Journal of Data Acquisition and Processing
ISSN: 1004-9037
https://sjcjycl.cn/
DOI: 10.5281/zenodo.755213

# NEW DEVELOPED METHODS USED FOR DATA SCIENCE OPTIMIZATION AS STATE-OF-THE-ART

**Mohamed Abdeldaiem Abdelhadi Mahboub**[1*]
[1*]Information Systems Department- Faculty of Information Technology, University of Tripoli- Tripoli-Libya. E-mail: m.mahboub@uot.edu.ly

***Corresponding Author:** Mohamed Abdeldaiem Abdelhadi Mahboub
*E-mail: m.mahboub@uot.edu.ly

**Abstract**
Nowadays, there is a potential need in the new era of data science for introducing newly developed methods and algorithms to be used to formulate data science models by optimization solutions. We are very much concerned with studying and improving extraordinary work as Stat-of-the-Art methods and/or algorithms for solving data science problems with respect to its scalability, and efficiency; which mainly include gradient-descent based algorithms, derivative free algorithms. We do really believe that the best method which has the ability to meet our goals for optimization solutions; is to benefit from using machine learning capabilities. Optimization formulations and algorithms are both possible to lead the development of new optimization approaches that make significant changes presented by machine learning applications.
**Keywords:** Optimization Solutions, Data Science, new Soft-Set theory applications, Machine Learning, Deep Learning.

## 1-An Introduction

Optimization methods have been developed and integrated by means of different algorithms, which are becoming essential methods in many scientific and technological fields [2],[3]; particularly in data science, where very large amounts of data need to be preprocessed very fast. We would start up from conventional optimization algorithms such as interior-point methods to the old and classic first-order gradient methods. Specific steps will make some new possible extensions, analyses being derived and/or rediscovered recently worthy [7], [8]. The main goal of this research work is to introduce our state-of-the-art methods and algorithms in data science optimization. The key to valuable research works will cover both theoretical analyses, and to understand which methods are the best methods for solving what kind of data Science problems have to be optimized [11]. In our research work we would like to reuse the mathematical concepts [5], operations, and notation required for studying formal language theory and automata theory which will be our taxonomy focused on the research for improving the data Science optimization based on some newly developed methods deals with the mathematical concepts. As it is known, formal language theory has its roots in linguistics, mathematical logic, and computer science for developing programming languages in terms of the science of finite state machines [1]. We have studied the soft set theory on both sides,

theatrically and in practice. We have also suggested some new ideas for the soft sets theory within its application.

## 2-Motivation

The most important facts about optimization in research works was how to improve the preprocessing phase to scale up the optimization for a targeted high degree of any performance systems in data science; which we have taken in advance as the motivation rules for satisfying our optimization methods [6].

## 3-Methodology

We have chosen the following optimized methods [table-1] for our research work to be the methodology for our case study of the preprocessing used by machine learning [14]. In our research paper, we have also studied the most widely used methods and techniques in preprocessing for data science as well as optimized methods. To understand those selected methods and techniques quite well in terms of mathematical and computational aspects, we have managed the entire data science into such simple described tables which include details of all studied methods within their attributes names, and values. Table-1 illustrates the methods organization for the next steps in the research work [6], [7].

## 4- Objectives

Our main goal in this research work was to find out the best optimization results from our new proposed methods, we have studied the soft set theory on both sides, theoretically and in practice [2], [3]. We have also suggested some new ideas for the soft sets theory within its applications [5]. It has brought some good and simple representations of the powerful tools as a state of the Art for decision-making about data science, data mining, and drawing conclusions from data, and the merge of total function within the soft set transformation can lead to a perfect result in optimization by the preprocessing methods used in our research work [6].

**Table-1: the most methods used for the preprocessing in Data Science optimization**

## 5-Related Works

| No | Method Name | Method description | Method parameters |
|---|---|---|---|
| **1** | **Data Cleaning** | **Data Cleaning is the first step in most techniques used to data preprocessing which usually consists the following steps as to, removing missing values, outliers, and any unnecessary data. This step helps to ensure that the data is accurate and the preprocessing is also a crucial step in data science for optimization.** | **a** |

| | | | |
|---|---|---|---|
| **2** | **Scaling and Normalization** | **Scaling and Normalization are both techniques must be used to transform features to a similar scale. This step of process helps to ensure that all features are quite equally important in the model.** | **b** |
| **3** | **Feature Selection** | **Feature Selection is the step of process of selecting the most relevant features for the model. This step helps to reduce the dimensionality of the data and improve the accuracy of the model.** | **c** |
| **4** | **Feature Engineering** | **Feature Engineering step has to creating new features from existing ones. This step helps to improve the accuracy of the model optimization by providing more information about the meta data.** | **d** |
| **5** | **Data Augmentation** | **Data Augmentation is the step of process used to increase the size of the dataset by creating new data from existing data. This technique can help to improve the accuracy of the model by providing more data for the model to learn from.** | **e** |
| **6** | **Parallel Processing** | **Parallel Processing is compulsory step to use some techniques to speed up the preprocessing phase by running multiple processes simultaneously. This technique can help to reduce the time required for preprocessing large datasets. By implementing these techniques, we can optimize the preprocessing stage and improve the accuracy and efficiency of our model.** | **f** |

Numerous researches have been conducted on optimization algorithms and techniques in the last recent years. The research works were focused on optimization models and frameworks that dealt with the methods used to improve the performance of most computer systems models. We have in our research work studied the importance of the soft set theory by which its applications in several domains have great results in the practice of information systems. Molodtsov [2], has presented some applications for the soft set theory in several directions namely; the study of the smoothness of functions, game theory, operations research, and theory of measurement, Maji [3], has also presented some new improved results as an application of

neutrosophic soft set in a decision making problem. Andreas et al [4], have studied the relationship between vector optimization and financial risk measures. Q. Zhong and X. Wang [5], they have introduced a new parameter reduction method based on soft set theory. Nasef et al [6], have presented a decision-making problem for a real estate marketing approach. Endert et al [7], have presented a summary of the progress and synthesizing select research advances. Kaiwen L et al [8], have introduced a systematic comparison of many algorithms covering various categories to solve many-objective problems. Radwa et al [9], have presented a comprehensive survey of the state-of-the-art efforts in tackling the CASH problem. Ebubeogu et al [10], have presented an assessment of the existing literature to identify the key issues related to data quality to provide a convenient collection of some techniques used to perform data preprocessing. Amir Ahmad, Shehroz S. Khan [11], have presented a taxonomy for the study of mixed data clustering algorithms by identifying some major research themes. Seba Susan et al [12], have presented in their paper surveys a plethora of conventional and recent techniques that address intelligent representations of samples from the majority and minority classes. T. Dharma et al [13], have presented some different optimization algorithms such as gradient descent, mini-batch gradient descent, momentum, NAG, RMS prop, Adagrad, and Adam. Abdu-rakhmon Sadiev et al [14] have presented federated learning (FL) as a framework for distributed learning and optimization where multiple nodes connected over a network try to collaboratively carry out a learning task. Syed Muzamil Basha et al [15], have presented in their research work the impact of time and space complexity of such optimization algorithms and their performance by implementing different learning strategies towards finding out the optimal solution. Ishaani Priyadarshini et al [16], have studied some machine learning algorithms like random forest (RF), decision trees (DT), k-nearest Neighbors (k-NN), and deep learning algorithms such as convolutional neural networks (CNN), long short-term memory (LSTM), and gated recurrent units (GRU)) for Human Activity Recognition HAR. Shubhkirti Sharma et al [17], they have introduced multi-objective optimization algorithms and their variants with pros and cons; and representative algorithms in each category were discussed in depth. Amit Sagu et al [18], have presented two novel metaheuristic optimization algorithms for optimizing the weights of deep learning (DL) models, which have used deep learning to detect and prevent cyber-attacks of this nature. Xiangning Chen et al [19], have presented a method to formulate algorithm discovery as a program search and apply it to discover optimization algorithms for deep neural network training. Yandong Sh et al [20], have in their work introduced a systematic review of the most representative “learning to optimize” techniques in diverse domains of 6G wireless networks by identifying the inherent feature of the underlying optimization problem by ML frameworks from the perspective of optimization. B.Lavanya et al [21], have studied automatic genre classification and stand non-trivial for its invaluable applications by improving web search results, information retrieval, and/or the notable trends in this domain and its several stages.

## 6- Math Preliminaries

The Set theory preliminaries has one of the most important algebra rules by which the total function was the key to our proposed model for optimization selected [1]. A mathematical function known as the Total function method in the set theory can be used to increase the data science adaptation by using new methods to improve the overall system optimization for data

set selection in the preprocessing phase. A new proposed application for total function properties as stated in the set theory: a Total function F from X to Y is a binary relation on X × Y; that satisfies the following two properties:

1- For each x ϵ X → y ϵ Y ; such that [x , y] ϵ f ….(1)

2- If [x1 , y1] and [x2 , y2] ϵ f ; then y1 = y2….(2)

We have used the benefit of the total function properties in our proposed model suggested for the optimization model by transforming the use of total function simplification into information system need [3], [4]. The outcome of our new proposed model will be the novelty of our research work which was, in fact, more mathematical improvement for soft set theory applications approach as state of the art rather than as a simple proof of total function in the real time systems; we have used an example dealing with our assumptions such that: let X = (1,2,3,4,5,6) and Y = (a,b,c,d,e,f ), thus to represent the relation between X and Y in total function from x to y; will be represented in the table-2 as following:

**Table—2: the representation of total function in set theory**

| F | Y1 | Y2 | Y3 | Y4 | Y5 | Y6 |
|---|---|---|---|---|---|---|
| X1 | a | a | a | a | a | a |
| X2 | b | b | b | b | b | b |
| X3 | c | c | c | c | c | c |
| X4 | d | d | d | d | d | d |
| X5 | e | e | e | e | e | e |
| X6 | f | f | f | f | f | f |

### 6.1- The combination of Total Function and Soft Set theory improvement.

In our new proposed model, we have made a combination of the benefit of properties of total function in set theory and the Soft Set theory application in terms of information System need [1], [2], [3]. We took the advantages derived from both theories to get new methods to deal with the preprocessing for data science as well as to improve the performance of data science applications in the real world [5], [6].) An example can be used as follows: given U = {u1; u2; u3; u4; u5; u6} be a universal set consisting of a set of six pre-processing methods under consideration. Let A = {e1; e2; e3; e4; e5; e6}; be a set of parameters, where {e1: satisfy 100% as optimal solution}; {e2: satisfy 80% as very good}; {e3: satisfy 60% as good}; {e4: satisfy 40% as poor}; {e5: satisfy 20% as very weak}; {e6: satisfy 00% as null}. A soft set of (F; A) describes the "Preprocessing Methods" in which our proposed model by using a machine learning approach, it will offer the best optimization methods for the system overall. Suppose that we have the following case as it was given in an example-1;

(F; e1) = {u1, u2, u3, u4, u5, u6}; means that, the function soft set, with (e1) parameter must satisfy all 6-methods. (F; e2) = {u2, u3, u4, u5, u6}; means that, the function soft set with (e2) parameter must satisfy only 5-methods. (F; c3) = {u3, u4, u5, u6}; means that, the function soft set, with (e3) parameter, must satisfy only 4- methods. (F; d4) = {u4, u5, u6}; means that, the function soft set, with (e4) must satisfy only 3-methods. (F; e5) = {u5, u6}; means that, the function soft set, with (e5} parameter must satisfy only 2-methods. (F; e6) = {u6}; means that, the function soft set, with (e6) parameter must satisfy only one-methods. [6].

**Table-3: the binary table used to represent the soft set table.**

| U | e1 | e2 | e3 | e4 | e5 | e6 |
|---|---|---|---|---|---|---|
| u1 | 1 | 1 | 1 | 1 | 1 | 1 |
| u2 | 1 | 1 | 1 | 1 | 1 | 0 |
| u3 | 1 | 1 | 1 | 1 | 0 | 0 |
| u4 | 1 | 1 | 1 | 0 | 0 | 0 |
| u5 | 1 | 1 | 0 | 0 | 0 | 0 |
| u6 | 1 | 0 | 0 | 0 | 0 | 0 |

Table-3 represents the transformation of preprocessing methods under consideration of the proposed model which has made the performance of dataset preprocessing available to assets and to evaluate in such a calculated manner. Our model has the ability to optimize the whole dataset before and after; in order to store soft set in the computer easily.

### 6.2- Taxonomy for Preprocessing Methods

In our developed new method, we have proposed a taxonomy for optimization in data science preprocessing [8]. There are a lot of new issues which has been studied as state–of–the–art in our research work. We are interested in most new issues by optimization such as for instance, the developed optimized mathematical methods which has been used to pre-process the data set of our proposed model. This can be as new aspects for realization by studying the new computational concepts and new soft set theory applications for how to optimize the preprocessing phase in general to get the optimal performance in data science [7],[9].

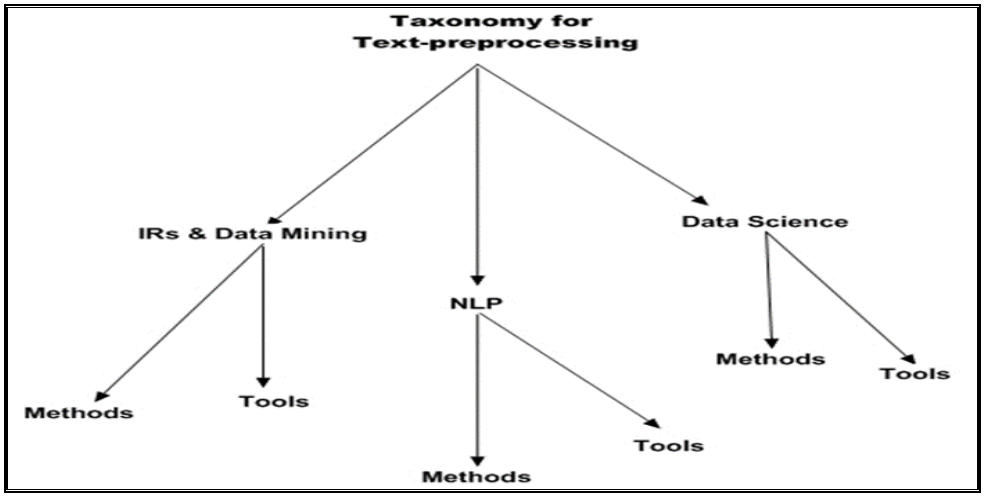


**Fig-1: Taxonomy for Preprocessing related to Data Science**

### 7- Our proposed Model

We have simplified the model design to be very useful to understand the idea behind our model which was developed to improve the performance of the preprocessing phase in data science and machine learning applications. The main goal is to get new methods used for optimization in the most important processes by dataset; and also will be the novelty of our research work which was in fact more mathematical improvement for soft set theory applications as state of the art. We have organized our methodology as shown in the figure-2 which illustrates the

main components of our proposed model [5], [6]. In general, the process of building any machine learning model, mostly; we need some iterative processes to get the work done within its process which involves a number of steps as shown in the figure-3. For instance, most data scientist need to select among specific methods and/or algorithms including; (Support Vector Machines, Neural Networks, Bayesian Models, Decision Trees) to make some extra tuning on parameters of the selected algorithm, the performance of the proposed model can also be judged by various metrics as well as by (accuracy, sensitivity, specificity, F1-score). [10], [14].

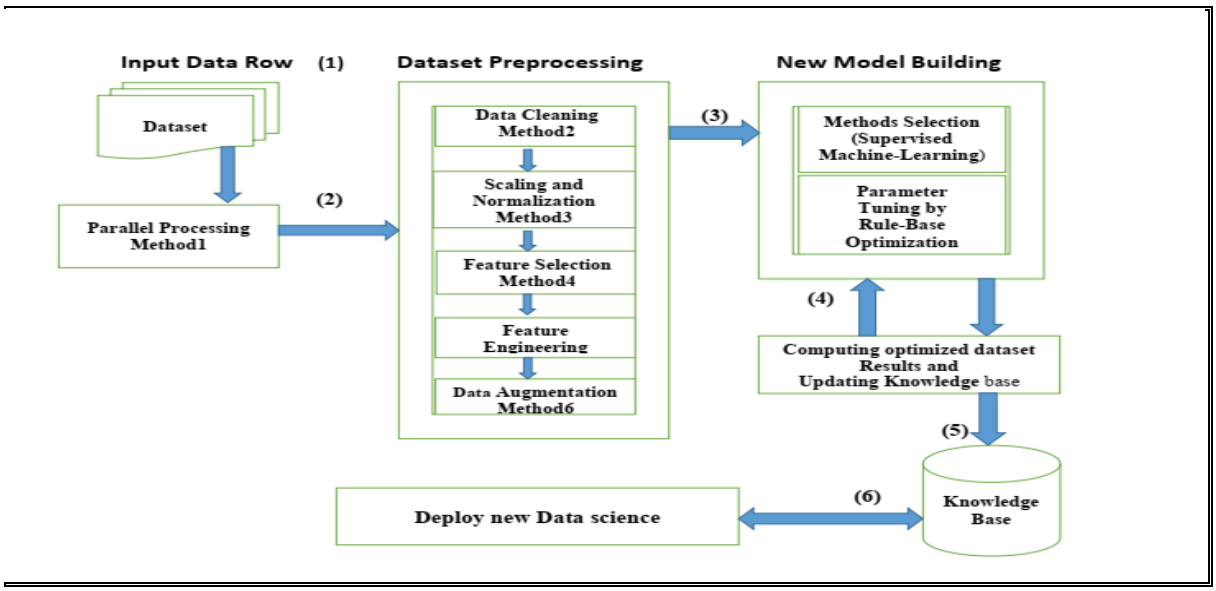


**Fig-2: Proposed model for the preprocessing phase in data science**

We have used in our research work, a machine learning model to evaluate the system performance in the preprocessing phase. We have used a training set by using a collected corpus of described Arabic dialects dataset. The corpus of the training dataset is composed of some dialects as shown in table-5, (Libya-1, Morocco-2, Egypt-3, Jordan-4, Palestine-5, and Sudan-6). As a matter of fact, we have found that our simple training model has provided good optimized results for our proposed model. We have randomly selected a small size of the corpus, which belongs to Arabic Text. We have also organized the dialects words manually in this phase, to investigate the model reliability.

### 7.1- Dataset

We have preprocessed a middle-size of Arabic Dialect dataset, which belongs to the Modern Standard Arabic Language. Our model was created upon the baseline of a developed model for the dataset based on a machine-learning approach [9], [10].

**Table-4. The corpus of training dataset is composed of some Arabic-dialect**

| No | Dataset | Country Codes | Text in Dialects -Total " Words" | Text in MSA -Total " Words" |
|---|---|---|---|---|
| 1 | Training | 6 | 1200 | 1200 |

| | | | | |
|---|---|---|---|---|
| **2** | **Test** | **6** | **600** | **600** |
| **3** | **Total** | **12** | **1800** | **1800** |

### 7.2- Rule-based table transformation

We have used a simple rules-based table to transform it into binary table that used to represent the soft set into another form to be as conditional rules based on the soft set theory manner. It is considered to be a more flexible and simple methods used for the preprocessing phase [15].

**Table-5. Rule-based into Transformation (binary table).**

| level | Rules-base selection | Parameters selection | m1 | m2 | m3 | m4 | m5 | m6 | 100% accuracy |
|---|---|---|---|---|---|---|---|---|---|
| **1** | **Rule-1** | **a1** | **1** | **1** | **1** | **1** | **1** | **1** | **90-100** |
| **2** | **Rule-2** | **b2** | **1** | **1** | **1** | **1** | **1** | **0** | **70-80** |
| **3** | **Rule-3** | **c 3** | **1** | **1** | **1** | **1** | **0** | **0** | **50-60** |
| **4** | **Rule-4** | **d4** | **1** | **1** | **1** | **0** | **0** | **0** | **30-40** |
| **5** | **Rule-5** | **e5** | **1** | **1** | **0** | **0** | **0** | **0** | **10-20** |
| **6** | **Rule-6** | **f6** | **1** | **0** | **0** | **0** | **0** | **0** | **00-10** |

For our model as a machine learning, we have used the Rule-base table-5 as illustrated, the accuracies of our different optimizers has calculated the optimization levels due to the parameters related to proposed optimization methods.

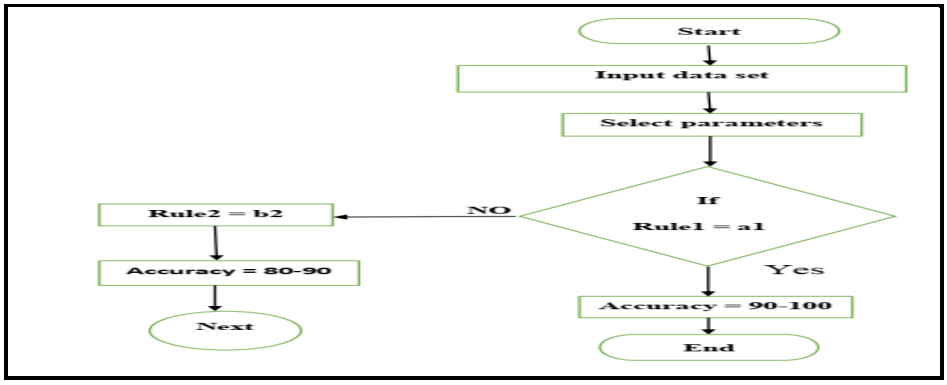


**Fig-3: The logical flow-chart for the dataset preprocessing evaluation**

### 7.3- Results Analysis

Our proposed model has performed the preprocessing methods that we have selected to be our optimizers for the proposed machine learning model [6], [9]. As shown in the result table-6 our dataset was a corpus of some different documents in Arabic text categorized into different subjects. In addition, we have trained the model using a simple rules based table to transform its rules into the binary table used to represent the soft set into another form to be conditional rules based on the soft set theory properties. It is considered to be the more flexible and simple method used for the preprocessing phase. The epoch is known as one of the iterative passes, where the entire dataset gets trained and for each [epoch parameters]; were updated and has improved the test accuracy. Table-6 shows all values, by which the proposed model has been used to train all the optimization methods.

**Table-6. Our proposed model results by different optimization methods on training data.**

| No | EPOCH | Method-1 Data Cleaning | Method-2 Scaling & Normalization | Method-3 Feature Selection | Method-4 Feature Engineering | Method-5 Data Augmentation | Method-6 Parallel Processing |
|---|---|---|---|---|---|---|---|
| 1 | 00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 |
| 2 | 20 | 0.893 | 0.923 | 0.881 | 0.883 | 0.876 | 0.832 |
| 3 | 40 | 0.899 | 0.926 | 0.871 | 0.920 | 0.894 | 0.836 |
| 4 | 60 | 0.901 | 0.944 | 0.912 | 0.927 | 0.900 | 0.921 |
| 5 | 80 | 0.924 | 0.968 | 0.913 | 0.936 | 0.922 | 0.951 |
| 6 | 100 | 0.941 | 0.944 | 0.955 | 0.957 | 0.958 | 0.961 |

Table-7 shows that; the benefit of our proposed methods gave a clear good sign on the 60th epoch, and has loss level almost constant. As we have trained the model from 0-100th epochs. Our model outcomes after accuracies tests on those optimized methods has shown that, the more improvement we have noted at every 20$^{th}$ epoch the more accuracies increasingly raised [13],[15].

**Table-7: Test accuracies by our propose model**

| No | Optimization Methods | Test accuracies in 100% |
|---|---|---|
| 1 | Data Cleaning | 0.941 |
| 2 | Scaling & Normalization | 0.944 |
| 3 | Feature Selection | 0.955 |
| 4 | Feature Engineering | 0.957 |
| 5 | Data Augmentation | 0.958 |
| 6 | Parallel Processing | 0.961 |

### 7.4- Conclusion and Future Scope

Optimization in data science related to high-performance systems which is totally dependent on machine learning techniques; and have become one of the most techniques used for having accurate and reliable information systems applications. The production pipeline of a machine learning model passes through different phases and stages that require a vast knowledge of several available tools, and algorithms. Big data produces daily huge data and has brought the need increasingly, and continuously at rapid digital lanes. Nowadays, AI is an essential domain to develop the methods used to scale up the preprocessing process. We have chosen these six optimized methods to experiment and observe how each method performs the trained data on a specific dataset. This research work has had a good impact on the research area through successive research and practice; many other methods can be developed in the future to improve the performance of the model. We have to mention that our research work still needs

to be extended to fulfill the second part of our research work in the next period of our research group activities.

**References**


1. Thomas A.Sudkamp, Languages and Machines, "An introduction to the Theory of Computer Science", eBook, 1997.
2. D. Molodtsov, "Soft set theory-first results, Computers Math", Applic, (1999), 19-31.
3. MAJI et al, "An Application of Soft Sets in a Decision Making Problem," PERGAMON-Computers and Mathematics with Applications", 2002.
4. Andreas et al, "Set optimization -a rather short introduction", arXiv: 1404.5928v2 [math.OC], 2 May 2014.
5. Q. Zhong and X. Wang, "A new parameter reduction method based on soft set theory", Vol. 9, No. 5 (2016), 99-108.
6. Nasef et al, "Soft Set Theory and Its Applications", https://www.researchgate.net/publication/326561107, July 2018.
7. Endert et al," The State of the Art in Integrating Machine Learning into Visual Analytics", 6arXiv:1802.07954v1 [stat.ML], 22 Feb 2018.
8. Kaiwen L et al," Evolutionary Many-Objective Optimization: A Comparative Study of the State-of-the-Art", June 5, 2018 Digital Object Identifier 10.1109/ACCESS.2018.2832181.
9. Radwa et al, "Automated Machine Learning: State-of-The-Art and Open Challenges", arXiv: 1906.02287v2 [cs.LG], 11 Jun 2019.
10. Ebubeogu et al, "Systematic literature review of preprocessing techniques for imbalanced data", doi/10.1049/iet-sen.2018.5193 October 2019.
11. Amir Ahmad, Shehroz S. Khan,"Survey of State-of-the-Art Mixed Data Clustering Algorithms", Digital Object Identifier 10.1109/ACCESS.2019.2903568.
12. Seba Susan et al," The balancing trick: Optimized sampling of imbalanced data sets, A brief survey of the recent State of the Art", DOI: 10.1002/eng2.12298, 7 September 2020
13. Dharma et al, "A Performance Comparison of Optimization Algorithms on a Generated Dataset", Chapter • January 2022, doi: 10.1007/978-981-16-3690-5_135.
14. Abdurakhmon Sadiev et al, "Federated Optimization Algorithms with Random Reshuffling and Gradient Compression", arXiv: 2206.07021v2 [cs.LG], 3 Nov 2022.
15. Syed Muzamil Basha et al, "A comprehensive Study on learning strategies of optimization algorithms and its applications", DOI: 10.1109/ICSSS54381.2022.9782200 ©2022, IEEE.
16. Ishaani Priyadarshini et al," Human activity recognition in cyber-physical systems using optimized machine learning techniques", doi.org/10.1007/s10586-022-03662-8,Springer Nature, 2022.
17. Shubhkirti Sharma et al," A Comprehensive Review on Multi-Objective Optimization Techniques: Past, Present, and Future", doi.org/10.1007/s11831-022-09778-9ne June, 2022
18. Amit Sagu et al, "Design of Metaheuristic Optimization Algorithms for Deep Learning Model for Secure IoT Environment", Sustainability, 2023, doi.org/10.3390/su15032204.
19. Xiangning Chen et al, "Symbolic Discovery of Optimization Algorithms", google, arXiv: 2302.06675v4 [cs.LG], 8 May 2023.

20. Yandong Shi et al, “Machine Learning for Large-Scale Optimization in 6G Wireless Networks”, IEEE, arXiv: 2301.03377v1 [eess.SP], 3 Jan 2023.
21. B.Lavanya et al, “Text Genre Classification: A Classified Study”, Eur. Chem. Bull, DOI: 10.31838 /ecb/ 2023.12.s1-B.383.