SVD, LLE, ISOMAP, LE, ICA, t-SNE)', *Computer Science Review* 40, 100378.

Anshori, M., Mar'i, F., Alauddin, M. W. and Bachtiar, F. A. (2018), 'Prediction Result of Dota 2 Games Using Improved SVM Classifier Based on Particle Swarm Optimization', *3rd International Conference on Sustainable Information Engineering and Technology, SIET 2018 - Proceedings* pp. 121–126.

Arnaiz-Gonzalez, A., Diez-Pastor, J. F., Rodriguez, J. J. and Garcia-Osorio, C. (2016), 'Instance selection of linear complexity for big data', *Knowledge-Based Systems* 107, 83–95.

Basir, M. A., Hussin, M. S. and Yusof, Y. (2020), 'Optimization of multi-objective ENORA and NSGA-II based on bio-inspired algorithms for classification

problem', *International Journal of Advanced Trends in Computer Science and Engineering* 9, 110–116.

Bello, R. and Verdegay, J. L. (2012), 'Rough sets in the Soft Computing environment', *Information Sciences* 212, 1–14.

Błaszczy, J., Greco, S., Matarazzo, B. and Słowi, R. (2013), 'jMAF - Dominance-Based Rough Set Data', *A. Skowron and Z. Suraj (Eds.): Rough Sets and Intelligent Systems, ISRL 42, Springer* pp. 185–209.

Bouhana, A., Fekih, A., Abed, M. and Chabchoub, H. (2013), 'An integrated case-based reasoning approach for personalized itinerary search in multimodal transportation systems', *Transportation Research Part C: Emerging Technologies* 31, 30–50.

Cai, J., Luo, J., Wang, S. and Yang, S. (2018), 'Feature selection in machine learning: A new perspective', *Neurocomputing* 300, 70–79.

Canbolat, A. S., Bademlioglu, A. H., Arslanoglu, N. and Kaynakli, O. (2019), 'Performance optimization of absorption refrigeration systems using Taguchi, ANOVA and Grey Relational Analysis methods', *Journal of Cleaner Production* 229, 874–885.

Chai, J. and Liu, J. N. K. (2014), 'A novel believable rough set approach for supplier selection', *Expert Systems with Applications* 41(1), 92–104.

Chen, K., Zhou, F. Y. and Yuan, X. F. (2019), 'Hybrid particle swarm optimization with spiral-shaped mechanism for feature selection', *Expert Systems with Applications* 128, 140–156.

Chen, Z., He, C., He, Z. and Chen, M. (2018), 'BD-ADOPT: a hybrid DCOP algorithm with best-first and depth-first search strategies', *Artificial Intelligence Review* 50, 161–199.

Chormunge, S. and Jena, S. (2018), 'Correlation based feature selection with clustering for high dimensional data', *Journal of Electrical Systems and Information Technology* pp. 4–11.

Choudhury, S. J. and Pal, N. R. (2019), 'Imputation of missing data with neural networks for classification', *Knowledge-Based Systems* 182, 104838.

Cong, Y., Wang, S., Fan, B., Yang, Y. and Yu, H. (2016), 'UDSFS: Unsupervised deep sparse feature selection', *Neurocomputing* 196, 150–158.

Das, S. and Kar, S. (2013), 'LNCS 8251 - Intuitionistic Multi Fuzzy Soft Set and its Application in Decision Making', pp. 587–592.

Ding, W., Lin, C. T., Chen, Senbo Zhang, X. and Hu, B. (2018), 'Multiagent-consensus-MapReduce-based attribute reduction using co-evolutionary quantum PSO for big data applications', *Neurocomputing* 272, 136–153.

Dong, H., Li, T., Ding, R. and Sun, J. (2018), 'A novel hybrid genetic algorithm with granular information for feature selection and optimization', *Applied Soft Computing Journal* 65, 33–46.

Durbach, I. N. and Stewart, T. J. (2012), 'Modeling uncertainty in multi-criteria decision analysis', *European Journal of Operational Research* 223(1), 1–14.

Erfani, S. M., Sutharshan, R., Shanika, K. and Christopher, L. (2016), 'High-dimensional and large-scale anomaly detection using a linear one-class SVM with deep learning', *Pattern Recognition* 58, 121–134.

Esposito, M., Minutolo, A., Megna, R., Forastiere, M., Magliulo, M. and De Pietro, G. (2018), 'A smart mobile, self-configuring, context-aware architecture for personal health monitoring', *Engineering Applications of Artificial Intelligence* 67, 136–156.

Fahmideh, M. and Beydoun, G. (2018), 'Big data analytics architecture design—An application in manufacturing systems', *Computers and Industrial Engineering* pp. 1–16.

Faraway, J. J. and Augustin, N. H. (2018), 'When small data beats big data', *Statistics and Probability Letters* 136, 142–145.

Fayed, H. A. and Atiya, A. F. (2019), 'Speed up grid-search for parameter selection of support vector machinesn', *Applied Soft Computing Journal* 80, 202–210.

Feng, F., Li, C., Davvaz, B. and Ali, M. I. (2010), 'Soft sets combined with fuzzy sets and rough sets: A tentative approach', *Soft Computing* 14(9), 899–911.

Feng, F., Liu, X., Leoreanu-Fotea, V. and Jun, Y. B. (2011), 'Soft sets and soft rough sets', *Information Sciences* 181(6), 1125–1137.

Frăsinaru, C. and Răschip, M. (2019), 'Greedy Best-First Search for the Optimal-Size Sorting Network Problem', *Procedia Computer Science* 159, 447–454.

Gadekallu, T. R., Bhattacharya, N. K. S., Maddikunta, S. S. P. K. R., Ra, I.-H. and Alazab, M. (2020), 'Early detection of diabetic retinopathy using PCA-firefly based deep learning model', *Electronics MDPI* 9, 274.

Gao, W., Hu, L., Zhang, P. and Wang, F. (2018), 'Feature selection by integrating two groups of feature evaluation criteria', *Expert Systems with Applications* 110, 11–19.

García-Gil, D., Sergio, R.-G., Salvador, G. and Francisco, H. (2018), 'Principal Components Analysis Random Discretization Ensemble for Big Data', *Knowledge-Based Systems* 108, 1977–1993.

Geng, S., Li, Y., Feng, F. and Wang, X. (2011), 'Generalized intuitionistic fuzzy soft sets and multiattribute decision making', *Proceedings - 2011 4th International Conference on Biomedical Engineering and Informatics, BMEI 2011* 4, 2206–2211.

Ghaddar, B. and Naoum-Sawaya, J. (2018), 'High dimensional data classification and feature selection using support vector machines', *European Journal of Operational Research* 265, 993–1004.

Gong, Z. T., Xie, T., Shi, Z. H. and Pan, W. Q. (2011), 'A Multiparameter Group Decision Making Method Based on the Interval-valued Intuitionistic Fuzzy Soft Sets', *Proceedings of the 2011 International Conference on Machine lLearning and Cybernetics* pp. 10–13.

Grama, A., Anshul, G., George, K. and Vipin, K. (2003), 'Principles of parallel algorithm design. Introduction to Parallel Computing. 2nd ed. Harlow: Addison Wesley'.

Greco, S., Matarazzo, B. and Słowiński, R. (2013), 'Beyond Markowitz with multiple criteria decision aiding', *Journal of Business Economics* 83(1), 29–60.

Guan, X., Li, Y. and Feng, F. (2013), 'A new order relation on fuzzy soft sets and its application', *Soft Computing* 17(1), 63–70.

Guyon, I. (2003), 'Design of experiments of the NIPS 2003 variable selection benchmark', *NIPS 2003 workshop on feature extraction* pp. 1–30.

Hall, M. A. (1999), *Correlation-based Feature Selection for Machine Learning*, The University of Waikato, Hamilton, New Zealand.

Harous, S., El Menshawy, M., Serhani, M. A. and Benharref, A. (2018), 'Mobile health architecture for obesity management using sensory and social data', *Informatics in Medicine Unlocked* 10, 27–44.

Hashem, E. M. and Mabrouk, M. S. (2014), 'A Study of Support Vector Machine Algorithm for Liver Disease Diagnosis', 4(1), 9–14.

Hassan, N. and Al-Qudah, Y. (2019), 'Fuzzy parameterized complex multi-fuzzy soft set', *Journal of Physics: Conference Series* 1212.

Herawan, T. and Deris, M. M. (2010), 'Soft Decision Making for Patients Suspected Influenza', *LNCS 6018 - Computational Science and Its Applications ... ICCSA 2010* pp. 405–418.

Herawan, T., Deris, M. M. and Abawajy, J. H. (2010), 'Matrices Representation of Multi Soft-Sets and Its Application', *LNCS 6018 - Computational Science and Its Applications, ICCSA 2010* p. 201–214.

Hernandez, E. A. and Uddameri, V. (2010), 'Selecting Agricultural Best Management Practices for Water Conservation and Quality Improvements Using Atanassov's Intuitionistic Fuzzy Sets', *Water Resources Management* 24, 4589–4612.

Hosseini, E. S. and Moattar, M. H. (2019), 'Evolutionary feature subsets selection based on interaction information for high dimensional imbalanced data classification', *Applied Soft Computing Journal* 82, 105581.

Houari, R., Bounceur, A., Kechadi, M. T. and Tari, A. Kamel Euler, R. (2016), 'Dimensionality reduction in data mining: A Copula approach', *Expert Systems with Applications* 64, 247–260.

Hu, J., Pan, L., Yang, Y. and Chen, H. (2019), 'A group medical diagnosis model based on intuitionistic fuzzy soft sets', *Applied Soft Computing Journal* 77.

Hu, M.-l., Shen, F.-f. and Chen, Y.-h. (2011), 'A multi-attribute decision analysis method based on rough sets dealing with uncertain information', *Proceedings of 2011 IEEE International Conference on Grey Systems and Intelligent Services* pp. 576–581.

Idris, I., Selamat, A. and Omatu, S. (2014), 'Hybrid email spam detection model with negative selection algorithm and differential evolution', *Engineering Applications of Artificial Intelligence* 28, 97–110.

Irfan Ali, M. (2011), 'A note on soft sets, rough soft sets and fuzzy soft sets', *Applied Soft Computing Journal* 11(4), 3329–3332.

Jain, I., Jain, V. K. and Jain, R. (2018), 'Correlation feature selection based improved-Binary Particle Swarm Optimization for gene selection and cancer classification', *Applied Soft Computing Journal* 62, 203–215.

Jing, Y., Li, T., Fujita, H., Wang, B. and Cheng, N. (2018), 'An incremental attribute reduction method for dynamic data mining', *Information Sciences* 465, 202–218.

Jothi, G. and Hannah, I. H. (2012), 'Soft set based quick reduct approach for unsupervised feature selection', *Proceedings of 2012 IEEE International Conference on Advanced Communication Control and Computing Technologies, ICACCCT 2012* pp. 277–281.

Kamacı, H. (2020), 'Selectivity analysis of parameters in soft set and its effect on decision making', *International Journal of Machine Learning and Cybernetics* 11, 313–324.

Keeney, R. L. (1982), 'Decision Analysis: An Overview', 30, 803–838.

Kewat, A., Srivastava, P. N. and Kumhar, D. (2020), 'Performance Evaluation of Wrapper-Based Feature Selection Techniques for Medical Datasets', *Advances in Computing and Intelligent Systems.* 11, 619–633.

Kim, K. J. and Jun, C. H. (2018), 'Rough set model based feature selection for mixed-type data with feature space decomposition', *Expert Systems with Applications* 103, 196–205.

Kim, Song Hyun, V., Thanh Mai, P. and Ho, C. (2017), 'A preliminary study on applicability of artificial neural network for optimized reflector designs', *Energy Procedia* 131, 77–85.

Koc, L., Mazzuchi, T. a. and Sarkani, S. (2012), 'A network intrusion detection system based on a Hidden Naïve Bayes multiclass classifier', *Expert Systems with Applications* 39(18), 13492–13500.

Kong, Z., Wang, L. and Wu, Z. (2012), 'Two Cases Based on Normal Parameter Reduction in Soft Sets', *2012 International Conference on Computer Science and Electronics Engineering* pp. 577–581.

Korf, R. E. (1993), 'Linear-space best-first search', *Artificial Intelligence* 62, 41–78.

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

Kushal, B., Illindala, T. R. and S, M. (2020), 'Correlation-based feature selection for resilience analysis of MVDC shipboard power system', *International Journal of Electrical Power and Energy Systems* 117, 105742.

Labbé, M., Martínez-Merino, L. I. and Rodríguez-Chía, A. M. (2018), 'Mixed integer linear programming for feature selection in support vector machine', *Discrete Applied Mathematics* pp. 276–304.

Lashari, S. A. and Ibrahim, R. (2013), 'A Framework for Medical Images Classification Using Soft Set', *Procedia Technology* 11, 548–556.

Lashari, S. A., Ibrahim, R. and Senan, N. (2012), 'Soft set theory for automatic classification of traditional Pakistani musical instruments sounds', *2012 International Conference on Computer Information Science (ICCIS)* 1, 94–99.

Li, A. D., Bing, X. and Mengjie, Z. (2020), 'Multi-objective feature selection using hybridization of a genetic algorithm and direct multisearch for key quality characteristic selection', *Information Sciences* 523, 245–265.

Li, H., Li, D., Zhai, Y., Wang, S. and Zhang, J. (2016), 'A novel attribute reduction approach for multi-label data based on rough set theory', *Information Sciences* 367–368, 827–847.

Li, P., Wu, J. and Qian, H. (2012), 'Ground water quality assessment based on rough sets attribute reduction and TOPSIS method in a semi-arid area, China', *Environmental Monitoring and Assessment* 184(8), 4841–4854.

Li, Z., Liang, P., Avgeriou, P. and Guelfi, N. (2014), 'A systematic mapping study on technical debt', *Under submission* 101, 193–220.

Li, Z. and Xie, T. (2014), 'The relationship among soft sets, soft rough sets and topologies', *Soft Computing* 18(4), 717–728.

Lin, Y., Li, Y., Wang, C. and Chen, J. (2018), 'Attribute reduction for multi-label learning with fuzzy rough set', *Knowledge-Based Systems* 152, 51–61.

Little, R. J. and Rubin., D. B. (2019), *Statistical analysis with missing data.*, Vol. 793, John Wiley and Sons.

Liu, J., Lin, Y., Li, Y., Weng, W. and Wu, S. (2018), 'Online multi-label streaming feature selection based on neighborhood rough set', *Computers in Industry* 84, 273–287.

Liu, J. and Zio, E. (2019), 'Integration of feature vector selection and support vector machine for classification of imbalanced data', *Future Generation Computer Systems* 75, 702–711.

Liu, P., Baoying, Z., Peng, W. and Mengjiao, S. (2020), 'An approach based on linguistic spherical fuzzy sets for public evaluation of shared bicycles in China', *Engineering Applications of Artificial Intelligence* 87, 103295.

Luan, C. and Dong, G. (2018), 'Experimental identification of hard data sets for classification and feature selection methods with insights on method selection', *Data and Knowledge Engineering* 118, 41–51.

Ma, X., Liu, Q. and Zhan, J. (2017), 'A survey of decision making methods based on certain hybrid soft set models', *Artificial Intelligence Review* 47, 507–530.

Ma, X., Qin, H., Sulaiman, N., Herawan, T. and Abawajy, J. (2014), 'The Parameter Reduction of the Interval-Valued Fuzzy Soft Sets and Its Related Algorithms', *IEEE Transactions on Fuzzy Systems* 22(1), 57–71.

Ma, X., Sulaiman, N. and Qin, H. (2011), 'Parameterization value reduction of soft sets and its algorithm', *2011 IEEE Colloquium on Humanities, Science and Engineering, CHUSER 2011* pp. 261–264.

Ma, X. and Wang, G. (2011), 'An extended soft set model: Type-2 fuzzy soft sets', *CCIS2011 - Proceedings: 2011 IEEE International Conference on Cloud Computing and Intelligence Systems* pp. 128–133.

Mahindru, A. and Sangal, A. L. (2020), 'Dldroid: Feature selection based malware detection framework for android apps developed during COVID-19', *International Journal on Emerging Technologies* 11, 516–525.

Massimiani, A., Palagi, L., Sciubba, E. and Tocci, L. (2017), 'Neural networks for small scale ORC optimization', *Energy Procedia* 129, 34–41.

Mathew, M., K., C. R. and J., R. M. (2020), 'A novel approach integrating AHP and TOPSIS under spherical fuzzy sets for advanced manufacturing system selection', *Engineering Applications of Artificial Intelligence* 96, 103988.

Maugis, P. A. G. (2018), 'Big data uncertainties', *Journal of Forensic and Legal Medicine* 57, 7–11.

Meena, K. A., R., M. and N., G. (2019), 'Correlation Based Feature Selection Algorithms for Varying Datasets of Different Dimensionality', *Wireless Personal Communications* 108, 1977–1993.

Meng, D., Zhang, X. and Qin, K. (2011), 'Soft rough fuzzy sets and soft fuzzy rough sets', *Computers and Mathematics with Applications* 62(12), 4635–4645.

Miao, D., Duan, Q., Zhang, H. and Jiao, N. (2009), 'Rough set based hybrid algorithm for text classification', *Expert Systems with Applications* 36(5), 9168–9174.

Minaei-Bidgoli, B., Barmaki, R. and Nasiri, M. (2013), 'Mining numerical association rules via multi-objective genetic algorithms', *Information Sciences* 233, 15–24.

Mirzaei, S., El Yacoubi, M., Garcia-Salicetti, S., Boudy, J., Kahindo, C., Cristancho-Lacroix, V. and Kerhervé, H. snd Rigaud, A. S. (2018), 'Two-Stage Feature Selection of Voice Parameters for Early Alzheimer's Disease Prediction', *Irbm* 39, 430–435.

Mohamad, M. and Selamat, A. (2017), 'A New Soft Rough Set Parameter Reduction Method for an Effective Decision-Making', *New Trends in Intelligent Software Methodologies, Tools and Techniques* 297, 691–704.

Mohamad, M. and Selamat, A. (2018*a*), 'A Two-Tier Hybrid Parameterization Framework for Effective Data Classification', *New Trends in Intelligent Software Methodologies, Tools and Techniques* 303, 321–331.

Mohamad, M. and Selamat, A. (2018*b*), 'Analysis on Hybrid Dominance-Based Rough Set Parameterization Using Private Financial Initiative Unitary Charges Data', *LNAI Asian Conference on Intelligent Information and Database Systems* pp. 318–328.

Mohsen, H., El-Dahshan, E.-S. A., El-Horbaty, E.-S. M. and Salem, A.-B. M. (2017), 'Classification using Deep Learning Neural Networks for Brain Tumors', *Future Computing and Informatics Journal* 3, 68–71.

Molodtsov, D. (1999), 'Soft set theory-first results', *Computers and Mathematics with Applications* 37(4), 19–31.

Moonsamy, V., Rong, J. and Liu, S. (2019), 'Ensemble feature selection using election methods and ranker clustering', *Information Sciences* 480, 365–380.

Mursalin, M., Zhang, Y., Chen, Y. and Chawla, N. V. (2017), 'Automated epileptic seizure detection using improved correlation-based feature selection with random forest classifier', *Neurocomputing* 241, 204–214.

Nguyen, H. S. and Skowron, A. (2013), 'Rough Sets: From Rudiments to Challenges', *Intelligent Systems Reference Library* 42, 75–173.

Ortega, J., János, T., Sarbast, M., Tamás, P. and Szabolcs, D. (2020), 'An integrated approach of analytic hierarchy process and triangular fuzzy sets for analyzing the park-and-ride facility location problem', *Symmetry* 12, 103988.

Oussous, A., Benjelloun, F. Z., Ait Lahcen, A. and Belfkih, S. (2018), 'Big Data technologies: A survey', *Journal of King Saud University - Computer and Information Sciences* 30, 431–448.

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

Palma-Mendoza, R. J., De-Marcos, L., Rodriguez, D. and Alonso-Betanzos, A. (2019), 'Distributed correlation-based feature selection in spark', *Information Sciences* 496, 287–299.

Paradarami, T. K., Bastian, N. D. and Wightman, J. L. (2017), 'A hybrid recommender system using artificial neural networks', *Expert Systems with Applications* 83, 300–313.

Pawlak, Z. (1997), 'Rough set approach to knowledge-based decision support', *European Journal of Operational Research* 99, 48–57.

Pawlak, Z. (1998), 'Rough set theory and its applications', *Journal of Telecommunications and Information Technology* 29, 7–10.

Pillans, J. (2021), 'Efficiency of evolutionary search for analog filter synthesis', *Expert Systems with Applications* 168, 114267.

Polat, H., Danaei Mehr, H. and Cetin, A. (2017), 'Diagnosis of Chronic Kidney Disease Based on Support Vector Machine by Feature Selection Methods', *Journal of Medical Systems* 41, 55.

Qian, Y., Liang, X., Wang, Q., Liang, J., Liu, B., Skowron, Andrzej andYao, Y., Ma, J. and Dang, C. (2018), 'Local rough set: A solution to rough data analysis in big data', *International Journal of Approximate Reasonings* 97, 38–63.

Rao, H., Shi, X., Rodrigue, A. K., Feng, J., Xia, Y., Elhoseny, M., Yuan, X. and Gu, L. (2019), 'Feature selection based on artificial bee colony and gradient boosting decision tree', *Applied Soft Computing Journal* 74, 634–642.

Rima, H., Ahcène, B., Tahar, K. M., Kamel, T. A. and Reinhardt, E. (2016), 'Dimensionality reduction in data mining: A Copula approach', *Expert Systems with Applications* 64, 247–260.

Rui, Z., Feiping, N., Xuelong, L. and Xian, W. (2019), 'Feature selection with multi-view data: A survey', *Information Fusion* 50, 158–167.

Sadiq, A. S., Tahir, M. A., Ahmed, A. A. and Alghushami, A. (2019), 'Normal parameter reduction algorithm in soft set based on hybrid binary particle swarm and biogeography optimizer', *Neural Computing and Applications* 2, 1–19.

Shabir, M., Irfan Ali, M. and Shaheen, T. (2013), 'Another approach to soft rough sets', *Knowledge-Based Systems* 40, 72–80.

Shah, T., Medhit, S. and Farooq, G. (2015), 'Intuitionistic Fuzzy Soft Set Decision Criterion for Selecting Appropriate Block Cipher', *3D Research* 6(3), 32.

Sheeja, T. K. and Kuriakose, A. S. (2018), 'A novel feature selection method using fuzzy rough sets', *Computers in Industry* 97, 111–121.

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

Singh, A., Rajan, P. and Bhavsar, A. (2020), 'SVD-based redundancy removal in 1-D CNNs for acoustic scene classification', *Expert Systems with Applications* 131, 383–389.

Singh, D. and Singh, B. (2019), 'Hybridization of feature selection and feature weighting for high dimensional data', *Applied Intelligence* 49, 1580–1596.

Singh, M. and Pamula, R. (2019), 'An outlier detection approach in large-scale data stream using rough set', *Neural Computing and Applications* 4, 1–15.

Singhal, N., Verma, A. and Chouhan, U. (2018), 'An Application of Similarity Measure of Fuzzy Soft Sets in Verndor Selection Problem', *Materials Today: Proceedings* 5, 3987–3993.

Soliman, O. S. and Rassem, A. (2012), 'Correlation based feature selection using quantum bio inspired estimation of distribution algorithm', *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)* 7694 LNAI, 318–329.

Sood, S. K., Sandhu, R., Singla, K. and Chang, V. (2018), 'IoT, big data and HPC based smart flood management framework', *Sustainable Computing: Informatics and Systems* pp. 1–16.

Sun, B. and Ma, W. (2011), 'Soft fuzzy rough sets and its application in decision making', *Artificial Intelligence Review* pp. 67–80.

Szelag, M., Greco, S. and Słowiński, R. (2014), 'Variable consistency dominance-based rough set approach to preference learning in multicriteria ranking', *Information Sciences* 277, 525–552.

Tao, Z., Huiling, L., Wenwen, W. and Xia, Y. (2019), 'GA-SVM based feature selection and parameter optimization in hospitalization expense modeling', *Applied Soft Computing* 75, 323–332.

Tavanaei, A., Ghodrati, M., Kheradpisheh, S. R., Masquelier, T. and Maida, A. S. (2018), 'Deep Learning in Spiking Neural Networks', *Neural Networks* 111, 47–63.

Teixeira de Lima, G. R. and Stephany, S. (2013), 'A new classification approach for detecting severe weather patterns', *Computers & Geosciences* 57, 158–165.

Teng, S. H., Lu, M., Yang, A. F., Zhang, J., Nian, Y. and He, M. (2016), 'Efficient attribute reduction from the viewpoint of discernibility', *Information Sciences* 326, 297–314.

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

Tsai, C. F., Eberle, W. and Chu, C. Y. (2013), 'Genetic algorithms in feature and instance selection', *Knowledge-Based Systems* 13, 240–247.

Tseng, T.-L. B., Huang, C.-C., Fraser, K. and Ting, H.-W. (2016), 'Rough set based rule induction in decision making using credible classification and preference from medical application perspective', *Computer Methods and Programs in Biomedicine* 127, 273–289.

Uddin, M. P., Md, A. M., Masud, I. A. and Hossain, M. A. (2021), 'Information-theoretic feature selection with segmentation-based folded principal component analysis (PCA) for hyperspectral image classification', *International Journal of Remote Sensing* 42, 286–321.

Vijayanand, R., Devaraj, D. and Kannapiran, B. (2018), 'Intrusion detection system for wireless mesh network using multiple support vector machine classifiers with genetic-algorithm-based feature selection', *Computers and Security* 77, 304–314.

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

Wang, H., Min, Z. Y. and Zhou, Y. (2019), 'A risk evaluation method to prioritize failure modes based on failure data and a combination of fuzzy sets theory and grey theory', *Engineering Applications of Artificial Intelligence* 82, 216–225.

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

Wang, Y. and Feng, L. (2018), 'Hybrid feature selection using component co-occurrence based feature relevance measurement', *Expert Systems with Applications* 102, 83–99.

Wason, R. (2018), 'Deep learning: Evolution and expansion', *Cognitive Systems Research* 52, 701–708.

Weng, C.-H., Huang, T. C.-K. and Han, R.-P. (2016), 'Disease prediction with different types of neural network classifiers', *Telematics and Informatics* 33, 277–292.

Xiao, Z., Chen, W. and Li, L. (2013), 'A method based on interval-valued fuzzy soft set for multi-attribute group decision-making problems under uncertain environment', *Knowledge and Information Systems* 34(3), 653–669.

Xu, W., Pan, Y., Chen, W. and Fu, H. (2019), 'Forecasting corporate failure in the Chinese energy sector: A novel integrated model of deep learning and support vector machine', *Energies* 12, 2251.

Yang, X. and Yao, Y. (2018), 'Ensemble selector for attribute reduction', *Applied Soft Computing Journal* 70, 1–11.

Yang, Z. and Chen, Y. (2013), 'Fuzzy soft set-based approach to prioritizing technical attributes in quality function deployment', *Neural Computing and Applications* 23(7-8), 2493–2500.

Yaya, L., Keyun, Q. and Luis, M. (2018), 'Improving decision making approaches based on fuzzy soft sets and rough soft sets', *Applied Soft Computing Journal* 65, 320–332.

Yi, L., Yinghui, Z., Jie, L. and Zhusong, L. (2018), 'Secure and fine-grained access control on e-healthcare records in mobile cloud computing', *Future Generation Computer Systems* 78, 1020–1026.

Zahedi, K. A. and Kılıçman, A. (2019), 'Multi-attribute decision-making based on soft set theory: a systematic review', *Soft Computing* 23, 6899–6920.

Zhan, J. and Alcantud, J. C. R. (2017), 'A survey of parameter reduction of soft sets and corresponding algorithms', *Artificial Intelligence Review* 52, 1–34.

Zhang, K., Zhan, J. and Wu, W. Z. (2019), 'Novel fuzzy rough set models and corresponding applications to multi-criteria decision-making', *Fuzzy Sets and Systems* 1, 1–35.

Zhang, Q., Yang, L. T., Chen, Z. and Li, P. (2018), 'A survey on deep learning for big data', *Information Fusion* 42, 146–157.

Zhang, T., Xia, D., Tang, H., Yang, X. and Li, H. (2016), 'Classification of steel samples by laser-induced breakdown spectroscopy and random forest', *Chemometrics and Intelligent Laboratory Systems* 157, 196–201.

Zhang, Z. (2012), 'A rough set approach to intuitionistic fuzzy soft set based decision making', *Applied Mathematical Modelling* 36(10), 4605–4633.

Zhou, P., Hu, X., Li, P. and Wu, X. (2017), 'Online feature selection for high-dimensional class-imbalanced data', *Knowledge-Based Systems* 136, 187–199.

Zhou, Q., Hongming, M. and Yong, D. (2020), 'A new divergence measure of Pythagorean fuzzy sets based on belief function and its application in medical diagnosis', *Mathematics* 8, 142.

LIST OF PUBLICATIONS

Journal with Impact Factor

1. **Mohamad, M.**, Selamat, A., Subroto, I. M., & Krejcar, O. (2019). Improving the classification performance on imbalanced datasets via new hybrid parameterisation model. Journal of King Saud University-Computer and Information Sciences. https://doi.org/10.1016/j.jksuci.2019.04.009.
2. **Mohamad, M.**, Ondrej Krejcar, Hamido Fujita, & Tao Wu (2020). An analysis on new hybrid parameter selection model performance over big dataset. Knowledge-Based Systems. https://doi.org/10.1016/j.knosys.2019.105441. (Q1, IF:5.101)

Indexed conference proceedings

1. **Mohamad, M.** & Selamat, A. (2019). An Analysis on Performance of Different Type Classifiers in Handling Big datasets. In SoMeT (pp. 298-309). IOS Press. https://10.3233/FAIA190057. (Indexed by SCOPUS)
2. **Mohamad, M.** & Selamat, A. (2019). An Analysis on Deep Learning Approach Performance in Classifying Big dataset. In 1st International Conference on Artificial Intelligence and Data Sciences (AiDAS) (pp. 35-39). IEEE. 10.1109/AiDAS47888.2019.8970980 (Indexed by SCOPUS)
3. **Mohamad, M.** & Selamat, A. (2018). A two-tier hybrid parameterisation framework for effective data classification. In SoMeT (pp. 321-331). IOS Press. https://10.3233/978-1-61499-900-3-321. (Indexed by SCOPUS)
4. **Mohamad, M.** & Selamat, A. (2018). Analysis on Hybrid Dominance-Based Rough Set parameterisation Using Private Financial Initiative Unitary Charges Data. In Asian Conference on Intelligent Information and Database Systems (pp. 318-328). Springer, Cham. https://doi.org/10.1007/978-3-319-75417-8_30. (Indexed by SCOPUS)

5. **Mohamad, M.** & Selamat, A. (2017). A new soft rough set parameter reduction method for an effective decision-making. In SoMeT (pp. 691-704). IOS Press. https://10.3233/978-1-61499-800-6-691. (Indexed by SCOPUS)

6. **Mohamad, M.** & Selamat, A. (2016). A new hybrid rough set and soft set parameter reduction method for spam e-mail classification task. In Pacific Rim Knowledge Acquisition Workshop (pp. 18-30). Springer, Cham. https://doi.org/10.1007/978-3-319-42706-5_2. (Indexed by SCOPUS)

7. **Mohamad, M.** & Selamat, A. (2016). Recent study on the application of hybrid rough set and soft set theories in decision analysis process. In International Conference on Industrial, Engineering and Other Applications of Applied Intelligent Systems (pp. 713-724). Springer, Cham. https://doi.org/10.1007/978-3-319-42007-3_61. (Indexed by SCOPUS)

8. **Mohamad, M.**, Selamat, A. Krejcar, O. & Kuca, K. (2015). A recent study on the rough set theory in multi-criteria decision analysis problems. In Computational Collective Intelligence (pp. 265-274). Springer, Cham. https://doi.org/10.1007/978-3-319-24306-1_26. (Indexed by SCOPUS)

9. **Mohamad, M.** & Selamat, A. (2015). An evaluation on the efficiency of hybrid feature selection in spam email classification. In 2015 International Conference on Computer, Communications, and Control Technology (I4CT) (pp. 227-231). IEEE. https://doi.org/10.1109/I4CT.2015.7219571. (Indexed by SCOPUS)

Non-Indexed conference proceedings

1. **Mohamad, M.** & Selamat, A. (2017, April). An analysis of rough set-based application tools in the decision-making process. In International Conference of Reliable Information and Communication Technology (pp. 467-474). Springer, Cham. https://doi.org/10.1007/978-3-319-59427-9_49.