

# **Unsupervised Feature Selection Based on Rough Set Theory Using Direct Dependency Classes (DDC)**



Author

**Saliha Hanif**

**00000172268**

**Ms-16 (CSE)**

Supervisor

**Dr. Usman Qamar**

Co Supervisor

**Dr. Muhammad Summair Raza**

DEPARTMENT OF COMPUTER AND SOFTWARE ENGINEERING COLLEGE OF  
ELECTRICAL & MECHANICAL ENGINEERING  
NATIONAL UNIVERSITY OF SCIENCES AND TECHNOLOGY,  
ISLAMABAD

SEPTEMBER, 2020



In the name of Allah most beneficent most merciful

-My success, my inspiration, my guidance, my accomplishment, my reconciliation in my reform work, my welfare, my adjustment, my adaptation, my prosperity can come only and only through Allah, and none else

وَلَا يُحِيطُونَ بِشَيْءٍ مِّنْ عِلْمِهِ إِلَّا بِمَا شَاءَ

*And they can't encompass anything from His knowledge, but to extend He wills [2:255]*

# Unsupervised Feature Selection Based on Rough Set Theory Using Direct Dependency Classes (DDC)

## Author

Saliha Hanif

00000172268

MS-16(CSE)

A thesis submitted in partial fulfillment of the requirements for the degree of  
MS Computer Software Engineering

## Thesis Supervisor:

Dr. Usman Qamar

## Thesis Co Supervisor:

Dr. M. Summair Raza

Thesis Supervisor's Signature:

---

Thesis Co Supervisor's Signature:



---

DEPARTMENT OF COMPUTER AND SOFTWARE ENGINEERING COLLEGE OF  
ELECTRICAL & MECHANICAL ENGINEERING NATIONAL UNIVERSITY OF  
SCIENCES AND TECHNOLOGY,

ISLAMABAD

SEPTEMBER, 2020

# Contents

Declaration .....	i
Language Correctness Certificate.....	ii
Plagiarism Certificate(Turnitin Report).....	iii
Copyright Statement.....	iv
Acknowledgements.....	v
Abstract.....	vi
List of Figures .....	7
List of Tables .....	8
Chapter 1.....	9
1 Introduction .....	9
<b>1.1 Research Objectives .....</b>	<b>13</b>
<b>1.2 Research Contribution.....</b>	<b>13</b>
<b>1.3 Thesis Organization .....</b>	<b>14</b>
Chapter 2.....	15
2 Related Work.....	15
<b>2.1 Machine Learning .....</b>	<b>15</b>
2.1.1 Machine Learning importance.....	15
2.1.2 Machine Learning Applications .....	15
2.1.3 Effectiveness of machine learning.....	17
<b>2.2 Rough Set Theory.....</b>	<b>19</b>
<b>2.3 Rough Set Based Feature Selection Techniques .....</b>	<b>23</b>
2.3.1 Hybrid Feature Selection Algorithm Based On Particle Swarm Optimization (PSO).....	23
2.3.2 Genetic Algorithm .....	25
<b>2.4 USQR Algorithm .....</b>	<b>25</b>
<b>2.5 Selecting Features in Supervised Learning.....</b>	<b>28</b>
<b>2.6 Filter Techniques.....</b>	<b>29</b>
2.6.1 FOCUS .....	29
2.6.2 SCARP.....	30
<b>2.7 Wrapper method .....</b>	<b>30</b>
<b>2.8 Unsupervised Feature Selection.....</b>	<b>31</b>
2.8.1 Unsupervised Filters .....	31
2.8.2 Univariate filter methods .....	33
2.8.3 Multivariate Filter Method.....	34
2.8.4 Unsupervised Wrappers .....	34
2.8.5 Sequential techniques .....	36
2.8.6 Bio-propelled techniques .....	37

2.8.7	Hybrids .....	37
2.8.8	Fuzzy c-means Clustering .....	38
<b>2.9</b>	<b>Supervised Feature Selection Using DDC.....</b>	<b>38</b>
2.9.1	Quick Reduct PSO:.....	38
2.9.2	Genetic Algorithm: .....	39
2.9.3	Incremental Feature Selection Algorithm: .....	39
2.9.4	Fish Swarm Algorithm:.....	39
2.9.5	Rough Set Improved Harmony Search Quick Reduct .....	39
2.9.6	Tolerance Rough Set Firefly based Quick Reduct .....	39
2.9.7	Improve Quick Reduction for Function Selection.....	40
<b>2.10</b>	<b>Clustering algorithm using rough set theory for unsupervised feature selection.....</b>	<b>40</b>
Chapter	.....	42
3	Research methodology .....	42
<b>3.1</b>	<b>Problem statement.....</b>	<b>42</b>
<b>3.2</b>	<b>Dataset collection.....</b>	<b>44</b>
<b>3.3</b>	<b>Preprocessing.....</b>	<b>44</b>
<b>3.4</b>	<b>Direct Dependency Calculation (DDC).....</b>	<b>44</b>
<b>3.5</b>	<b>Feature selection.....</b>	<b>51</b>
<b>3.6</b>	<b>Implementation Environment.....</b>	<b>51</b>
3.6.1	Tools and programming languages used:.....	51
<b>3.7</b>	<b>DDC Algorithm .....</b>	<b>54</b>
Chapter 4.....	.....	55
4	Dataset, Implementation and outcome .....	55
<b>4.1</b>	<b>Dataset Collection .....</b>	<b>55</b>
<b>4.2</b>	<b>Feature Selection Accuracy and Execution Time .....</b>	<b>55</b>
<b>4.3</b>	<b>Comparison with other algorithms .....</b>	<b>58</b>
<b>4.4</b>	<b>Time Complexity using Big O notation .....</b>	<b>62</b>
<b>Chapter 5</b> .....	.....	<b>64</b>
5	Conclusion and Future Work.....	64
<b>5.1</b>	<b>Conclusion .....</b>	<b>64</b>
<b>5.2</b>	<b>Future work.....</b>	<b>66</b>

## List of Figures

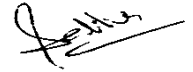
Figure 2.1 : Approximation Diagram.....	200
Figure 2.2: PSO-QR Algorithm.....	24
Figure 3.1: Feature Selection from unsupervised dataset using DDC .....	43
Figure 4.1: Classification Accuracy for Unsupervised Feature Selection .....	556
Figure 4.2: Execution time of DDC Algorithm with parallel and without parallel processing .....	57
Figure 4.3: Accuracy Comparision.....	60

## List of Tables

Table 2.1: Decision System Example .....	26
Table 3.1: Calculates dependency using DDC .....	44
Table 3.2: Unsupervised dataset .....	45
Table 4.1: Set of Datasets used.....	555
Table 4.2: DDC Algorithm classification accuracy.....	556
Table 4.3: DDC algorithm execution time with and without parallel processing .....	57
Table 4.4 : Comparison with other algorithm .....	58
Table 4.5 : Accuracy Comparison with other classifier .....	6059
Table 4.6: Execution Time of different techniques on different datasets .....	6061
Table 4.7 : Time Complexity using Big O notation .....	63

## Declaration

I certify that this research work titled — “Unsupervised Feature Selection Based on Rough Set Theory Using Direct Dependency Classes (DDC)” is my own work. The work has not been presented elsewhere for assessment. The material that has been used from other sources is properly acknowledged / referred.



Signature of Student

Saliha Hanif

00000172268

MS-16 (CSE)



## Language Correctness Certificate

This thesis has been read by an English expert and is free of typing, syntax, semantic, grammatical and spelling mistakes. Thesis is also according to the format given by the university.



Signature of Student

Saliha Hanif

00000172268

MS-16 (CSE)

Signature of Supervisor

Dr. Usman Qamar



Signature of Co Supervisor

Dr. Muhammad Summair Raza

## Plagiarism Certificate (Turnitin Report)

This thesis has been checked for Plagiarism. Turnitin report endorsed by Supervisor is attached.



Signature of Student

Saliha Hanif

Registration Number

000001712268

Ms-16 (CSE)

Signature of Supervisor

Dr. Usman Qamar



Signature of Co Supervisor

Dr. M. Summair Raza

## **Copyright Statement**

- Copyright in text of this thesis rests with the student author. Copies (by any process) either in full, or of extracts, may be made only in accordance with instructions given by the author and lodged in the Library of NUST College of E&ME. Details may be obtained by the Librarian. This page must form part of any such copies made. Further copies (by any process) may not be made without the permission (in writing) of the author.
- The ownership of any intellectual property rights which may be described in this thesis is vested in NUST College of E&ME, subject to any prior agreement to the contrary, and may not be made available for use by third parties without the written permission of the College of E&ME, which will prescribe the terms and conditions of any such agreement.
- Further information on the conditions under which disclosures and exploitation may take place is available from the Library of NUST College of E&ME, Rawalpindi.

This page is intentionally left blank

## **Acknowledgements**

This is by the Grace of my Lord to test me whether I am grateful or ungrateful! And whoever is grateful, truly, his gratitude is for (the good of) his own self, and whoever is ungrateful, (he is ungrateful only for the loss of his own self). Certainly! My Lord is Rich (Free of all wants), Bountifull [An-Naml: 40]

I am indebted to NUST College of Electrical and Mechanical Engineering, for providing me an opportunity for Masters Research. First and foremost I offer my sincerest gratitude to my thesis supervisor, Dr. Usman Qamar, who has supported me throughout my thesis, with his patience and knowledge.

I would also like to thank Dr. Sumair, Brig Dr. M. Abbas and Dr. Wasi Haider Butt for being on my thesis guidance and examination committee and for guidance and cordial support, which helped me in completing this task through various stages.

Finally, I would like to thank my family, friends that have been a constant source of love, support and inspiration they provided me throughout my entire education and life. I would also like to give my special thanks to my beloved cousin Iqra Basharat for her support and prayers throughout my life.

Most importantly, I would like to thank my father, mother, husband and mother in law who always motivated for work and pushed me up by saying yes you can do that because of you today I am here thanks a lot my mentor. I would also want to extend my appreciation to those who could not be mentioned here but well played their role to inspire the curtain.

Last but not least I would like to give my special thanks to my friend Sana Abid for helping me out throughout my thesis. Without your help and guidelines I was not able to accomplish that work.

## **Dedication**

*I am dedicating my thesis to my beloved parents, M.Hanif and Zaib un Nisa, who has always been a positive role model in my life. In the face of adversity and many challenges along the way, they always pushed through and did what they had to do to make things work. My father has been a lifelong example of hard work and perseverance. He taught me the importance of hard work, honesty and integrity. My mother is a strong lady, beautiful person inside and out and I can only hope to be the kind of person she is. I would also like to dedicate my Aunt Sheraz Begum who taught me being humble in every situation and never giving up. I could never have done this without your faith, support and constant encouragement. Thank you for teaching me to believe in myself, in Allah and in my dreams. Thank*

## Abstract

The growth of data is considered to be at exponential rate in today's advanced technological world where data is all around us in different forms. With the increase in number of records in a dataset there is increase seen in horizontal dimension also i.e. the no attributes are also being added up. When we perform data analysis or decision making activity, these attributes plays an important role. Sometimes, more the number of attributes vaguer the results are! Therefore, we have to select those attributes which contribute more in producing better results and leave those which are irrelevant. That is what we call dimensionality reduction or feature selection.

Based on rough-set approach, direct dependency calculation algorithm is the primary procedure that reduces the no of attributes while preserving the key information. In this research study ,a direct dependency class calculation algorithm on unsupervised dataset is proposed, which has not been done before. The main goal is to extract useful features, reduce the code complexity and execution time while calculating dependencies of attributes on each other in a given dataset. This technique successfully performs feature selection by using two set of rules of direct dependency calculation. To verify the reduced execution time and algorithm complexity we carried out the experiment on standard datasets take from the UCI library. The results show great improvement in terms of feature selection accuracy and execution time with parallel processing. UDDC provides the accuracy above 95% when compared with other feature selection algorithms.

**Keywords:** Direct Dependency Class (DDC), rough set theory, dependency rules, positive region, feature selection.

# Chapter 1

## 1 Introduction

In various fields that involves data, especially huge datasets, feature selection is considered to be of sheer importance. In the procedure of feature selection, a subset of features is extracted that have massive contribution in the problem domain and problem analysis [1]. This process eradicates the noisy, redundant and misleading features that considerably distress the data analysis and causes the biasness in the model classification [2] [3].

Considerable range of methods proposed for focus on the issues of insignificant attributes that does not contribute in efficient and accurate results. Several methods are available to perform feature selection, some of them includes; evaluation of all potential feature subsets and their practicability from the unified dataset. For this exhaustive search would be required and that is not favorable in case of larger values of  $n$  [4] . Therefore, random search can be used in substitution of exhaustive search that supports the random generation of candidate feature selection. Another frequently used approach for feature selection is known as heuristic approach [5]. Filter and wrapper approaches are used to resolve the difficulty of feature selection that uses preprocessing steps and optimal feature selection respectively [6]

The goal of feature selection is to eliminate obsolete features typically referred to as irrelevant features and unnecessary features [7][8] [outliers have little effect on the target description, while large - scale do not add any new information to the target concept but have a detrimental impact on classification efficiency and processing time [4]. An informative feature is one that has a high correlation with the concept(s) of the decision but is highly uncorrelated with other functions. Similarly, if it is highly relevant and non-redundant a feature subset is considered useful.

Selection of features [9] and reduction of generation [10][11] are the pre-processing techniques used in data mining [10][11][12] to discover knowledge from data stored. Optimum subsets of features are chosen according to a certain evaluation criteria



It is an exciting research area that has been revealed as being very successful in removing obsolete and redundant traits. Parameters such as its relevance and redundancy assess the importance of a feature subset. Most of the decision function is predictive of a task associated with it; otherwise it is deemed irrelevant for apps. A functionality is unnecessary when closely associated with other functions.

We have implemented a direct approach to the calculation of dependence based on rough-set theory. RST is systematic numerical technique, well applicable for selection of features and can also be used to identify objects. Rough-set theory is considered to be capable of minimizing dimensionality, while supplying the full detail. Direct dependency calculation algorithm is the primary procedure based on rough-set approach which reduces attribute no while preserving the key information. In this research report, we have extended direct reliance groups on miscellaneous data sets for unsupervised collection of features and have obtained better outcomes. Collection of features is the collection method for a collection of individual dataset features to display the selected subset on behalf of the entire data. Choosing subgroups of features thus enables to lessen sets of data by removing unnecessary and redundant information to a manageable scale.

Over the last 20 years, the dimensionalities of the knowledge sets used in computer learning and data processing applications has explosively increased. For such a rise in data dimensionality, there are two main approaches: attribute reduction, and selection of features. Attribute reduction transforms underlying semance of results, because the name implies. Selection driven reduction, i.e. selection of features, chooses the parts to characterize the information as an alternative of converting the primary semics. Consequently, the fundamental semance is retained. Type picked in these domains support attenuate the dimensionalities of the thing gap, increase classification algorithms' analytical performance, and increases the representation and interpretation of the thoughts induced here. Choice of features involves only not a discount in dimensionality, i.e. a discount within the attributes number ought to remember while building a pattern, attribute may be chosen or useless on the idea of criterion that specify the utility. In the real world data consist of information more than the actual required, Thus, the choice of features develop into a compulsory stride

in the direction to create the study convenient and to find out output [13]. It is necessary within the study of high-profile results [14], various uses function, like dropping dimensional data set, reducing the instant interval needed per registration and reclassification a classifier's sorting precision through eliminating unnecessary with inaccurate not to avoid errors [15].

In the literature , different selection methods for features were suggested. These include selection of correlation-based features [16,17], shared function collection dependent on details [18,19], varied range of apps [20], selection of features based on consistency [21], chart theoretical technique [22], function collection based o ACO [23], possible modeling [24], and SVM-based feature selection [25].

Pawlak 's suggested Rough Set Theory (RST) [26,27] is a statistical method for information processing. Reduction of RST based attributes methods [28–31] and function selection [32–36] were prevalent. RST provides a positive dependence measure for the region for selecting a feature, called "attribute dependence." Dependence on attributes determine the worth of a dependent element is calculated by the particular The attribute importance. A Dependent 's value variable varies from 0 to 1, 0 defines one variable is not dependent on another, or one (1) means that one attribute is dependent entirely on another. However, positive area is used to find out dependency by this technique, that is a long and compound stride that negatively affects the act of selecting functionality and makes choice practically impossible. works when datasets exceed size.

Applications for machine learning are about extracting information from data. Inherently, data sets obtained from realworld applications are prone to contain both vague and incomplete data. In evaluating vagueness can be found in the subjective definitions such as beautiful, moist, intelligent, similar and so on. Insufficient data can be found when discerning between the There is no adequate feature set describing data samples, i.e. the dataset contains elements with the same values for all features but different values for the related result A typical example applies to patients with the same symptoms but with different diseases. It means that no definitive diagnosis can be made based on the

characteristics (symptoms) at hand.

Such main source of ambiguity is integrated, respectively, into the fuzzy set theory A membership value, carried from the unit interval, communicates to a given group the degree to which elements belong to a category rather than a crisp yes or no membership of elements. Rough sets treat missing knowledge, on the other hand. You are On the basis of the assumption that ideas are not necessarily modelable

In the quest for unknown information, the hybridization of these two models into fuzzy rough set theory was first suggested in[43], and has since been widely used and extended. It involves approximating a vague and incomplete term by means of two blurry sets, lower and upper blurred rough approximations. This allows elements even to be discernible to others degree from one another Extension, and not discernible or not. Across a wide range of machine learning domains, the exercise of fuzzy rough set theory is met.

The most prominent emphasis in the text has been on algorithms for characteristic variety, which reduces the amount of attribute that identify the elements in a dataset to achieve a speed-up and potential act gain from later knowledge algorithms.

Fuzzy rough sets used to shape strength of each attribute and direct the quest for an optimal subset of attributes. The history of those methods and their creation are checked, taking into account different quality measurements of features and search approaches. In addition to selecting attributes, we further study the orthogonal instance selection system, pre-processing data sets by deleting instances rather than functions.

The proposed algorithms, in this case use the Fuzzy rough set theory notions to determine the usefulness of every element. Preprocessing methods are mutually attribute selection and instance selection. In our analysis we also consider learning algorithms themselves, building models based on the data at hand. They assess controlled, unmonitored, and semi-supervised fields. Our former definition includes the classification and regression approaches, as well as neural networks. We consider Self-organizing maps and clustering algorithms to the unsupervised

domain. Finally, we study a Fuzzy Rough Sets program Semi-supervised self-training.

## **1.1 Research Objectives**

There are different methods to choosing functionality, one of them is rough-set theory, which offers a basis for minimizing dimensionality while retaining data semics and producing more precise outcomes. There is however a complexity of code or the complex operations of rough-set theory and complexity of time that makes it difficult. Hence the key aim of this research study is:

- Minimize code complexity and improve the time. The rough-set theory-based function selection algorithm is efficient.
- Selection of unsupervised data set results obtained after execution of the algorithms on different datasets showed that our proposed method yields satisfactory ends up in conditions of amount of features selected, calculation instance and categorization accuracy of a variety of classifiers.

## **1.2 Research Contribution**

Rough-set theory is considered to be capable of minimizing dimensionality, while supplying the full detail. Selection of unsupervised feature is measured a a great deal issue because of the difficulties in determining trait relevance. Rough-set theory feature selection strategies use two primary methods; heuristic approach and conventional methods use constructive region-dependence measures. Heuristic is considered to be more efficient out of these two.

Based on rough-set approach, direct dependence calculation algorithm is the primary technique that reduces the no of attributes while retaining the key information. For unsupervised feature selection, in this research study, we have applied direct dependency classes on miscellaneous data sets and achieved better results.

In our research we have proposed unsupervised direct dependency class calculation feature selection algorithm that works for unsupervised dataset. Through this approach we do not have

to calculate the positive region calculations as that is the time consuming effort instead we use direct dependency calculation to calculate the dependency measures. This helps in examining straight away the number of unique classes. This direct method is suitable for larger datasets as compare to positive region.

### **1.3 Thesis Organization**

The whole thesis document is divided into five chapters.

**Chapter 1: Introduction.** This chapter submitted a description of rough set theory and the collection of features by means of rough set theory. Calculation strategies for the direct and indirect dependence groups are also included in this chapter. Part of this chapter is also a target for study

**Chapter 2:** The Study of Literature deals with the work pertaining to selection of features using rough set theory approaches.

**Chapter 3: Research Methodology.** This chapter briefly describes our research methodology that what methods we used to address our research problem. It also describes our research design along with system flow. It also elaborates on our proposed conceptual framework including the concepts and components associated with it.

**Chapter 4. Dataset, Implementation and Results.** This chapter revolves around validating the suggested conceptual structure for selecting features using rough set theory approach. Our analytical structure is checked and it explains findings.

**Chapter 5. Conclusion and Work Forward.** provides the overall review and concludes the thesis along with future study.

## **Chapter 2**

### **2 Related Work**

#### **2.1 Machine Learning**

Machine learning systems auto-learn computer Applications. Designing them manually is often especially appealing to alternate, machine learning has spread rapidly across the computer science and beyond over the last decade.

In Internet search, spam filtering, recommendation systems, , credit rating, deception detection, stock trading, product plan, machine learning is used.

##### **2.1.1 Machine Learning importance**

Most companies using the large data came to know the importance of machine learning. Enterprises can work more efficiently and thus become more competitive advantage by gleaning secret insights from the data. In addition, reasonable and good processing and low cost enabled the development of models that analyze huge chunks of complex data quickly and accurately. In addition to enabling organizations to recognise patterns and patters from a variety of data sets, ML also helps businesses to automate analyzes that have historically been performed by human. The companies can deliver customized product ad service which differentiate specifically respond to changeable customer requests. However, ML also allows businesses to pursue prospects that could be profitable in the long run.

##### **2.1.2 Machine Learning Applications**

Companies across several industries have recognized the importance of machine learning technology which handles vast data amount. By using the insight gained from the data, companies can work efficiently in managing costs and maintain a lead over their rivals. It is how other fields / domains apply learning machines-

### **2.1.2.1 Financial Services**

Financial-sector businesses are able, with the aid of machine learning technology, to recognise key trends in financial data and avoid any incidence of financial fraud. The platform also helps to recognize investment and trade opportunities. Using cyber surveillance helps identify certain persons that are vulnerable to economic risk, that get the appropriate measures to avoid fraud in good time.

### **2.1.2.2 Sales and Promotion**

The industries use machine learning technology to evaluate their customers ' purchasing patterns. The future of sales and marketing is the skill to track, evaluate and use consumer data to have a modified skill.

### **2.1.2.3 Government**

Public departments such as safety and other also need machine learning, as they have numerous data source that to be analyzed to recognize valuable trends and observations. Sensor data , for example, to analyse to find behavior to cut expenses and improve performance. However, it avoid data breaches and discourage fraud.

### **2.1.2.4 Healthcare**

ML is becoming a fast-growing phenomenon Wearable devices and software are implemented in healthcare and use data to monitor patient well-being in real time. Wearable apps offer medical information in real time, respiration, high bp, critical parameter. Physicians and medical professionals may use this knowledge to evaluate an individual's health status, derive a trend from the patient's past and anticipate any possible diseases. In addition , the skill empowers health expert to evaluate data and determine patterns that enable improved diagnosis care.

### **2.1.2.5 Transportation**

Regarding the traffic history and travel pattern ,problems can be highlighted by machine learning. People can choose longer path. In this field machine learning teciques are used to analyze and process data. People can make better choice through it.

### **2.1.3 Effectiveness of machine learning**

Although controlled and unsupervised learning by companies today are two of the most commonly recognized forms of machine learning, numerous other techniques available. Below is a list of few of the ML method used most frequently.

#### **2.1.3.1 Supervised Learning**

It include a position of contribution of instructions with the exact tests needed. The algorithm then equates the real result to the expected outcome and, if any discrepancy exists, marks a error. Utilizing various method such as regression, classification, gradient enhancement and estimation, supervised learning use various patterns to proactively predict extra mark values unmarked information. The above approach is widely used in environments where historical data are used to forecast likely events.

#### **2.1.3.2 Unsupervised Learning**

The ML approach finds its use in areas where data does not have any past label. The machine will not be given the "correct answer," and the algorithm will classify what is displayed. Primary want is to find a way ,plan for the data available. Transactional data functions as a strong source of data collection for unsupervised learning. Likewise, Features that differentiate consumer segments from each other may also be recognized. It's a matter of finding a particular structure in the available data collection, either way. Additionally, outliers can be solved in the given data. Among the commonly used unsupervised methods of knowledge are-

- i. Clustered k-means
- ii. Autonomous maps
- iii. The decomposition of values
- iv. Map of nearest neighbor

While all of these methodologies have a common aim to extract observations, pattern, technologies to have known decision, they have different techniques. A few can be seen



#### **2.1.3.2.1 Data Mining**

This approach is a generalized form of various approaches to extract valuable insights from the available data, which may include machine learning and conventional statistical methods. It is mainly used to discover certain patterns that were previously unknown in a data set.

This methodology covers machine learning, mathematical algorithms, study Regression analysis, text classification, and other areas of computation. Furthermore, data mining includes data manipulation, storage studies and practices.

#### **2.1.3.2.2 Machine Learning**

The key goal of machine learning on the market to establish, better recognize the secret data structure and patterns. However, to achieve a deeper understanding, statistical distributions are applied to the data sets.

Almost every prediction test is backed by theories which are mathematically proven. Nevertheless, machine learning is largely dependent on the capacity on the ability of computers to dig deeper into the data available to unleash a structure, even in the lack of a principle of what the file system is.

The techniques are assessed with a mistake of validation, as opposed to passing a academic test that supports a unacceptable suggestion. Because machine learning is iterative in nature, it is easy to automate the learning process in terms of data learning, and the data is analyzed until a consistent pattern is discovered. Machine learning models are tested on new data sets using a validation error, as opposed to Theoretical test which establishes the hypothesis of nullity. Since machine learning is iterative in nature, the learning process can be simply programmed with respect to data learning; and the data will be analyzed in anticipation of a consistent guide is establish.

### **2.1.3.2.3 Deep Learning**

The method is to classify terms contained by sound and artifacts within images, through the ability to combine computational power and special neural networks in vast amounts of data to learn complex patterns. Many scholars are trying to replicate the success in identifying patterns to solve difficult problem such as health finding, industry challenge, verbal communication transformation etc. Such a crucial technical innovation, machine learning is embraced by numerous businesses across the globe.

Several researchers are trying to replicate the success Through identifying patterns, more complex activities such as medical diagnosis, market issues, language translation and other social problems are resolved.

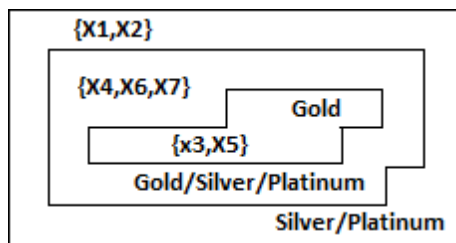
## **2.2 Rough Set Theory**

It is considered to be capable of minimizing dimensionality, while supplying the full detail. Owing to the difficulties of determining function significance, unsupervised feature collection is considered a much harder problem [37]. Rough-set theory feature selection techniques use two main approaches; heuristic approach based and traditional approaches use positive measurements of region dependence. Heuristic is considered to be more efficient out of these two, positive region is a costly approach which make it unsuitable on the way to exercise for large datasets. Direct dependency calculation algorithm is the primary procedure based on rough-set approach which reduces attribute no while preserving the key information. In this research report, we have extended direct reliance groups on miscellaneous data sets for unsupervised collection of features and have obtained better outcomes. In our research we suggested an unsupervised calculation of direct dependency type function selection algorithm that works for unsupervised dataset. Through this approach we don't have to calculate the positive region calculations as that is the time-consuming effort instead we use the calculation of direct dependency to calculate the dependency measures. This allows the number of distinct groups to be tested right away. This direct approach is ideal when compared with positive area for larger datasets. Over the last decade, rough set is interesting topic that is successfully applied by researchers to many

domains. A smaller collection of attributes (called reduct) that includes much of the information that be used for a given dataset. Thus attributes other than reduct set with minimal loss of information can be removed from the dataset.

Pawlak proposed RST on the discovery of knowledge in datasets[7,66]. Unlike standard discrete sets, RST, as discussed below, is base on the principles of upper and lower approximations. There may be redundant attributes in a dataset which can be eliminated without much of the essential loss of information. Rough sets [7,66] allow one to identify high and weak levels of significance in order to exclude redundant attributes. In RST the reduct principle is fundamental.

Being a subset of attributes, it will differentiate all the objects in a dataset that are discernible in comparison to the entire set of attributes. Another essential notion of center is in RST. A core is common set of attributes in all a dataset's reducts. Reduction as well as core are important concepts used in selection of features and reduction of dimensionality. The following section discusses reductions and cores in more detail.



*Figure 2.1: Approximation Diagram*

Rough set theory former suggested a theory for data analysis to address uncertain knowledge[27][66]. The classical theory of rough set is based on uniformity relationships. It is formed in same classes created by the relations of equality. The minor and higher approximations are created by adding the granules of information for attribute reduction.

However, the functions for separate data. The unprocessed information have to be discreted foremost for continuous information. Discretization of data leads to great loss of information.

Consequently the discrete information donot tell the actual result. Thus, the current rough set theory was comprehensive as of different perspectives [67], rough set neighborhood models [68], rough set models [69,70], etc. The rough set model of the neighborhood was implemented for dealing with numerical attribute reduction. The degree of dependence was established on the basis of the neighbor-hood relationships to determine the importance of attribute. And characteristic lessening algorithms were developed. In [71] two-form taxonomy method is proposed that at first classified the data using the lower approximation and then classified the non-classified data using the rough membership functions from the upper approximation collection at the first stage.

In [72] neighborhood k-step models explored as extensions to Pawlak model [73] distinct the positive region as a set of samples that can be classified without uncertainty, built the degree of dependence as the positive region's cardinality ratio to the sample space, and applied their degree of dependence to reduce various attributes. In other words, numerical and categorical attributes[74] explored a multi-district granularity model and proposed a method for selecting the appropriate granularity by optimizing margin distribution.. [75] developed an efficient rough range neighborhood model and developed a cost-sensitive feature selection backtracking algorithm based on a trade-off between the cost of research and the risk of misclassification. [76] The gene selection method was based on a rough set of neighborhoods and entropy measures to address uncertainty and noise in the gene data.

Machine learning can present a data mining problem. From a set of functions (attributes) that define the measurable assets gathered during a phase. Such attributes of enable the extraction of information by evaluating the classification, regression, and pattern clustering. The bulk of the real globe troubles have great dimensional characteristics, but only a small proportions of these are significant or appropriate features to outline the operation. Some algorithms for learning a machine high deficiencies present when the collection of features is substantial good, such as decreasing precision, computational increase Belasting, and biasing, among others. These issues fall under the curse of the consequences of dimensionality . The curse of dimensionality was dealt with according to purpose Selection methods which can most discriminate features of the process which use different methods. It shown where the authors

define the functions set in the modules Filter, Cover, and Embedded. Search approaches are based on ranking techniques which provide a score the deletion of irrelevant features is set for each function and threshold; correlation and reciprocal analysis of information belong to that group. The wrapper approach looks for relationships among the characteristics and function to preserve the feature actually influence the purpose job. Methods of choosing features are often split into two key groups, such as those regulated and unsupervised. The samples are classified in the Game Supervised based techniques, and the discrimination of main features is fairly significant Fast. Using supervised techniques, the wrapper approach and other embedded approaches could be called unregulated.

We are especially interested in unmonitored function select algorithms. An unsupervised approach is a clustering of attributes, which attempts to collect alike attribute in cluster; only one feature is defined at the end of the process as a symbol for each cluster, this method allows the characteristic space to be every. A summary of the job ,a Genetic Algorithm on attribute clustering, using two Metrics as health feature ensuring the right composition for such clusters. An unsupervised functional algorithm Selection based on a clustering based on local learning is proposed.

Writers use clustering of swarm intelligence and consensus as the Choosing feature. Some methods combine clustering algorithms by diverse theory to get feature selection devoid of a priori informing about the objective function, but progress is still limited in this field. Smart in studying theory and information technology applications the transmission of information is a hot matter. Owing to Computer Development Science, in particular the creation of a computer network, provide people with a large amount of knowledge. With the that amount of information, the information analysis tools requirement is also getting higher and higher, and people are hoping to automatically gain the possible knowledge from the data. In the field of artificial intelligence in particular, information discovery (rule extraction, data mining , machine learning, etc.) has attracted considerable attention over the past 30 years. The rough set theory, introduced by Professor Pawlak in 1982, is an important mathematical method for treating imprecise, contradictory, incomplete knowledge and information. The basic concept of the rough set theory can be divided into two parts, derived from the simple knowledge model. The first step consists

of creating definitions and laws by classifying relational databases. Section two is the exploration of insight by classifying the equivalence relationship and classifying it For Objective approximation. The rough set theory, as a theory of data analysis and processing, is a new mathematical tool for dealing with uncertain information following probability theory, fuzzy set theory, and evidence theory.

However, membership function selection is unclear. Hence the fuzzy set theory is, in a way, an ambiguous mathematical method for solving the ambiguous problems. Two specific boundary lines are defined for representing the imprecise concepts in rough set theory. The rough set theory is thus, in a sense, a certain mathematical tool for solving the unsure Discussion.

The rough collection because of creative thought, special process and simple operation theory has been an important method for handling information in the area of intelligent Data processing information[2-3]. And it has been widely used in machine learning , knowledge discovery, data mining, decision support, analysis and so on. The first International Conference on rough set theory was organized by Poland in 1992.

## **2.3 Rough Set Based Feature Selection Techniques**

Rough set theory was effectively used for the techniques of selecting apps. The underlying concepts provided by RST help by eliminating the redundant ones to find representative features. We will now be presenting numerous selection techniques of apps using RST concepts.

### **2.3.1 Hybrid Feature Selection Algorithm Based On Particle Swarm Optimization (PSO)**

Hanna et al . presented a supervised Particle Swarm Optimization ( PSO) and RST selection algorithm for the hybrid features. Algorithm computes reductions without generating all possible subsets exhaustively. The algorithm begins with an empty set, and adds one by one attributes. It creates a population of particles in S dimensions with a random position and velocity. In space for problem. This then uses RST-based dependency measure to determine fitness function of each particle. The highest-dependency feature is selected and the combination of all other features is constructed with this one. Every of these combinations is selected for its fitness.

When it is better than gbest then the gbest position is set to current the position of the current particle with the best fitness improved worldwide. This location represents the best subset of features found up to now, and is stored in R. The algorithm then updates the velocity of each particle and its position. This persists until conditions to avoid are met and in typical cases is the maximum number of iterations. The dependence of each subset of attributes is calculated based on the dependence on the decision attribute and the best particle is selected according to the algorithm. Algorithm uses positive region based dependency measure and is QuickReduct algorithm enhancement.

```

Input: C, the set of all conditional features;
          D, the set of decision features.
Output: Reduct R
Step 1: Initialize X with random position and Vi with random velocity
    ∀: Xi ← random Position();
    Vi ← random Velocity();
    Fit ← 0; globalbest ← Fit;
    Gbest ← Xi; Pbest(1) ← Xi
    For i = 1..S
        pbest(i) = Xi
        Fitness (i) = 0
    End For
Step 2: While Fit ≠ 1 //Stopping Criterion
    For i = 1..S //for each particle
        ∀: Xi;
        // Compute fitness of feature subset of Xi
        R ← Feature subset of Xi (I's of Xi)
        ∀x ∈ (C-R)
         $\gamma_{R \cup \{x\}}(D) = \frac{|POS_{R \cup \{x\}}(D)|}{|U|}$ 
        Fit =  $\gamma_{R \cup \{x\}}(D)$   ∀x ∈ R,  $\gamma_x(D) \neq \gamma_C(D)$ 
    End For
Step 3: Compute best fitness
    For i = 1:S
        if (Fitness(i) > globalbest) // if current fitness is greater than
            global best fitness
            globalbest ← Fitness(i); //assign current fitness value as
            global best fitness
            gbest ← Xi;
            getReduct(Xi)
            Exit
        End if
    End For
    UpdateVelocity(); //Update Velocity Vi's of Xi's
    UpdatePosition(); //Update position of Xi's
    //Continue with the next iteration
    End {while}
Output Reduct R

```

Figure 2.2: PSO-QR Algorithm

### 2.3.2 Genetic Algorithm

A rough set-based genetic algorithm (GA) for selecting features is provided in [24] authors. The selected set of features had been given to the artificial neural network classifier for further study. The method employs optimistic region-based measure of dependence as fitness in the proposed framework for developed candidates.

## 2.4 USQR Algorithm

For data mining implementations the judgment type identifiers are often unclear or missing. In this situation, the uncontrolled selection of features plays a vital role in selecting features. There are two input parameters of the current supervised QR algorithm: uncertain characteristic and decision attribute, and its determination of the degree of dependency value contribute to the decision attribute. But there is only one input parameter in the proposed USQR which is a conditional attribute.

Here the calculation of the degree of dependency value for the subset of a function refers to each conditional attribute and calculates the mean of the dependency values for all conditional attributes.

The USQR algorithm attempts to determine a reduct by creating All feasible subsets are exhaustive. This begins from an empty set and integrates certain attributes in turn, one at a time, which leads to the greatest rise in the rough set dependence metric before the highest possible value is generated for the dataset.

For each subset of attributes the mean dependency is calculated and the better alternative is selected according to the algorithm:

$$\gamma_p(a) = \frac{|POS_p(a)|}{|U|}, \forall a \in A \quad (2.6)$$

Algorithm : The USQR algorithm



USQR ( C )

C, the set of all conditional features

(1)  $R \{ \}$

(2) do

(3)  $T \leftarrow R$

(4)  $\forall x \in (C - R)$

(5)  $\forall y \in C$

(6)  $-\frac{|POS_{R \cup \{x\}}(y)|}{|U|}$

(7) if  $\overline{\gamma_{R \cup \{x\}}(y)}, \forall y \in C > \overline{\gamma_T(y)}, \forall y \in C$

(8)  $T \leftarrow R \cup \{x\}$

(9)  $R \leftarrow T$

(10) until  $\overline{\gamma_R(y)}, \forall y \in C = \overline{\gamma_C(y)}, \forall y \in C$

(11) return R

Table 2.1: Decision System Example

$x \in U$	$a$	$b$	$c$	$d$
1	1	0	2	1
2	1	0	2	0
3	1	2	0	0
4	1	2	2	1
5	2	1	0	0
6	2	1	1	0
7	2	1	2	1

The method seeks to calculate a reduct without generating sets exhaustively. This begins from an empty set and integrates certain attributes in turn, one at a time, which leads to the greatest rise in

the rough set dependence metric before the highest possible value is generated for the dataset. The WEKA method is used to classify data and the performance of classification is analyzed using precision classification and absolute mean error. This method compares with an existing supervised program, as it reveals that inefficiently removing redundant features may.

It is generated by this unsupervised approach are similar in volume to that of the supervised method, and the decreased data grouping indicates that the system chooses usable features of equal consistency. In the future, the same method for breast cancer detection can be applied to mammogram image datasets. In [38] authors introduced a new unsupervised algorithm for selecting features specifically designed To treat the large data collections. This approach uses a weight of the feature identified using samples from a given collection of data to model the degree of significance of each feature at each cluster. To order to determine the features should be chosen and the should be omitted, Then, a defined level applies to certain weights of the function. We empirically show that our method is able to eradicate unnecessary components, resulting in lower mean entropy and more condensed units. The key benefits of our approach is that it produces these outcomes by storing just A proportion of a given set of data and not requiring that the whole collection of data be high sufficient to fit into a computer's memory. These techniques are valuable for those involved in evaluating the broader data sets by removing relevant features from them. In a complex context, the newly created data group will be evaluated along For the determination of the most significant and important features of the entire data collection, the information derived from previous data. Consequently the efficiency and acceptability of the incremental feature selection model in the field of data mining increases. [39] suggested an incremental sorting algorithm base on a genetic algorithm for the collection of the optimized and appropriate subset of functions, The optimal solution for the genetic algorithm used to incrementally select features is defined by the use of rough set theory principles, the reduced and positive area of the target set that was previously generated. The method can be applied in the dynamic environment on a regular basis after The low to moderate amount of data is applied to the framework and hence the processing time, the key issue of the genetic algorithm, does not affect the proposed process. Test findings on baseline methods illustrate that the proposed solution provides satisfactory results in terms of the number of features selected, the calculation time and the results in terms of accuracy of the various algorithms.

## 2.5 Selecting Features in Supervised Learning

Feature Space is found by feature subset selection, candidates are the result and then on the basis of criteria they are evaluated in unsupervised learning. Redundancy and relevancy measures the value of a feature[7]. If a decision feature 's significance is determined then the feature is relevant otherwise it is irrelevant. The one that is highly correlated to other features is redundant. A best subset of apps, connected but not connected to each other with the decision function.

Mostly with rapid growth of digital technology, increasingly large volumes of data like film , images, manuscript, tone and common media interaction, Internet Things and the growth of cloud computing have been generated by massive new computer and internet technologies. These data often have high-dimensional characteristics which present a great The data analysis and decision-making challenge. Variety of Functions Efficient processing of high-dimensional data and increasing learning performance has been proved in both theory and practice. These data often have high-dimensional characteristics which present a great challenge for data analysis and decision-making. Selection of Functions Efficient processing of high-dimensional data and increasing learning performance has been proved in both theory and practice. Selection of features Refers to the process of obtaining a subset based on a selection from the original set of features Choose criterion for the relevant features for the dataset. It plays a role in compressing the huge amount of data processing, where unnecessary even obsolete properties.

Selection of features can enhance studying accuracy, lessen learning time and simplify the learning effects of raw input learning techniques and the outcomes of better selecting features [4–6]. Notably, collection of the apps and Extraction of the function [7,8,9] be two methods of minimizing dimensionality.

Depending on performance, features to be separated into preference app ranks (weightings) and sub-sets. The filter model find relationship among an attribute and a class name. Compared to the standard wrapper it has the Least in cost of computing. A assessment criterion is important for the Plan filter. Meanwhile, in the learning model recruitment process, the embedded model[54–56] selects feature and the features selection automatically results when the training process is in progress terminated.

The evaluation scheme being used in supervised and unsupervised approach of selecting features can be separated in two category[7, 48]:

1. Filter technique
2. Wrapper technique

## **2.6 Filter Techniques**

Features in filter technique are selected without learning algorithm. For selection of features the term rank or ranking is used. For example, the definition of the term given to the individual characteristic is called a score, entropy measure, consistency measure[49]. Type selection strategies related to the function filter have been proposed in literature[50-54]. Univariate and multivariate are the filter based approach according to Alelyani et al. Univariate method / ranking UFS method, uses certain ways to find out its features to find the last list and the ultimate subcategory of features is chosen by sort. These methodologies can not take away redundant ones, as they do not recognize potential dependencies between apps. But certain methods can identify and delete unnecessary features efficiently. Whereas the methods of multivariate filters measure the features' applicability jointly. Multivariate methods are also able to handle redundant and irrelevant features as univariate method, but in most of the problems, the precision achieved through the algorithm by means of multivariate selected features is much better than UFS .

### **2.6.1 FOCUS**

First search large is used to find subsets of features in View. It provides the training and data labeling. FOCUS[55] uses breadth-first search to find subsets of features which provide clear training data labelling. This compares all of the existing size subsets (initially one) and excludes those with the least sum Ignorance. The process continue pending a reliable subset is establish and all of the evaluated

## 2.6.2 SCARP

SCRAP[56] conducts sequential searches to determine the pertinence of features in instance space. This aims to classify certain characteristics that alter the boundaries of decision in the dataset. These characteristics are considered most informative by taking into account one entity at a point. The algorithm starts with the selection of an accidental item, that is the first instance. Then next point is selected normally the closest point with unlike group tag. After this closest object has a distinct class mark which is the next PoC. It demonstrates a boundary of neighborhood of decision and dimensionality among the two groups is defined by the characteristics to modify among them. Then consideration is given if only one characteristic changes between them.

If only one function varies between them, it is considered completely important and is integrated in the group of features if not incrementing the related point of significance (which is initially zero).

## 2.7 Wrapper method

The critique that filter approaches have faced the characteristics to select are independent of the learning algorithm. Wrapper approaches use performance to tackle this problem of the classifier to direct the search. [57].

Four common strategies [7,57] were:

1. Forward Selection : preliminary from an blank group of features, all features are compared then best is chosen, mixed among others.
2. Backward Elimination : In the start, all features are selected, edited unless the best is selected..
3. Genetic algorithm is used in genetic search to find for space on the feature. Every state is described by chromosome, which in fact represents a subset of features. For

finding the features Genetic search is very easy with this representation. However, evaluating the classification accuracy is costly.

4. Contrary to Genetic Algorithm that maintain the chromosome people. Simulated Annealing (SA) considers just single answer. Advance elimination and backward elimination stop as new features are introduced or removed will not affect the accuracy of classification. Simulated Annealing and Genetic Algorithm possibly advanced tools that might be used to understand search space properly.

## 2.8 Unsupervised Feature Selection

It is designed to wrap the expected categorization of facts plus progress cluster correctness by identifying a subset of features based on either clustering for doing so. Unattended methods of selection of apps may be unsupervised methods of selection of filters or wrappers. In the last part it was mentioned that feature selection unsupervised approach can be classified to the policy used to select groups such like wrapper, filter and approaches which are synthetic. First of all, in this part we set the UFS methods into the taxonomy.. After that we define all the methods and their characteristics and ideas.

However, it can be difficult to pick features of unsupervised learning, as the requirements for quality are not clearly established. The literature suggested many unsupervised feature selection methods, e.g.[58-62]. Unattended learning feature selection was classified as supervised learning .

### 2.8.1 Unsupervised Filters

The selection methods for the unsupervised filter function pick features according to the design of the data. In the process of selection, knowledge and cluster algorithms are not used, minimizing the time and the complication of the algorithms. The unsupervised collection technique of the filter function explicitly uses the statistical output of the training data as an assessment tool, that is well scalable, ideal for huge datasets.

Generally the selected function sub-set is lower than the wrapper model since the assessment parameters are self-governing of the particular algorithm suggested to use entropy to determine the value of features and to choose the most important sub-set of features using the trace criterion. An unsupervised array of filter functions.

It is devised for regular class distribution and average class-scatter distance. In every case the value of the decision functions the subset of features is determined, and the subset of features to optimize The function value of the decision is chosen for deciding the candidate work. The function and function selected is calculated. Unless the value of determination is greater than 0.75, then the nominee function is discontinued. The distribution of the function was analysed by Alibeigi et al. data using the density of probability of different function spaces in uncontrolled climate. The function is preferred through the information relative of distribution . Mitra et al established an unsupervised set of features system which uses the highest compression index for information measure similitude between characteristics. That method works quickly.

That extends to datasets of different types, as it does not scan. Chan , Zhou planned an unmonitored algorithm for clustering attributes, along with an unmonitored one task selection process. To construct an attribute distance matrix, determines the maximum in sequence Ratio to every group of qualities, and afterwards cluster all characteristics again using cluster maximum K. Whereas the number of clusters is automatically determined. Li et al proposed an unsupervised clustering based Function selection system called FSFC which follows the same process as clustering supervised selection models. FSFC is also good for high-dimensional datasets.

Techniques of unsupervised collection of features, including Laplacian-based filtering techniques were also suggested. Such techniques pertain to the local data topology creating clusters.

He et . suggested an approach that was base on the assumption with the aim that work would be similar to each other within the same class. In addition, the Laplacian Score used to determine the quality of the functions. Saxena et chosen a genetic algorithm that uses features with the stress function of Sammon, thus the topology structure of the original data is retained in a reduced space for application.

Features of filter based strategies are which they select groups that rely on a certain level or ranking that stays separate from the cycle of grouping or clustering. One example of this technique

is the Laplacian Score (LS), which can be added to DR when reinforcement reaches the locality. The LS is using the concept for the collection of unsupervised features[63]. Features are selected by LS by keeping space among the objects in input and output area. It ensures that all details are to the point and may be unnecessary,

Graph G is used to find out LS. In it the closest neighborly relation is seen.

square matrix  $S = G$

$S_{ij} = 0$  if not neighbors are  $x_i$  and  $x_j$ , whereupon:

$$S_{ij} = e^{-\frac{\|x_i - x_j\|^2}{t}} \quad (2.1)$$

$t$  is a constraint for the bandwidth here

$L = D - S$  represent graph Laplacian and  $D =$  diagonal matrix level as shown below

$$D_{ii} = \sum_j S_{ij}, \quad D_{ij, i \neq j} = 0 \quad (2.2)$$

To find out LS, we use following

$$\widetilde{m}_i = m_i - \frac{m_i^T D \mathbf{1}}{\mathbf{1}^T D \mathbf{1}} \mathbf{1} \quad (2.3)$$

$$LS_i = \frac{\widetilde{m}_i^T L \widetilde{m}_i}{\widetilde{m}_i^T D \widetilde{m}_i} \quad (2.4)$$

Where  $m_i$  is the value vector for  $i$ th feature, and where  $\mathbf{1}$  is a length  $n$  1s.

On this criterion all the features can be found out. The concept could be suitable for domain where protection of localities is an important inspiration[63].

## 2.8.2 Univariate filter methods



In univariate technique, two fundamental to be underlined: strategies to evaluate the significance of every component dependent on Data Hypothesis , what's more, the techniques which assess highlights dependent on Phantom Investigation utilizing the likenesses amongst things. The previous go after the thought of surveying the level of scattering of the information through measures, for example, entropy, dissimilarity, and common data, among others, to distinguish group structures in the information. On the other hand, techniques dependent on Unearthly Examination—Closeness, otherwise called Phantom Highlight Determination techniques, follow displaying or recognizing the neighborhood or worldwide information structure utilizing the eigen-arrangement of Laplacian .

### **2.8.3 Multivariate Filter Method**

This method is classified to three fundamental gatherings: Measurable/Data, Bio-enlivened, and Otherworldly strategies. Previously, the name proposes, incorporates Univariate Feature Selection strategies which play out the determination utilizing measurable or potentially data hypothesis measures, for example, change covariance, straight connection, entropy, shared data, among others. The subsequent gathering, then again, incorporates UFS strategies that utilization stochastic search procedures dependent on the multitude insight worldview for finding a decent subset of highlights, which fulfills some rule of value. At long last, the third gathering incorporates those UFS strategies dependent on Otherworldly Examination or on a mix of Ghostly Investigation and Scanty Learning . It is significant that a few regularly label the final strategies as implanted on the grounds that highlight choice is accomplished as some portion of the learning procedure, normally through the streamlining of an obliged relapse model. Be that as it may, in this examination, we like to order them as channel multivariate, since in expansion to together assess highlights, the essential goal is to perform include choice (or positioning) instead of finding the bunch marks. In addition, we imagine that implanted techniques might be measured a sub-class inside the fundamental methodologies (channel, covering, cross breed), not thwarting the chance to have implanted strategies among three methodologies.

### **2.8.4 Unsupervised Wrappers**

Wrapper-based approaches use the method of grouping or clustering as part of the collection of

features to test the subsets. One of those techniques is suggested in [64]. The concept of a CU (category unit) [65] has been used by authors presenting unsupervised Subcategory optimization for selecting subsets like wrapper. The CU was used as an optimization technique to direct the process of concept development and can be listed as follows:

$$CU(C, F) = \frac{1}{k} \sum_{C_i \in C} \left[ \sum_{f_i \in F} \sum_{j=1}^{r_i} p(f_{ij} | C_i)^2 - \sum_{f_i \in F} \sum_{j=1}^{r_i} p(f_{ij})^2 \right] \quad (2.5)$$

$C = \{C_1, \dots, C_k\}$  is the set of cluster

$F = \{F_1, \dots, F_i, \dots, F_p\}$  is the set of feature.

Category Unit measures the breakdown among the qualified likelihood of a function I in cluster I having value j and its preceding likelihood. The deepest number is above r characteristic value, the center is above p characteristics, the last is above k clusters. It is used as the main term in a wrapper like quest to rank the consistency of the clustering.

UFS techniques dependent on the covering approach can be separated into three general classifications as indicated by the element search technique: successive, bio-roused, and iterative. In the previous, highlights are included or evacuated consecutively. Strategies dependent on consecutive inquiry are anything but difficult to execute and quick. Then again, bio-motivated strategies attempt to consolidate haphazardness into the inquiry procedure, expecting to escape from nearby optima. At long last, iterative techniques address the solo component choice issue by giving it a role as an estimation issue and hence maintaining a strategic distance from a combinatory inquiry.

Unsupervised selection approach to the wrapper feature Is now using a clustering algorithm to modify search reliability of features. The sub-set of feature with the best clustering output will be known as the ultimate optimal subset of features. In the past, subset clustering efficiency of features chosen using the wrapper way is frequently higher than selected feature using the filter method. Every subset features however the clustering algorithm needs to be tested, this approach has a high computational complexity and could present a problem when additionally, have looked through the wrapper system Using EM clustering to pick subsets of functions. This

algorithm is used to approximate the parameter of a finite Gaussian combination with full probability. Separability and scattering is then used with Maximum probability to assess sub-sets of candidate characteristics. Gennari built-in feature selection into CLASSIT, an algorithm for hierarchical information computational clustering. Unsupervised set of functions approach searches for the best feature subset based on the clustering functionality of the features from the most important features. The quest for the feature continue until the current clustering feature you have selected can no longer be changed findings. To find subsets of functions Using clustering algorithm determine the sub-set of features, and clustering accuracy selects the best subset of features.

Dom ,Vaithyanathan developed use the Bayesian statistical evaluation model to select a sub-set of features. To find optimum number of clusters in the document clustering problem. For each cluster They constructed a polynomial model, and extended the clustering concept algorithm.

### **2.8.5 Sequential techniques**

In Brodley work, two component choice models were assessed: the measure of utmost Likelihood and the disperse distinguishableness basis . The technique look throughout the gap of highlight subsets, assessing every applicant subset ,bunching calculations are functional on the information depicted by every applicant subset. At that point, the got bunches are assessed with the ML or TR rules. The strategy utilizes a forward choice quest for producing subsets of highlights that will be assessed as portrayed previously. The technique closes when the adjustment in the estimation of the pre-owned rule is littler than a given limit.

A strategy which utilizes another enhancement measure for, individually, limiting and augmenting the intra-bunch and between group idlenesses was proposed in Luchain. The creators suggest a capacity, impartial . the quantity of groups and highlights, base minimization augmentation of the change of dissipate frameworks acquired from the groups worked by the k-implies bunching calculation. This capacity allots a positioning gain to every segment that might be characterized in the hunt room of every single imaginable subset of highlights and number of bunches. The rule proposed in this technique gives both a positioning of pertinent highlights also, an ideal parcel.

### **2.8.6 Bio-propelled techniques**

In this grouping a Delegate UFS technique was proposed in Kim et al. (2002) where a transformative neighborhood determination calculation (ELSA) was proposed to look through subsets of components as well as the number of groups dependent on the Gaussian Mixture 's k-implies and bunching calculations. Every arrangement gave by the bunching calculations is related with a vector whose components speak to the nature of the assessment standards, which depend on the attachment of the groups, between class division, and greatest probability. Those highlights that enhance the target capacities in the assessment stage are chosen. Another strategy, additionally dependent on a transformative calculation, highlight choice perform whereas the information is bunched utilizing a multi-objective hereditary calculation. The strategy propose a multi-target wellness work that limits the intra-group separation (consistency) and expands the between bunch division. Every chromosome speaks to an answer, which is created by a set of k group centroids (bunch community for constant highlights and group mode for straight out highlights) portrayed by a subset of highlights. The quantity of highlights utilized for every centroid in every chromosome is haphazardly produced, and the group communities and bunch methods of chromosomes in the underlying populace are made by producing irregular numbers, and highlight values from a similar component area, individually. At that point, for assigning group centroids again, MOGA utilizes the k-models bunching calculation that acquires the contributions among underlying populace produced in the past advance. A short time later, the hybrid, change, and replacement administrators be practical, and the procedure is rehashed until a prespecified stop rule is met. In end, strategy restores component group which improves the wellness work mutually along the groups that are delivered.

### **2.8.7 Hybrids**

Using a measure the inherent property of the facts, the features are ranked or selected in a filter stage to take benefit of the filter and wrapper approach, hybrid methods. Some subsets of features are tested by a common clustering algorithm to find the right one in wrapper phase. Hybrid methods of two type may be distinguished: method focused on rating, and method not focused on function rating. Here , we describe certain method belonging to this method, of both

types. It was one of the first methods for picking unsupervised hybrid function dependent on rating. The approach function, one by one, is extracted from the entire collection of features in the filter level, and after removing the function, the entropy produced in the dataset is measured.

A sorted list of characteristics is created to the quantity of disorder generated by each characteristic when removed from the whole set of functions. When every functions are sorted a forward selection test is conducted in the wrapper stage in conjunction with the k-means clustering algorithm to create clusters and find out using the separability criterion for scattering

### **2.8.8 Fuzzy c-means Clustering**

The goal for clustering approaches, the function space partition must be formed, assigning each training point set to a single cluster. Furthermore, the fuzzy C-Means clustering methods, a technique of fuzzy clusters build as a fuzzy-set for each of the  $c$  cluster. In which all the training instances have some degree of accession. The main aim of the clustering algorithm is to divide the dataset into classes of similar features. To evaluate the similarity of the fuzzy clusters between clusters, the fuzzy C-Means technique was supplemented by a fluffy rough assessment check. Using this indicator, the results of many runs of the process can be compared with varying values of  $c$ , to obtain the optimum number of clusters that needs to be created. However, clustering technique also tries to certify that two features have a high degree of membership in the same fluffy cluster.

## **2.9 Supervised Feature Selection Using DDC**

DDC can also be used in any selection algorithm, simply replace the measurement of positive region based dependency with DDC. Just as in IDC, we re-implemented all the different algorithms discussed in the related section of the research using DDC approach. In short, we'll discuss these algorithms here, as these have already been discussed in detail in previous sections.

### **2.9.1 Quick Reduct PSO:**

PSO on Quick Reduct [32], PSO-quick reduct was originally designed to use the measure of positive region based dependency. The algorithm was implemented again using the method of

calculating direct dependence.

### **2.9.2 Genetic Algorithm:**

Genetic Algorithm is same as described above. The only improvement made was that models based on DDC were used in fitness model, in comparison to the initial approach of measuring positive regional dependency.

### **2.9.3 Incremental Feature Selection Algorithm:**

IFSA[32] use the calculation of positive region-based dependency. Like IFSA based IDC, implemented it again with DDC based process. Every measures measuring positive area dependent dependency have been replaced by DDC based process. Rest of algorithm details have been kept intact.

### **2.9.4 Fish Swarm Algorithm:**

Fish Swarm [44] used swarm based optimisation on the way to select apps. For all searching, swarming and subsequent behaviour, positive region based dependency measure is used. Criteria meant for stopping were also based on a positive approach based on region. We substituted all the measures focused on positive area with DDC.

### **2.9.5 Rough Set Improved Harmony Search Quick Reduct**

In [45], a technique is submitted combined with a better search algorithm for harmony based on the Rough set theory. The positive region-based methods, it also uses a predictable measure of positive region-based dependence. Yet we did make this algorithm work with DDC-based process.

### **2.9.6 Tolerance Rough Set Firefly based Quick Reduct**

Another method was developed and implemented for mind imagery with MRI [46] that follow Tolerance rough firefly. An integrated clever device seeks to leverage the profit of the simple model while still moderating their limits.

The attributes taken from brain tumor Images are actual concepts. Therefore Rough Tolerance collection is added in this work. In this analysis, the imperative characteristics of the brain tumor are chosen using a combination of two techniques, Resistance Rough Range and Firefly Algorithm . TRSFFQR efficiency is compared to Colony of Artificial Bees , Cuckoo Search Algorithm.

### **2.9.7 Improve Quick Reduction for Function Selection**

Researchers in [47] have suggested new approaches for reducing FRFS estimates. This new method was proposed based on Fuzzy Lower Approximation-Based Feature Selection, which selects smaller subsets of features, improves classification accuracy and runs faster than the base method , especially on large datasets.

This is done using a threshold based stop criterion that prevents the addition of additional features in the QuickReduct algorithm. Performance and effectiveness of our proposed method are confirmed by experimental results on UCI datasets.

Supervised selection of hybrid features based on PSO and rough medical diagnostic sets FS is vital element of the detection of knowledge. A novel methods is proposed Particle Swarm Optimization (PSO) hybridization, for the diagnosis of diseases. The experimental results on more than a few regular medical datasets demonstrate the competence of the projected method and improvements above the presented techniques for selecting features.

### **2.10 Clustering algorithm using rough set theory for unsupervised feature selection**

In this an unsupervised feature selection algorithm is proposed using: (a) Relative dependence on finding similarities between characteristics, (b) a clustering algorithm for grouping related characteristics, and (c) a method for selecting mainly ambassador characteristic to get a cheap space. The comparative quantity of dependence among characteristic pairs is used to calculate a similitude calculation. A clustering algorithm utilizes this measure to perform clustering attributes By KNN and clustering based on prototype. The concept is experienced using familiar

criteria and matched to traditional methods of selecting apps that are supervised and unsupervised. In fact, our plan assesses a real-world framework for spinning machines in fault diagnosis. One of the accepted feature selection algorithm is K-means [40-43] . It makes K clusters of N-entities related features. For example: It allows the number of clusters, K, to be determined beforehand; (ii) it can be stuck in local minima; (iii) it expects the degree of significance of each element to be the same; (iv) it can not cluster wide data sets at an appropriate level.



## Chapter 3

### 3 Research methodology

*"Methodology of science is the formal, analytical review of the procedures applied to a study area. Methodology requires methods to describe, explain and forecast phenomena in order to solve a problem; it is the "how;" the analysis method or techniques." (Kothari, 2004)*

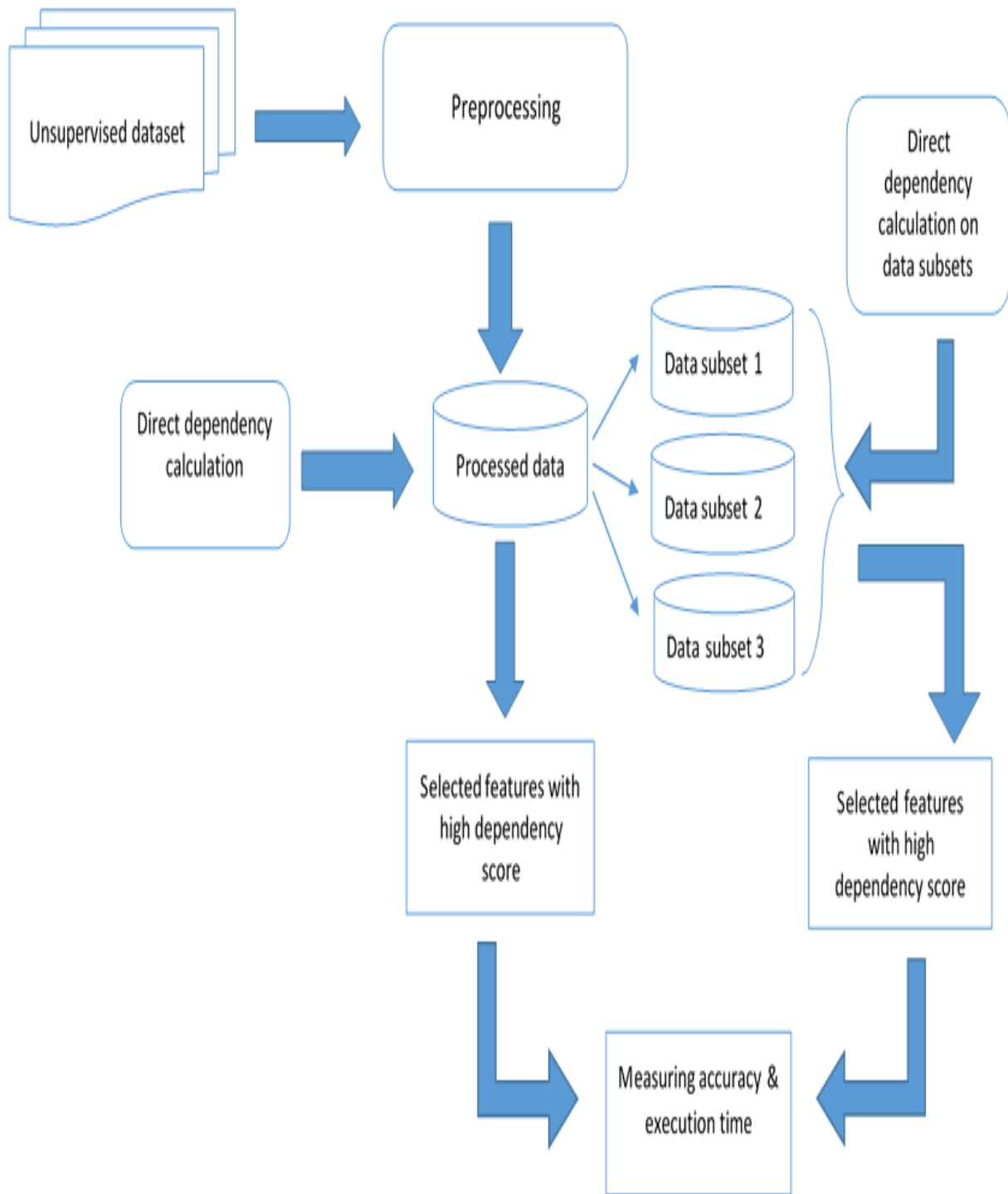
#### 3.1 Problem statement

In previous literature direct dependency calculation (DDC) technique was proposed for feature selection of supervised data set. In this research study I have carried out feature selection for unsupervised data set using direct dependency calculation. This approach provides accurate and efficient results. DDC was applied on 5 different types of datasets of different domains. All the datasets were unsupervised. This approach shows better results as compare to supervised DDC.

Proposed technique mainly consists of 4 steps

- *Dataset collection*
- *Preprocessing*
- *Direct Dependency Calculation (DDC)*
- *Feature selection*
- *Measuring execution time and accuracy*

These steps are represented in figure below.



*Figure 3.1: Feature Selection from unsupervised dataset using DDC*

### 3.2 Dataset collection

Data collection is one of the most significant activity in any kind of research, especially in machine learning where everything we do is dependent on data. Therefore for this fundamental step we have collected few data sets of different domains from UCI machine learning repository and Kaggle. The figure below shows the sample datasets used in our study.

### 3.3 Preprocessing

- a. The redundant and noisy data was normalized in this step.
- b. The textual attributes in the datasets were mapped to numerical form as the DDC algorithm works well on numerical data therefore to get accurate calculation we mapped the textual data to numerical one.

### 3.4 Direct Dependency Calculation (DDC)

- a. In this step we have applied two rules of direct dependency class calculation whereas, in incremental dependency calculation there are four rules. Using these two rules gives precise calculations over incremental approach.
- b. DDC calculates the amount of single groups in a data collection for an trait C. A special class reflects the values of the attributes that refer to a particular class of decisions in the dataset, and this classification may be used to correctly classify the class of decisions. Non-unique groups reflect the meanings of attributes belonging to specific groups of decisions, but they can not be used specifically to evaluate the class of decisions.

*Table 3.1: Calculates dependency using DDC*

<b>Dependency</b>	<b>amount of groups that are unique / non-unique</b>
0	no unique class
1	no class which is not unique
n	Where otherwise $0 < n < 1$

DDC Dependency to be determined using the procedure below:

If we take into account the number of unique classes :

$$\gamma(\{att\}, D) = \frac{1}{N} \sum_{i=1}^m (1) \quad (1)$$

If we consider the classes which are not unique:

$$\gamma(\{att\}, D) = 1 - \frac{1}{N} \sum_{i=1}^n (1) \quad (1)$$

Where

$(\{att\}, D)$  is the "D" attribute dependency on the  $\{att\}$  attribute.

$\{att\}$  is the existing attribute of decision (Decision class) under consideration D.

M is sum of values which lead to specific classes of decisions

N is the whole of data records leading to the non-unique decision class N

The assumed attribute D here is dependent on attribute C with a measure of K.

Table 3.2 gives an example of an unsupervised data set

*Table 3.2: Unsupervised dataset*

<b>U</b>	<b>a</b>	<b>b</b>	<b>c</b>	<b>d</b>
11	1	0	2	1
12	1	0	2	0
13	1	2	0	0
14	1	2	2	1
15	2	1	0	0
16	2	1	1	0
17	2	1	2	2

Take into account the decision system set out in the table above  $(\{att\}, D) = \frac{1}{N} \sum_{i=1}^m (1)$

**For Unique Class:**

**First Iteration:**

U	a'	b'	c'	d'
11	1	0	2	1
12	1	0	2	0
13	1	2	0	0
14	1	2	2	1
15	2	1	0	0
16	2	1	1	0
17	2	1	2	2

If we consider attribute  $\{b\}$ ,

$$\gamma(\text{Attribute}, d) = \frac{1}{N} \sum_{i=1}^N (\gamma'_i)$$

$$\gamma(\{a\}, b) = \frac{1}{7}(1)$$

$$\gamma(\{a\}, b) = \frac{1}{7}(1+1)$$

**Second Iteration:**

U	a'	b'	c'	d'
11	1	0	2	1
12	1	0	2	0
13	1	0	0	0
14	1	2	2	1
15	2	1	0	0
16	2	1	1	0
17	2	1	2	2

If we consider attribute {b},

$$\gamma(\text{Attribute}, d) = \frac{1}{N} \sum_{i=1}^N (\gamma'_i)$$

$$\gamma(\{a\}, b) = \frac{1}{7}(0+0+0)$$

$$\gamma(\{a\}, b) = \frac{1}{7}(0+0+0+0)$$

**Third Iteration:**

U	a'	b'	c'	d'
11	1	0	2	1
12	1	0	2	0
13	1	2	0	0
14	1	2	2	1
15	2	1	0	0
16	2	1	1	0
17	2	1	2	2

If we consider attribute {b},

$$\gamma(\text{Attribute}, d) = \frac{1}{N} \sum_{i=1}^N (\gamma'_i)$$

$$\gamma(\{a\}, b) = \frac{1}{7}(0+0+0+0+1)$$

$$\gamma(\{a\}, b) = \frac{1}{7}(0+0+0+0+1+1)$$

$$\gamma(\{a\}, b) = \frac{1}{7}(0+0+0+0+1+1+1) = \frac{3}{7}$$

For non unique class

First Iteration:

U	a'	b'	c'	d'
11	1	0	2	1
12	1	0	2	0
13	1	2	0	0
14	1	2	2	1
15	2	1	0	0
16	2	1	1	0
17	2	1	2	2

If we consider attribute {b},

$$\gamma(\text{Attribute}, d) = 1 - \frac{1}{N} \sum_{i=1}^N (\gamma_i')$$

$$\gamma(\{a\}, b) = 1 - \frac{1}{7}(0)$$

$$\gamma(\{a\}, b) = 1 - \frac{1}{7}(0+0)$$

$$\gamma(\{a\}, b) = 1 - \frac{1}{7}(1+1+1)$$

$$\gamma(\{a\}, b) = 1 - \frac{1}{7}(1+1+1+1)$$



Second Iteration:

U	a'	b'	c'	d'
11	1	0	2	1
12	1	0	2	0
13	1	2	0	0
14	1	2	2	1
15	2	1	0	0
16	2	1	1	0
17	2	1	2	2

If we consider attribute {b},

$$\gamma(\{a\}, b) = 1 - \frac{1}{7}(1+1+1+1+0)$$

$$\gamma(\{a\}, b) = 1 - \frac{1}{7}(1+1+1+1+0+0)$$

$$\gamma(\{a\}, b) = 1 - \frac{1}{7}(1+1+1+1+0+0+0)$$

$$\gamma(\{a\}, b) = 1 - \frac{1}{7}(4)$$

$$\gamma(\{a\}, b) = 1 - \frac{1}{7}(1+1+1+1+0+0+0) = \frac{3}{7}$$

No unique classes + no non uniqueness classes = universe size

So we have to either compute the amount of single classes or non-single classes

Dataset further divided into three subsets on which direct dependency class calculation was performed in parallel. Through this, the execution time will be enhanced without effecting the accuracy of the results obtained in step b.

### **3.5 Feature selection**

We get dependency score along each attribute and the relative attribute of the decision making attributes will be the one whose score is high. As per the rules of DDC, the greater the dependency the more relevant that attribute is. That illustrates the selected features we get from this whole process.

### **3.6 Implementation Environment**

#### **3.6.1 Tools and programming languages used:**

The following python libraries were used to implement direct dependency classes calculation algorithm: Numpy, Pandas, Operator, Scikitlearn, Dask, Scipy, Time, Matplotlib, XGBoost.

##### **3.6.1.1 Python:**

Python is one of the programming languages most widely used, and has overshadowed other languages in the industry.

There are several reasons why developers are popular with Python, and one of them is that they have a large set of libraries with which users can use them.

Here's just a handful major causes why Python is widely known:

- Python has extensive library collections
- Python is regarded as a programming language for beginners because of its simplicity and ease of use

- From deployment to maintenance Python requires the developer to be more competitive
- Flexibility is just another reason for the enormous success of Python
- Python syntax is easy to understand, and strong in contrast to C , Java and C++

This allows for the development of new software by script smaller amount of code.

Simplicity in Python has involved many developers into building new libraries for machine learning. Python is becoming more and more popular among experts in machine learning because of the large selection of libraries.

### **3.6.1.1.1 Numpy**

Numpy is well-liked machine-learning library of Python's . Tensor Flow and other libraries use Numpy internally to perform procedures on trigonometric functions. Numpy 's best, most important component is the interface to an array.

#### **3.6.1.1.1.1 Numpy Features**

1. Interactive: Numpy is very interactive and very user friendly.
2. Mathematics: Makes complex implementations of mathematics very simple.
3. Easy to understand: Coding really simple and it's easy to grasp the concepts.
4. Communication: Widespread use, thus much open source participation..

Such a module is used to express imagery noise waves etc as an collection of N-dimensional real numbers. For full stack developers it is important to implement machine learning library, which has Numpy expertise.

### **3.6.1.1.2 Pandas**

It is Python machine learning library providing a wide array of high-level analytical tools and data structures. One of this library's great features is its ability to use one or two commands to translate complex data-based operations. Pandas have many streamlined methods for classification, merging and manipulating data and the features of the time series.

Pandas make sure the entire data manipulation process gets easier. Operations help such as Re-indexing, Iteration, Sorting, Aggregations, Concatenations and Visualizations are among Pandas' function highlights.

#### **3.6.1.1.2.1.1 Use of Pandas**

There are currently fewer pandas library releases that include fresh features , Bugs are fixed, modifications and updates to the API. The enhancements to the pandas relate to its ability to combine and sort data, chose the best output for the application process, and support custom activity forms. If it comes to the use of Pandas, data analysis takes the spotlight over everything else. Pandas still maintains high levels when used in conjunction with other libraries and tools accessibility and versatility.

#### **3.6.1.1.3 Scikit-learn**

It is a compatible Python library with NumPy, and SciPy. It is considered one of the best libraries for working with complex data. There are plenty of improvements in this library. One update is the cross-validation feature which allows more than one metric to be used. Many training methods such as logistics regression and nearest neighbors have been moderately improved.

##### **3.6.1.1.3.1 Scikit-Learn Features**

- 1.Cross-validation: Different methods exist for checking The accuracy of tracked models on different data sets
- 2.Unsupervised learning algorithms: the selection of algorithms is once again widespread – starting with clustering, factor analysis, key component analysis, and uncontrolled neural network.
- 3.Extraction feature: Its helpful to remove image and text functions ( e.g. word bag).

##### **3.6.1.1.3.1.1 Use of Scikit-Learn**

It has several algorithms for Basic machine learning and data mining activities such as noise removal, classification , deterioration , cluster and selection of model.

#### **3.6.1.1.4 SciPy**

A library which Provides programmers and experts with data science. The difference between the library and stack, however, tio be identified. This library includes the modules for performance tuning, algorithm design, mathematics and implementation.

### 3.6.1.1.4.1 SciPy Features

The major characteristic of the library is to built by NumPy, it set uses NumPy to the max. Furthermore, It use own sub module to give the important arithmetical routine such as optimisation, numerical incorporation and several more. Many of the functions are well defined in all SciPy submodules.

#### 3.6.1.1.4.1.1 Use of SciPy

SciPy is a library that utilizes NumPy to solve math functions. SciPy employs NumPy sequences as the simple data structure, and provides modules for different commonly used science programming tasks.

SciPy easily handles activities include mathematical principles, calculus integration, standard differential equation solution, and signal processing.

## 3.7 DDC Algorithm

```
Step 1: U= Universe Set with all features
Step 2: Distribution of U
|U| = |U1|, |U2|, |U3|,....., |Un| such that number
of subsets depend on number of instances in |U|
Step 3: Assume D = C such that D ∈ C
Where C= Conditional attributes
Such that C = {C1,C2,C3,.....Cn}
Step 4: Calculate attribute dependency based on two
pre-defined DDC rules:
Unique Classes
p= Y(C,D)=  $\frac{\text{unique classes sum}}{I U I}$ 
Non-Unique Classes
k = Y(C,D)=  $1 - \frac{\text{non-unique classes sum}}{I U I}$ 
Step 5: Generate Reduct R  $\forall C$  with highest K score
Where,
p = 1; maximum dependency
p = 0; not dependent
0<p<1; partial dependency
```

## Chapter 4

### 4 Dataset, Implementation and outcome

The section is about the tools, environment used, dataset, the specific implementation steps which are taken, and the results of the model search and transfer learning tasks.

#### 4.1 Dataset Collection

For the purpose of dependency calculation we have used python. It has seen that Python become most popular language of data sciences in the year of 2018. Then a UCI dataset is used for that purpose. From UCI Repository Machine Learning Database, the datasets are collected. The details of the datasets are given in Table 4.1.

*Table 4.1: Set of Datasets used*

S.No	Datasets	Attributes	Instances
1	Student Assessment	5	173912
2	Telecom	19	3333
3	Adult census income	15	32560
4	Australian weather	24	142193
5	Breast cancer	33	569

#### 4.2 Feature Selection Accuracy and Execution Time

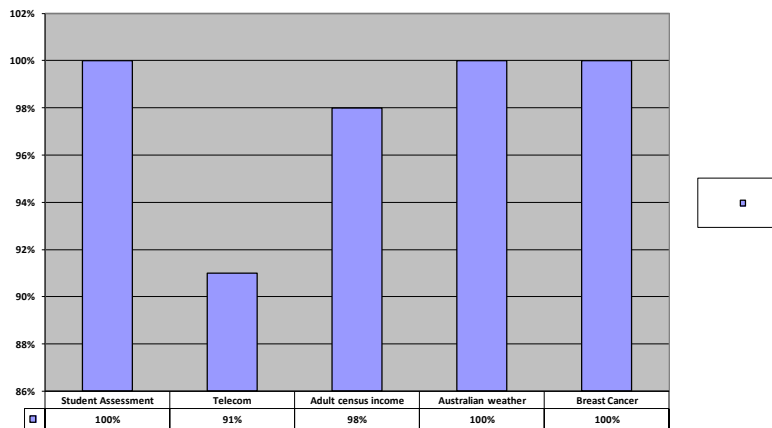
Table 4.2 shows the accuracy of different datasets based on implemented feature selection method. A given dataset contains both redundant and relevant features where the relevant features needs to be extracted for further processing. The irrelevant features are those features which has minimum dependencies and does not seem useful for gaining useful information. On the other hand, attributes with highest dependencies in a given data set

illustrates that they are highly useful and provides knowledgeable information. These dependencies can be easily calculated in supervised datasets with given class labels and decision attribute. In unsupervised datasets dependencies calculation poses a great problem due to absence of D. Code and time complexity increases since D has to be assumed multiple times depending on the size of the dataset. Therefore, DDC algorithm is able to calculate dependencies by executing only two rules where a set of unique and non-unique classes are determined on attribute values. The datasets were distributed into multiple subsets depending on the size of the overall dataset. The algorithm was implemented on these subsets in a parallel way so that execution time can be further minimized as shown in Table 4.3.

Table 4.2 shows feature selection accuracy of different datasets when DDC was applied on them. Using NumPy we have done preprocessing on the selected datasets and removed the redundant and duplicated data.

*Table 4.2: DDC Algorithm classification accuracy*

S.No	Datasets	Accuracy
1	Student Assessment	100%
2	Telecom	91%
3	Adult census income	98%
4	Australian weather	100%
5	Breast Cancer	100%



*Figure 4.1: Unsupervised Feature Selection Classification Accuracy*

Table 4.3: DDC algorithm execution time

S.No	Datasets	DDC Execution Time With Parallel Processing	DDC Execution Time Without Parallel Processing
1	Student Assessment	2.92052	4.953837
2	Telecom	92.510057	156.139446
3	Adult census income	7.158889	11.231004
4	Australian weather	48.050717	63.365123
5	Breast Cancer	113.943083	126.713946

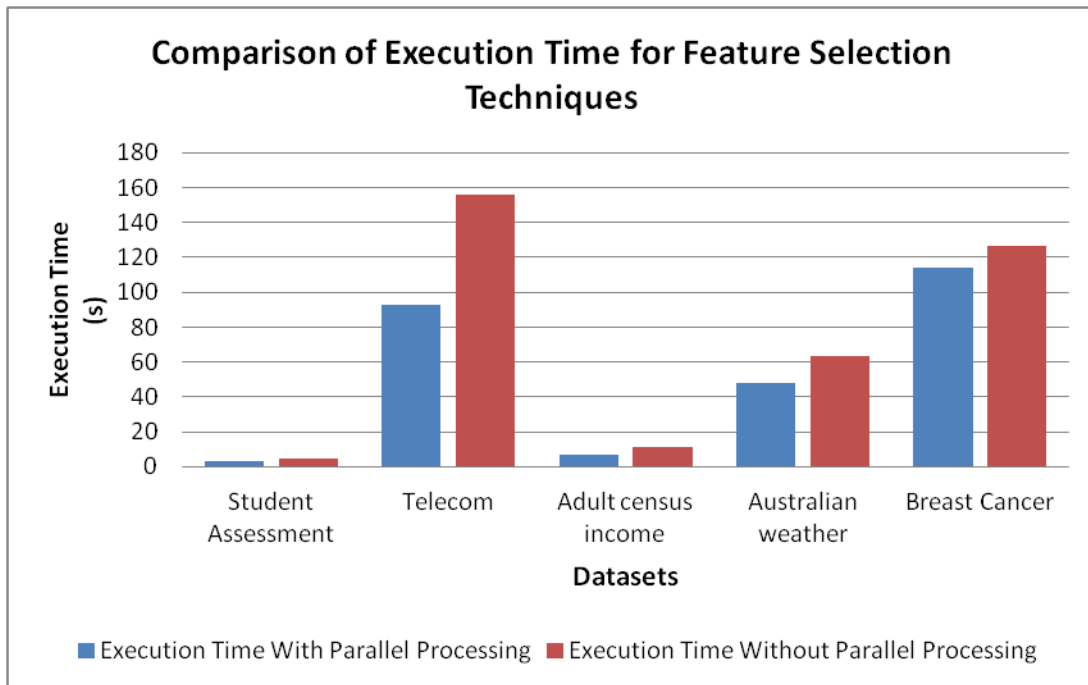


Figure 4.2: Execution time of DDC Algorithm with parallel and without parallel processing



### 4.3 Comparison with other algorithms

Table 4.4 : Comparison with other algorithm

No.	Datasets	UDDC	Decision Tree	KNN	Random forest	MLP
1.	Breast cancer	100%	92% [80]	95.27 [80]	95.61% [85]	95.42 [80]
2.	Diabetes	99%	87% [79]	0.78 [78]	91% [82]	97.65 [88]
3.	Wine quality	99.9%	94.51 [86]	98.93 [88]	81.96 [81]	78.78% [81]
4.	Census	99.8%	85.5% [83]	93.934 [87]	88.71 [84]	

The method is contrasted with findings of popular standard classification algorithms already published in literature to assess the efficacy and efficiency of the proposed DDC system. Such algorithms include the Random Forest, Decision Tree, KNN, MLP etc. The algorithms mentioned were applied to the miscellaneous dataset, and the same datasets were used in DDC.

Results of the proposed DDC method are assessed and compare on the base of sorting accuracies on the condensed classifier datasets. In our work, , K-nearest neighbors, Multi-layer Perceptron (MLP) ,Random Forest, Decision Tree are considered classifiers. The worth of K-NN is set inside the square root of the data element. In the table are already calculated the Number of original attribute, a after the proposed one DDC and the accuracies (percent) of the reduced data sets of the above classifiers.. It is observed from Table 2, that in most cases the proposed DDC Method selects fewer classification attributes of greater accuracy than other methods. The suggested solution known as Direct Dependency Calculation measures dependence estimate directly without the time-consuming positive area analysis being done.

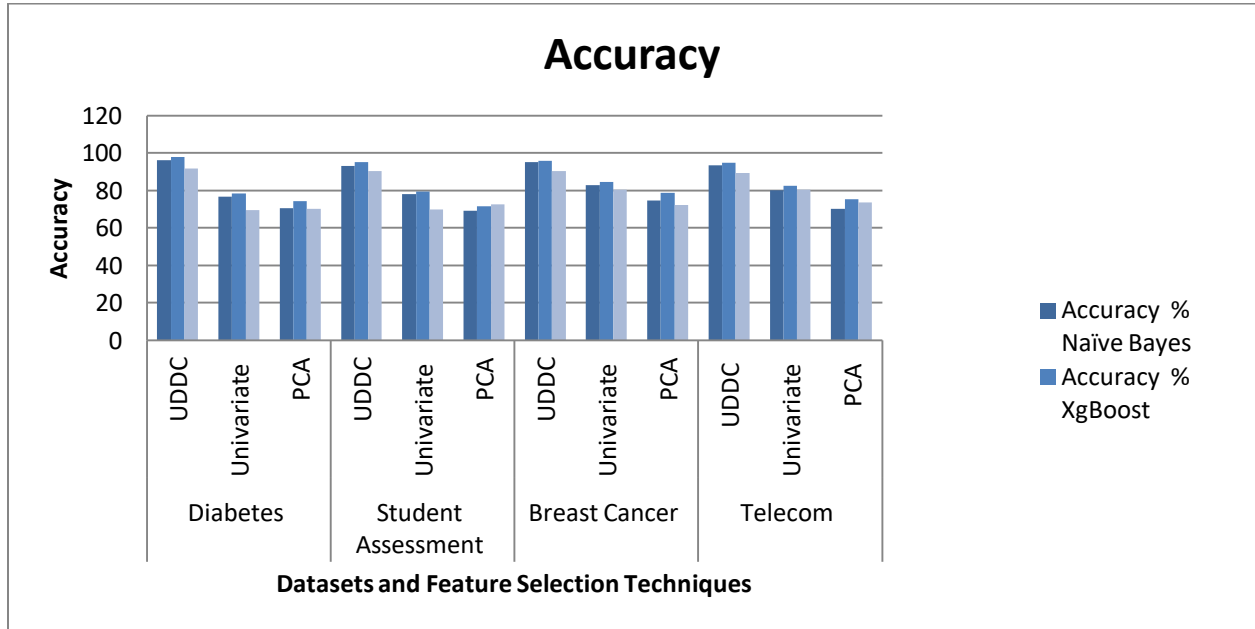
It checks the amount of specific groups in a dataset by means of characteristic value explicitly, and measures dependency. Dependency estimate in this way allows us to keep away from the positive region, making DDC based selection feature algorithms appropriate

for regular and better datasets. In this way measuring dependency helps us to escape the positive zone, allowing DDC based choice function algorithms ideal for normal and better datasets. The planned method is an option Calculates regular optimistic region-based dependence, and a rough set-based dependence measure can be used safely in any selection of features algorithm. Calculation of dependencies in unsupervised datasets creates a major problem due to the absence of D. Complexity of code and time increases because D has to be inferred several times depending on the dataset size. Therefore, the DDC algorithm can determine dependencies by executing only two rules while evaluating a set of unique and non-unique classes based on attribute values. Depending on the amount of the overall information set, the datasets were distributed into multiple subsets. The algorithm has been implemented in parallel on these subsets, so that execution time can be further reduced as seen in Table 4.4 above.

*Table 4.5 : Accuracy Comparison*

<b>Accuracy</b>						
<b>Datasets</b>	Feature Selection Methods	# of Instances	# of Features	Accuracy (%)		
				Naïve Bayes	XgBoost	kNN
<b>Diabetes</b>	UDDC	768	9	96.2	98	91.77
	Univariate	768	9	76.62	78.35	69.69
	PCA	768	9	70.562	74.45	70.129
<b>Student Assessment</b>	UDDC	173912	5	93.1	95	90.5
	Univariate	173912	5	78.2	79.32	70
	PCA	173912	5	69.14	71.52	72.5
<b>Breast Cancer</b>	UDDC	569	32	95.17	96	90.45
	Univariate	569	32	82.7	84.5	80.4
	PCA	569	32	74.58	78.7	72.3
<b>Telecom</b>	UDDC	3333	19	93.5	94.74	89.2
	Univariate	3333	19	80.3	82.4	80.6
	PCA	3333	19	70.3	75.2	73.7

Figure 4.3: Accuracy Comparison



Another comparison was drafted to assess the accuracy of proposed feature selection algorithm. In this method, we have selected feature selection techniques for unsupervised datasets. These techniques include; principal component analysis for feature selection and univariate feature selection along with proposed UDDC feature selection technique. These feature selection techniques were applied on five miscellaneous datasets taken from UCI machine learning data repository. With the help of these techniques important features were extracted on the basis of their scores. the extracted features were then passed to classifiers. The classifiers we used include; Xgboost, kNN and Naïve Bayes.

Table 4.6 : Execution Time of different techniques on different datasets

<b>Execution Time</b>				
<b>Datasets</b>	Feature Selection Methods	# of Instances	# of Features	Execution Time (s)
<b>Diabetes</b>	UDDC	768	9	6.75
	Univariate	768	9	19.4
	PCA	768	9	17.12
<b>Student Assessment</b>	UDDC	173912	5	14.1
	Univariate	173912	5	22.5
	PCA	173912	5	18.4
<b>Breast Cancer</b>	UDDC	569	32	4.73
	Univariate	569	32	14.7
	PCA	569	32	8.35
<b>Telecom</b>	UDDC	3333	32	9.1
	Univariate	3333	32	12.5
	PCA	3333	19	11.4
<b>Australian weather</b>	UDDC	142193	24	12.4
	Univariate	142193	24	40.4
	PCA	142193	24	26.7

Results of the proposed DDC method are assessed and compare on the base of sorting accuracies on the condensed classifier datasets. The worth of K-NN is put to the square root of the sample data element. The table 4.3 presents the results of our comparison by displaying the names of the

datasets, feature selection techniques used for comparison along with proposed UDDC, the total no of instances and amount of features in the dataset and then the quantity of selected with each technique. The accuracy achieved with each classifier upon all datasets is also presented in the said table.

It is observed from Table 4.3, that in most cases the proposed DDC method gives advanced categorization accuracy compared to the other method. The highest accuracy for DDC with XgBoost classifier is calculated as 98% then Univariate given second highest value of 78 % using diabetes dataset. Similarly, this trend can also be seen with other datasets as well.

The suggested solution known as Direct Dependency Calculation measures dependence estimate directly without the time-consuming positive area analysis being done. It checks the amount of specific groups in a dataset by means of characteristic value explicitly, and measures dependency. In this way , dependency calculation allows us to keep away from the positive region, making DDC based selection feature algorithms appropriate for regular and better datasets. In this way measuring dependency helps us to escape the positive zone, allowing DDC based selection function algorithms ideal for normal plus better datasets. The planned method is an option to the predictable positive region-based dependence calculates, and a rough set-based dependence measure used carefully in feature selection algorithm. Calculation of dependencies in unsupervised datasets creates a major problem due to the absence of D. Complexity of code and time increases because D has to be inferred several times depending on the dataset size. Therefore, the DDC algorithm can determine dependencies by executing only two rules while evaluating a set of unique and non-unique classes based on trait standards.

#### **4.4 Time Complexity using Big O notation**

Big O notation is used to describe performance, or complexity, of an algorithm. Big O specifically describes the worst-case scenario and can be used to describe the execution time or space required for the algorithm

Table 4.7 : Time Complexity using Big O notation

Step by Step Algorithm	Big-O
U=1000	
df_=df[cols].iloc[:U]	
count=1	
final_dict=dict()	
scores_=dict()	
U_name='U'	
for D in df_.columns:	O(U)
# D='a'	
if D!=U_name:	
for x in df_.columns:	O(U <sup>2</sup> )
dic=dict()	
if x!=U_name and x!=D:	
for i,j in zip(df_[D],df_[x]):	O(U <sup>2</sup> )
if (i,j) in dic.keys():	
dic[(i,j)]=dic[(i,j)]+1	
else:	
dic[(i,j)]=count	
temp_df=pd.DataFrame(columns=['coll'])	
for x1 in dic.keys():	O(U <sup>2</sup> )
temp_df=temp_df.append({'coll':x1[1]},ignore_index=True)	
li=list(temp_df.coll.value_counts().reset_index(name="count").query("count == 1")["index"])	
final_li=[]	
for x2 in dic.keys():	O(U <sup>2</sup> )
if x2[1] in li:	
final_li.append(x2)	
sum1=0	
for item in final_li:	O(U <sup>2</sup> )
sum1=sum1+dic[item]	
final_dict[x]=sum1	
score=(sum(list(final_dict.values()))/U)/(len(df.columns)-1)	
scores_[D]=score	

To compute the Big-O time complexity we will consider the highest degree term. Therefore, time complexity of proposed algorithm is:

$$\begin{aligned} \text{Big-O Time Complexity} &= O(U) + O(U^2) + O(U^2) + O(U^2) + O(U^2) + O(U^2) + O(U^2) \\ &= O(U^2) \end{aligned}$$

## Chapter 5

### 5 Conclusion and Future Work

#### 5.1 Conclusion

In this research study, direct dependency class algorithm was implemented on unsupervised and unlabeled datasets for the purpose of characteristic selection based on rough set theory. The foremost goal be toward to reduce code complexity and execution time while calculating dependencies of attributes on each other in a given dataset. The attributes with maximum dependencies can be extracted for further processing methods in machine learning. Decision attribute has to be selected from the given conditional attribute set due to which execution time and algorithm complexity becomes a challenge. DDC is able to calculate dependencies of attribute for an assumed  $D$  from the conditional attributes dataset by utilizing two rules of determining unique and non-unique classes. The results show great improvement in provisions of feature selection accuracy and execution time with parallel processing. Parallel processing was also implemented in addition to further reduce the execution time of DDC algorithm.

Direct dependency classes specify how dependency value change as new record is read in dataset. It means that reading a record belonging to unique class will increase the dependency and reading a record belonging to non-unique class will decrease dependency.

In our research we have proposed unsupervised direct dependency class calculation feature selection algorithm that works for unsupervised dataset. Through this approach we do not have to calculate the positive region calculations as that is the time consuming effort instead we use direct dependency calculation to calculate the dependency measures. This helps in examining straight away the number of unique classes. This direct method is suitable for larger datasets as compare to positive region.

We have proposed a direct dependency class calculation algorithm on unsupervised dataset,

which has not been done before. The main goal is to extract useful features, reduce the code complexity and execution time while calculating dependencies of attributes on each other in a given dataset. This technique successfully performs feature selection by using two set of rules of direct dependency calculation. To verify the reduced execution time and algorithm complexity we carried out the experiment on standard datasets take from the UCI library. The results show great improvement in conditions of feature selection correctness and execution time with parallel processing. UDDC provides the accuracy above 95% when checked with other algorithm of feature selection.



## 5.2 Future work

Collection of features helps the learning strategies to work more efficiently by reducing the impact of unnecessary knowledge to increase classification efficiency. Unlike controlled and semi supervised feature selection, unmonitored characteristic selection is deemed a a great deal tougher trouble owing to the difficulties of determining feature relevance[37].

In our research we proposed a non-supervised calculation of direct dependency class feature selection algorithm that works for unsupervised dataset. Our future research will attempt to refine our method's memory management to handle sparse datasets better, the detail that it consider at that time. There are next few more problems related to this research to be addressed; We don't have sufficient former information regarding the bunch structure of the facts in Unsupervised Feature Selection. While some recent efforts have been made to analyze the stability of feature selection methods in the unsupervised contexts there is much effort to be done in this path.

One big problem in Unsupervised Feature Discovery is to pick the appropriate features for troubles when both numeric and non-numeric features (mixed data) define data at the same time. Mixed data is very popular and appears in many real life problems.. like, industry, software cost estimates ,in medicines and health care system etc.

However, as we saw in this study, the majority of the existing methods has been considered for mathematical figures only. Unsupervised Feature Selection Approaches for mixed results can also be developed.

Unsupervised feature selection base on rough set theory is able to generate very effective results. The dependencies of feature illustrate how relevant and significant their inclusion is for future analysis. In this thesis, we aim to generate feature subsets based on their dependency calculations for unsupervised data sets. The datasets used were mostly numerical/ integer based whereas preprocessing was done on feature with real attribute type. Datasets with other attribute

characteristics were not suitable for hid type of algorithm where calculations are more numerical type data. The algorithm can be further adapted for such datasets with diverse characteristics. There are several other unsupervised algorithms which aim to generate features base on rough set theory. There are other prospects for the extension of this research work where algorithm and proposed framework will be performed on more datasets and other techniques will be included for the purpose of feature extraction and optimization.

In future, the goal would be to compare the results with other unsupervised feature selection methods and apply the proposed modified algorithm on larger datasets.

## References

- [1] Robust Unsupervised Feature Selection. Mingjie Qian and Chengxiang Zhai.
- [2] J. A. Sáez, E. Corchado: KSUFS: A Novel Unsupervised Feature Selection Method Based on Statistical Tests.
- [3] P. Mitra, C. A. Murthy and S. K. Pal, "Unsupervised feature selection using feature similarity," in IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 24, no. 3, pp. 301-312, March 2002.[Duda et al.,][Huawei et al., 2011]
- [4] Raza, M. S., & Qamar, U. (2018). Feature selection using rough set-based direct dependency calculation by avoiding the positive region. International Journal of Approximate Reasoning, 92, 175-197.
- [5] J. Hua, D. T. Waibhav and E. R. Dougherty. "Performance of feature-selection methods in the classification of high-dimension data." Pattern Recognition 42.3 (2009): 409-424.
- [6] Zhong, N., Dong, J., & Ohsuga, S. (2001). Using rough sets with heuristics for feature selection. Journal of intelligent information systems, 16(3), 199-214.].
- [7] R. Jensen and Q. Shen. "Computational intelligence and feature selection: rough and fuzzy approaches." Vol. 8. John Wiley & Sons, 2008
- [8] E. Hancer et al. "A multi-objective artificial bee colony approach to feature selection using fuzzy mutual information". 2015 IEEE Congress on Evolutionary Computation (CEC). IEEE, 2015.].
- [9] P.A. Devijver, J. Kittler, Pattern Recognition: A Statistical Approach, Prentice-Hall, Englewood Cliffs, NJ, 1982
- [10] N. Zhong, A. Skowron, A rough set-based knowledge discovery process, Int. J. Appl. Math. Comput. Sci. 11 (3) (2001) 603–619.]
- [11] [K. Thangavel, A. Pethalakshmi, Dimensionality reduction based on rough set theory: a review, J. Appl. Soft Comput. 9 (1) (2009) 1–12
- [12] E. Alpaydin, Introduction to Machine Learning, 2nd edition, PHI, New Delhi, 2010
- [13] N. Dessì, B. Pes, Similarity of feature selection methods: an empirical study across data intensive classification tasks, Expert Sys. Appl. 42 (10) (2015) 4632–4642.
- [14] T.P. Hong, C.H. Chen, F.S. Lin, Using group genetic algorithm to improve

performance of attribute clustering, *Appl. Soft Comput.* 29 (2015) 371–378.

- [15] S. Paul, S. Das, Simultaneous feature selection and weighting – an evolutionary multi-objective optimization approach, *Pattern Recognit. Lett.* 65 (2015) 51–59.
- [16] K.O. Akande, T.O. Owolabi, S.O. Olatunji, Investigating the effect of correlation-based feature selection on the performance of support vector machines in reservoir characterization, *J. Nat. Gas Sci. Eng.* 22 (2015) 515–522.
- [17] I. Koprinska, M. Rana, V.G. Agelidis, Correlation and instance based feature selection for electricity load forecasting, *Know.-Based Sys.* 82 (2015) 29–40.
- [18] W. Qian, W. Shu, Mutual information criterion for feature selection from incomplete data, *Neurocomputi.* 168 (2015) 210–220.
- [19] M. Han, W. Ren, Global mutual information-based feature selection approach using single-objective and multi-objective optimization, *Neurocomput.* 168 (2015) 47–54.
- [20] M. Wei, T.W.S. Chow, R.H. Chan, Heterogeneous feature subset selection using mutual information-based feature transformation, *Neurocomput.* 168 (2015) 706–718.
- [21] M. Dash, H. Liu, Consistency-based search in feature selection, *Artif. Intell.* 151 (1) (2003) 155–176.
- [22] P. Moradi, M.A. Rostami, Graph theoretic approach for unsupervised feature selection, *Eng. Appl. Artif Intell.* 44 (2015) 33–45.
- [23] P. Moradi, M. Rostami, Integration of graph clustering with ant colony optimization for feature selection, *Knowl.-Based Sys.* 84 (2015) 144–161.
- [24] S.A. Bouhamed, I.K. Kallel, D.S. Masmoudi, B. Solaiman, Feature selection in possibilistic modeling, *Pattern Recognit.* 48 (11) (2015) 3627–3640.
- [25] M.L. Samb, F. Camara, S. Ndiaye, Y. Slimani, M.A. Esseghir, A novel RFE-SVM-based feature selection approach for classification, *Int. J. Adv. Sci. Tech.* 43 (2012) 27–36.
- [26] Z. Pawlak, A. Skowron, Rudiments of rough sets, *Information sciences* 177 (1) (2007) 3–27.
- [27] Z. Pawlak, Rough sets, *Int. J. Comp. Info. Sci.* 11 (1982) 341–356.
- [28] P.R.K. Varma, V.V. Kumari, S.S. Kumar, A novel rough set attribute reduction based on ant colony optimization, *Int. J. Intell. Sys. Tech. Appl.* 14 (3–4) (2015) 330–353.
- [29] C. Wang, M. Shao, B. Sun, Q. Hu, An improved attribute reduction scheme with covering based rough sets, *Appl. Soft Comp.* 26 (2015) 235–243.

- [30] X. Jia, L. Shang, B. Zhou, Y. Yao, Generalized attribute reduct in rough set theory, *Knowl.-Based Sys.* 91 (2016) 204–218.
- [31] Y. Kusunoki, M. Inuiguchi, Structure-based attribute reduction: a rough set approach, *Feat. Sel. Dat. Pattern Recognit.* (2015) 113–160.
- [32] H.H. Inbarani, A.T. Azar, G. Jothi, Supervised hybrid feature selection based on PSO and rough sets for medical diagnosis, *Comput. Meth. Prog. Biomed.* 113 (1) (2014) 175–185.
- [33] K. Zuhtuogullari, N. Allahvardi, N. Arikan, Genetic algorithm and rough sets based hybrid approach for reduction of the input attributes in medical systems, *Int. J. Innov. Comput. Info. Cont.* 9 (2013) 3015–3037.
- [34] W. Qian, W. Shu, B. Yang, C. Zhang, An incremental algorithm to feature selection in decision systems with the variation of feature set, *Chinese J. Elect.* 24 (2015) 128–133.
- [35] Y. Chen, Q. Zhu, H. Xu, Finding rough set reducts with fish swarm algorithm, *Knowl.-Based Syst.* 81 (2015) 22–29.
- [36] H.H. Inbarani, M. Bagyamathi, A.T. Azar, A novel hybrid feature selection method based on rough set and improved harmony search, *Neural Comput. Appl.* 26 (8) (2015) 1859–1880.
- [37] Dy and Brodley 2004 Dy JG, Brodley CE (2004) Feature selection for unsupervised learning. *JMach Learn Res* 5:845–889
- [38] de Amorim, R. C. (2019). Unsupervised feature selection for large data sets. *Pattern Recognition Letters*, 128, 183-189
- [39] Das, A. K., Sengupta, S., & Bhattacharyya, S. (2018). A group incremental feature selection for classification using rough set theory based genetic algorithm. *Applied Soft Computing*, 65, 400-411
- [40] B. Hans-Hermann , Origins and extensions of the k-means algorithm in cluster analysis, *J. Electron. dHistoire des Probabilités et de la Statistique Electron. J. History Probab. Stat.* 4 (2) (2008)
- [41] A. Jain, Data clustering: 50 years beyond k-means, *Pattern Recogn. Lett.* 31 (8) (2010) 651–666, doi: 10.1016/j.patrec.2009.09.011
- [42] B. Mirkin , *Clustering: A Data Recovery Approach*, Computer Science and Data Analysis, CRC Press, London, UK, 2012 .
- [43] D. Steinley , K-means clustering: a half-century synthesis, *Br. J. Math. Stat. Psy- chol.*

59 (1) (2006) 1–34 .

[44] K. Zuhtuogullari, N. Allahverdi and N. Arikan. “Genetic Algorithm and Rough Sets Based Hybrid Approach for Reduction of the Input Attributes in Medical Systems.” *International Journal of innovative computing and information control* 9 (2013) 3015-3037.

[45] Q. Wenbin, W. Shu, B. Yang and Z. Changsheng, “An Incremental Algorithm to Feature Selection in Decision Systems with the Variation of Feature Set.” *Chinese Journal of Electronics*, 24 (2015) 128-133.

[46] Jothi, G. (2016). Hybrid Tolerance Rough Set–Firefly based supervised feature selection for MRI brain tumor image classification. *Applied Soft Computing*, 46, 639-651

[47] Anaraki, J. R., & Eftekhari, M. (2011, May). Improving fuzzy-rough quick reduct for feature selection. In 2011 19th Iranian Conference on Electrical Engineering (pp. 1-6). IEEE

[48] P. Cunningham. “Dimension Reduction.” University College Dublin, Technical Report, 2007.

[49] b. Tang, S. Kay and H. Haibo. "Toward optimal feature selection in naive Bayes for text categorization." In: *IEEE Transactions on Knowledge and Data Engineering* 28.9 (2016): 2508- 2521.

[50] F. Jiang, S. Yuefei and Lin Zhou. "A relative decision entropy-based feature selection approach." *Pattern Recognition* 48.7 (2015): 2151-2163.

[51] D. Singh et al. "Feature Selection Using Rough Set For Improving the Performance of the Supervised Learner." *International Journal of Advanced Science and Technology* 87 (2016): 1-8.

[52] J. Xu et al. "L1 graph based on sparse coding for feature selection." *International Symposium on Neural Networks*. Springer Berlin Heidelberg (2013): 594–601.

[53] N. Hamdi, K. Auhmani and M. M. Hassani. “Quantum Clustering-Based Feature Subset Selection for Mammographic Image Classification.” *International Journal of Computer Science & Information Technology* 7.2 (2015): 127-133

[54] G. Roffo, S. Melzi and M. Cristani. "Infinite Feature Selection." In: *Proceedings of the IEEE International Conference on Computer Vision* (2015): 4202-4210.

[55] H. Almuallim and T.G. Dietterich. Learning with many irrelevant features. In the 9th National Conference on Artificial Intelligence, MIT Press, (1991): 547–552.

[56] B. Raman and R. I. Thomas. "Instance-based filter for feature selection." *Journal of*

Machine Learning Research 1.3 (2002): 1-23.

[57] H. Liu, H. Motoda, Computational Methods of Feature Selection, Chapman & Hall/Crc Data Mining and Knowledge Discovery Series, 2007.

[58] L. Du and S. Yi-Dong. "Unsupervised Feature Selection with Adaptive Structure Learning." In: Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (2015): 209-218.

[59] J. Li et al. "Unsupervised Streaming Feature Selection in Social Media." In: Proceedings of the 24th ACM International on Conference on Information and Knowledge Management (2015): 1041-1050.

[60] D. A. A. G. Singh, S. A. A. Balamurugan and E. J. Leavline. "An unsupervised feature selection algorithm with feature ranking for maximizing performance of the classifiers." International Journal of Automation and Computing 12.5 (2015): 511-517.

[61] P. Zhu et al. "Unsupervised feature selection by regularized self-representation." Pattern Recognition 48.2 (2015): 438-446.

[62] N. Zhou et al. "Global and local structure preserving sparse subspace learning: an iterative approach to unsupervised feature selection." Pattern Recognition 53 (2016): 87-101.

[63] X. He, D. Cai and P. Niyogi. "Laplacian score for feature selection." Advances in Neural Information Processing Systems 18 (2005): 21-26.

[64] M. Devaney and R. Ashwin. "Efficient feature selection in conceptual clustering." In: Proceedings of the Fourteenth International Conference on Machine Learning (1997): 92-97.

[65] M. Gluck. "Information, uncertainty and the utility of categories." In: Proceedings of the Seventh Annual Conf. on Cognitive Science Society (1985): 283-287.

[66] Z. Pawlak , Rough Sets: Theoretical Aspects of Reasoning About Data, Kluwer Academic Publishers, Boston, 1991.

[67] Y. Huang , T. Li , C. Luo , H. Fujita , S. Horng , Matrix-based dynamic updating rough fuzzy approximations for data mining, Knowl. Based Syst. 119 (2017) 273–283.

[68] S. Lun, Z. Xiaoyu, X. Jiucheng, W. Wei, L. Ruonan, A gene selection approach based on the fisher linear discriminant and the neighborhood rough set, Bio- engineered (2017).

[69] G. Lang , D. Miao , M. Cai , Z. Zhang , Incremental approaches for updating reducts in dynamic covering information systems, Knowl. Based Syst. 134 (2017) 85–104]

[70] C. Luo , T. Li , H. Chen , H. Fujita , Z. Yi , Incremental rough set approach for hier-

archical multicriteria classification, *Inf. Sci.* 429 (2018) 72–8.

[71] D. Kim , Data classification based on tolerant rough set, *Pattern Recognit.* 34 (8) (2001) 1613–1624.

[72] W.Z. Wu , W.X. Zhang , Neighborhood operator systems and approximations, *Inf. Sci.* 144 (14) (2002) 201–217.

[73] Q.H. Hu , D. Yu , J.F. Liu , C. Wu , Neighborhood-rough-set based heterogeneous feature subset selection, *Inf. Sci.* 178 (18) (2008) 3577–3594.

[74] P. Zhu , Q.H. Hu , Adaptive neighborhood granularity selection and combination based on margin distribution optimization, *Inf. Sci.* 249 (2013) 1–12 .

[75] H. Zhao , P. Wang , Q.H. Hu , Cost-sensitive feature selection based on adaptive neighborhood granularity with multi-level confidence, *Inf. Sci.* 366 (2016) 134–149.

[76] Y. Chen , Z. Zhang , J. Zheng , Y. Ma , Y. Xue , Gene selection for tumor classification using neighborhood rough sets and entropy measures, *J. Biomed. Inform.* 67 (2017) 59–68.

[77] K. P. Bennett and O. L. Mangasarian: "Robust Linear Programming Discrimination of Two Linearly Inseparable Sets", *Optimization Methods and Software* 1, 1992, 23-34.

[78] Zhu, C., Idemudia, C.U. and Feng, W., 2019. Improved logistic regression model for diabetes prediction by integrating PCA and K-means techniques. *Informatics in Medicine Unlocked*, 17, p.100179.

[79] Thomas, J., Joseph, A., Johnson, I. and Thomas, J., 2019. Machine Learning Approach For Diabetes Prediction. *International Journal of Information*, 8(2).

[80] Gharibdousti, M.S., Haider, S.M., Ouedraogo, D. and Susan, L.U., 2019. Breast cancer diagnosis using feature extraction techniques with supervised and unsupervised classification algorithms. *Applied Medical Informatics.*, 41(1), pp.40-52.

[81] Shaw, B., Suman, A. K., & Chakraborty, B. (2019). Wine Quality Analysis Using Machine Learning. *Emerging Technology in Modelling and Graphics*, 239–247. doi:10.1007/978-981-13-7403-6\_23 .

[82] Mujumdar, A. and Vaidehi, V., 2019. Diabetes Prediction using Machine Learning Algorithms. *Procedia Computer Science*, 165, pp.292-299.

[83] Moran, M. and Gordon, G., 2019. Curious feature selection. *Information Sciences*, 485, pp.42-54.

[84] Mollas, I., Tsoumakas, G. and Bassiliades, N., 2019. LionForests: Local Interpretation



of Random Forests through Path Selection. arXiv preprint arXiv:1911.08780.

[85] Chauhan, H., Kumar, H. and Mittal, S., 2020. Breast Cancer Classification using Random Forest Classification. *Our Heritage*, 68(30), pp.7867-7872.

[86] Aich, S., Al-Absi, A.A., Hui, K.L., Lee, J.T. and Sain, M., 2018, February. A classification approach with different feature sets to predict the quality of different types of wine using machine learning techniques. In 2018 20th International Conference on Advanced Communication Technology (ICACT) (pp. 139-143). IEEE.

[87] Salvador–Meneses, J., Ruiz–Chavez, Z. and Garcia–Rodriguez, J., 2019. Compressed kNN: K-nearest neighbors with data compression. *Entropy*, 21(3), p.234.

[88] Das, A.K., Sengupta, S. and Bhattacharyya, S., 2018. A group incremental feature selection for classification using rough set theory based genetic algorithm. *Applied Soft Computing*, 65, pp.400-411.

## Completion Certificate

*It is certified that the contents of thesis document titled “Unsupervised Feature Selection Based on Rough Set Theory Using Direct Dependency Classes (DDC)” submitted by Ms. Saliha Hanif with Registration No. 00000172268 have been found satisfactory for the requirement of degree.*

*Thesis supervisor: \_\_\_\_\_*

*(Dr. Usman Qamar)*



*Thesis Co- supervisor: \_\_\_\_\_*

*(Dr. Summair Raza)*