International Journal of Advanced Science and Technology
Vol.129 (2019), pp.29-42
http://dx.doi.org/10.33832/ijast.2019.129.03

APPLICATION OF ROUGH SET THEORY IN MEDICAL HEALTH CARE DATA ANALYTICS

Indrani Kumari Sahu[1*], G K Panda[2] and Susant Kumar Das[3]

[1]*Research Scholar, Dept. of Comp. Sc., Berhampur University, India*
[2]*Professor, MITS School of Biotechnology, Utkal University, India*
[3]*P.G. Dept. of Comp. Sc., Berhampur University, India*
[1*]indranisahu@gmail.com, [2]drgkpmail@gmail.com, [3]dr.s.k.das.1965@gmail.com

***Abstract*—Rough Set theory (RST) is a mathematical tool and used to deal with vagueness, impreciseness, inconsistence and uncertain type knowledge. RST-based research has been applied in machine learning, inductive reasoning, decision support systems and knowledge discovery applications. Popular methods like finding of reducts, core, feature selection and reduction through the concepts of approximations have attracted researchers to use RST further in the field of high dimensional data like social networks, IoT applications and Big data analytics. In this article we make an attempt to summarize the basic concepts, characteristics of RST, some evolutionary extensions of RST and applications limited to Medical data analysis. In keeping the view of learners, a survey on RST based software tools and packages outlined with their exhaustive functionalities. It also identifies the importance of RST in the domain of medical or clinical data analytics, and also exhibits the strengths and limitations of the respective underlying approaches.**

***Keywords*— Rough Set, Reduct, Core, Medical Data Analytics, Clinical Dataset**

1. INTRODUCTION

Smart Healthcare system is a new healthcare paradigm and evolving with high expectations where physicians, professionals, integrators, stake holders and researchers across the globe are looking for cost-effective, innovative and technology-driven collaborative solutions to the patient community. The complement of high configured, lightweight and low-cost intelligent bio-sensors has also given a best choice for mobile medical diagnosis systems and mobile health monitoring systems. IoT based such sensors are capable of sensing, processing and communicating vital observatory signs of patients during diagnosis, examinations and alert processes into a network of IoT based computer system.

Today's medical system facilitate one layer solution to both caregivers and consumers in adopting applications like, doctor-on-demand, video calls with doctors, Wi-Fi-enabled blood pressure monitors etc. The back-end of such system is powered with a strong base of cloud infrastructure with real time computing techniques that integrates the right operational and clinical decision makers at the right time. These processes produce large quantities of data, in real time. The recent adoption of electronic health records (EHR) and electronic medical records (EMR) systems also taking driver's role in making sense of huge collected data from prescriptions, diagnostic tests, patient care records and insurance claims.

Received: April 25, 2019
Reviewed: June 9, 2019
Accepted: June 28, 2019
* Corresponding Author

ISSN: 2207-6360 IJAST

Thus, new generations Smart Health Care system generate a high volume of data, every second of the day but not all information stored by health care organizations is useful. This is due to there are no adequate resources, tools and skills to integrate such clinical/medical data into analytics. Research on knowledge acquisition from datasets has been intended to identify the functions of non-trivial extraction of implicit, previously unknown and potentially useful information. Data mining, rule extraction, constraints and finding of regulations of regularities from data are some of the major outcomes.

Rough Set Theory (RST) [34] introduced by Professor Zdzislaw Pawlak is one of the well known classification and knowledge discovery techniques and has pulled in much enthusiasm and attention from researchers. This is due to its strength in dealing with wide variety of application areas like, classification, clustering, fault detections, plant control. RST is considered as one of the first non-statistical approaches in data analysis. RST and its variances have been used to cater the problems on representation of vague-type knowledge (data with uncertainity/impreciseness), analysis of information/knowledge/ reducts, consistent evaluation of qualitative information, attribute reduction, rule acquisition, intelligent algorithms.

The extent of RST applied nowadays is much extended than in the past, principally in the areas of Artificial Intelligence, cognitive sciences, machine learning, decision analysis, expert systems, inductive reasoning and pattern recognition, engineering, banking, financial and market analysis, process control, social networks and many more. One common aspect among all these applications is to discard the incomplete/imprecise data from the source, which would shrink the information sphere. Hence, it is essential to devise an appropriate data processing and knowledge acquisition tool in catering these issues. In [3][4][7][34][44], authors discussed data processing tools for computational intelligence and reasoning systems. In [34], Pawlak's approach gave new dimension in this regard. In [17][24] authors developed RST-based more computationally efficient methods in addressing these issues. These are popularly termed as "pattern discovery from databases", "formulation of decision rules", "reduction of data", "principal component analysis", "inference interpretation based on available data". Software tools developed for formal modeling, reasoning and computing and evolved from the conventional approach of deterministic or precise type to impreciseness, due to nondeterministic nature of some real life applications. In this paper we explore some RST enabled software tools/packages being used in general to automate the above processes/analysis.

In Section 2 we discuss basic concepts of RST. In machine learning environment, classification solves many problems including the problem of identifying the category to which a new data point belongs. In real dataset analysis, sometimes it becomes difficult to classify items automatically. In Section 3 we discuss extensions of RST and some generalizations. In Section 4, we discuss some software packages available for RST analysis. Section 5 emphasizes on RST in clinical datasets. Finally, we end up with a section on concluding remarks followed by a bibliography.

2. BASICS OF ROUGH SET THEORY

Objects that possess similar characteristics happen to be a common set. In real life datasets, objects are termed to be crucial and rough information is associated with each object through associated data. The dataset contains a set of conditional attribute value sets. In medical systems such datasets may differ for different types of observations. Based on the available information, it is possible to classify and make conclusion that some of the objects are apart, while others are impossible to distinguish. It is hard to indiscernible such objects and possible to represent as a set. Such set is termed as a *"knowledge granule"* that plays a key role in building the blocks of knowledge in relational datasets.

Dempster-Shafer Theory, Fuzzy set theory and RST are some of the most common approaches used in decision making systems. The first two methods use probability assignments and membership values where approximations and granularity structure of the data are used in RST. Many conventional methods use the threshold information to operate the dataset whereas RST does not need it. Moreover, the embedded uncertainty in real data is handled through the approximations of RST.

2.1. EQUIVALENCE RELATION

In simple mathematical term, it is a binary relation which satisfies reflexive, symmetric and transitive properties. In RST, knowledge is related to the classification ability of objects and deeply associated with the patterns of classification with respect to the specificity of *'U'* (Universe of Discourse). A group of such classifications over *'U'* is termed as the "knowledge-base over *U*, *K=(U, R)*". *'U/R'* is said to be "*group of equivalence class"* of *'R'*, if *'R'* is an equivalence relation over *'U'*. *'[x]*$_R$*'* is a category in the '*equivalence relation*' having an element *'x'* which in turn an element of the Universe.

2.2. INDISCERNIBILITY

A relation *IND(A)* is said to be an *Indiscernibility* if it is associated with two or more objects and values of the objects are identical in relation to a particular attribute or more attributes. Indiscernibility is used to find minimal representative subset of the attributes from the application dataset. In other words, if $A \subseteq R$ and, $A \neq \phi$ then, $\bigcap A$ is the Indiscernibility relation of *A*.

2.3. APPROXIMATIONS

In the intersection domain, the "*approximation of specs*" refers the formal classification of knowledge/information. It is due to the availability of granularity of information in the knowledge-base, RST cannot be characterized directly from such knowledge. Hence, there are two approximations (crisp sets) associated with each rough set 'S'. It also associate with a partition which cannot be classified uniquely to the set or its complement. Given any set $S \subseteq U$ and $R \in IND(A)$, the two types of approximations and the region that cannot be classified are:

- *Lower approximation* describes a positive region for the selected target *'S'*. The region includes the group of objects that can be classified as belonging to the desired target of concepts *'S'* with full certainty.

$$\underline{R}S = \bigcup \{T \in I(R) : T \subseteq S\}$$

- *Upper approximation* describes about the objects that possibly belong to the subset-of -interest.

$$\overline{R}S = \bigcup \{T \in I(R) : T \cap S \neq \phi\}$$

- *Rough-boundary* describes about the imperfect knowledge, which is the partition of objects and cannot be classified, as belonging to S and -S.

$$B_R(S) = \overline{R}S - \underline{R}S$$

- *Positive/Negative Region*

$$POS_P(Q) = \bigcup_{X \in U/Q} \underline{R}S, \qquad NEG_P(Q) = U - \bigcup_{X \in U/Q} \overline{R}S$$

If $\underline{R}S \neq \overline{R}S (\text{or } B_R(S) \neq \phi)$ then '*S'* is said to be '*rough over R'*. If, $\underline{R}S = \overline{R}S$ $(or\ B_R(S) = \phi)$ then '*S*' is '*Rough-definable'*.

2.4. ACCURACY OF COEFFICIENT

Accuracy and error approximations induce the effect of calculated values. In this case, accuracy coefficients like imprecision and quality are defined as below.

Imprecision coefficient $$\alpha R(S) = \frac{|\underline{R}(S)|}{|\overline{R}(S)|}$$

Quality coefficient $$\alpha R((\overline{R}(S)) = \frac{|\overline{R}(S)|}{|TotalObjects|}$$

2.5. ROUGH DEPENDENCY

Attribute dependency plays a vital role in data analysis. It becomes challenging to find the dependencies between two and more attributes, automatically in a data set. In a data set if *A* and *B* are two sets of attribute(s), *B* can depend on *A (A=>B)* i.e., if all values of *B* are uniquely determined by values from *A*. There associate three means of the degree of dependency (*'k'*). If (*k=0*), it represents as *B* doesn't depend on *A*, if (*k=1*), it represents *B* fully depends on *A* and if (*0 < k < 1*), it represents that *B* partially depends on *A*. The degree of dependency '*k*' is:

$$k = \gamma(A, B) = \frac{|POSc(B)|}{|U|}$$

$$where,\ POSc(B) = \bigcup_{X \in U/B} \underline{A}(X)$$

'$POS_c(B)$' is the *"positive region"* of the partition *'U/B'* with respect to *'A'*. It has all objects of *'U'* that can be classified uniquely to the blocks of the above partition.

2.6. REDUCTS

Decision centric information system contains more than one conditional attributes, where all conditional attributes do not signify same importance, rather there are some important attribute(s) which signify the knowledge in the equivalence class structure leaving aside other attributes. The subset of attributes that fully characterize the knowledge is known as *'reduct'*. An ideal reduct justifies the same classification accuracy as that of the entire dataset. Hence reduct offers the best choice in optimizing the data usage complexity.

Mathematically, an attribute subset will be a reduct if:

$$\gamma(A, B) = \gamma(A', B) \quad for\ A' \subseteq A$$

i.e., for decision class 'B', an attribute-set $A' \subseteq A$ will be a 'reduct' if the dependency of 'B' on 'A' is equal to the entire set of conditional-attributes.

3. SOME EXTENSIONS/GENERALIZATION OF RST

Classical RST is fully dependent on discrete data, where some real-life applications contain real-value data which made significant drawback in the successful operations of classical RST. In medical or clinical datasets, the diagnosis objective must not be limited to discrete values like 'yes (1)' or 'no (0)'. This is due to the fact that, these discrete

values may not be most-appropriate to derive for a conclusion in terms of decision making with high degree of accuracy. In contrary, screening tests or medical surveillance exhibits its significance to access the chances of symptoms having a particular disease. For example, *"pap smear"* for cervical cancer, *"mammography"* for breast cancer, *"PSA"* for prostate cancer, *"Cholesterol level"* for heart disease and many more. The results of these tests depend not only on the technical parameters of the test, but also on its sensitivity and specificity. So, sometimes decision makers derive with a subjective decision from the set of objective measured data-sets. To be more specific, attributes like *'Blood Pressure'*, *'Blood Glucose level'* etc., are quantified as "Normal", "High", "Low" for establishing boundaries for the measured values.

Many researchers have been extending or generalizing the classical RST as per the application diversity with a general aim of handling the *"inconsistent data"* or *"uncertainty"* in the data set. Such criteria generally happens when we do not have enough information to describe the underlying concept in order to be able to say that, *"which object belongs to the desired class with certainty"*. In Table-I we present some extensions to Classical RST. It is observed that, initial extensions focused on the aspects of similarities and difference factors of relationship.

Table I. Extension and Generalization of Rough Set

Methods	Year	Authors	Purpose
Decision-theoretic Rough Sets	1990	Yao et al. [41]	DTRS is a probabilistic extension of RST. This method uses Bayesian decision procedure to find the minimum risk in the process of decision making system. In the process, all elements are partitioned in two approximations (lower/upper) according to the conditional probability (CP). The CP is measured through a threshold value (α and β) that help to find the region for inclusion of the elements. DTRS is said to be strong method because the threshold parameters like α and β are calculated automatically using a set of 6 loss functions exhibiting the classification risk.
Fuzzy Rough Sets	1992	Dubous et al. [11]	FRS handles real-valued data. Fuzzy equivalence classes are employed to cater the real valued data.
Variable Precession Rough Sets	1993	Ziarko et al. [45]	VPRS is one of the generalizations of RST which uses a "*controlled-degree*" of mis-classification in relaxing "*subset operator*" in the method.
Tolerance Rough Sets	1996	Skowron et al. [38]	TRS is an extension of RST which handles real-valued data. In TRS, similarity relations are used instead of Indiscernibility relations. This is confined with a "*limited degree of variability*" in attribute-values.
Neighborhood Rough sets	1988 2001	Lin et al. [25] [26]	This is one of the generalizations of RST. This method deals with numerical data in updating the approximations of any concept with 24 neighborhood operators.
Dominance based Rough sets	2001	Greco et al. [15]	DRS is an extension of RST and used for "*multi-criteria decision*" analysis. In this method, the equivalence relation is replaced by dominance relation and outperforms the inconsistencies.
Covering based Rough Sets	2002	Zhu et al. [46]	CBRS is an extension to RST where covering approximation spaces (CAS) are used which is termed as a generalization of equivalence-based RST. The CAS comprises of 4-types of "*covering lower approximation operators*" and 3-types of "*covering upper approximation operators*". Thus CBRS has 12 types covering based approximation operators.
Game-theoretic Rough Sets	2011	Herbert et al. [20]	GTRS is an extension of RST. This method uses game-theory based concept to find an effective size of region for inclusion by optimizing the criteria set in the classification/decision making system.

4. TOOLS AND SOFTWARE SYSTEMS FOR RST

Software tools play a major role in validating experimental researches, designed concepts and algorithms. In decision making system too, researchers like to use such software tools for thorough and multi directional investigations so as to focus on the most

essential matters of their study. Academicians and scientists across the globe have already put their effort in designing such software tools in the domain of data mining, knowledge discovery, rule inductions, feature/instance selection, missing value completion, decomposition, cross validation, nearest-neighborhood classification and many more operational concepts of RST. In this section we discuss such commonly used software tools/ packages concerned to RST application. Table-II presents an overview of some software or tools with their source of availability, mainly designed on the above domains.

Table II. RST Based Software Systems/ Tools

Tools	Source	Description	Advantages
WEKA (1993)	www.cs.waikato.ac	WEKA tool represents for "Waikato Environment for Knowledge Analysis", and developed at "University of Waikato, New Zealand".	It is compatible with Java and easy to process some of machine learning tasks like, 'association', 'regression', 'visualization', 'data processing', 'classification', 'clustering'. Researchers find it to use for machine learning analysis.
RSES (1994)	http://logic.mimuw.edu.pl/~rses/	RSES represent as "Rough Set Exploration System", that facilitates RST operations for Data analysis.	Some of the important operations being facilitated in this tool are, Data set fragmentation (Decompose large dataset into smaller units satisfying similar properties), Attribute Discretization (for numerical data), Reduct Discovery, Identification of hidden patterns in data, Identification of decision rules, etc.
TAS (1994)	Warsaw Univ. of Tech.	TAS is based on RST and Petri nets and used for concurrent processing.	TAS provides utilities like, identification of concurrent models from user input data-tables and finding of decision making based parallel programs.
PRIMEROSE (1994)	http://www.cardiff.ac.uk/biosi/research/biosoft/	PRIMEROSE tool supports RST methods to find Probabilistic Rule inductions.	This tool supports for clinical databases and in general it is used for rule-extraction for expert systems.
R (1997)	www.r-project. org	Scripting interface support RST features	compatible with C/C++, Java, Python
Rosetta (1998)	www.lcb.uu.se http://rosetta.lcb.uu.se/general/	ROSETA represents for "RST Toolkit for Analysis of data". It is used to make dataset-analysis in tabular form using basic operations of RST.	Some of its supporting functions are used for, "rule induction","Discretization","clustering","classification", "rule pruning", "classifier evaluation", "reduct identification", and "knowledge discovery". It is also ccompatible with languages like, C++, Python and Perl scripts.
ROSE2 (1998)	http://idss.cs.put.poznan.pl/site/60.html	ROSE2 represents as "Rough Sets Data Explorer". It is successor software to "Rough DAS" & "Rough Class systems".	Some of its important functions are, "Rule/ Knowledge Discovery", "Discretization", "Data reduction by Core/ Reducts", "Approximation based Decision Rules" and "Classifications".
PROB ROUGH	--	A System for Probabilistic Rough Classier Generation.	This tool is used to Classify rules with probabilistic approach.
4eMka	http://idss.cs.put.poznan.pl/site/60.html	It is an implementation of multiple criteria decision support in combination with RST and dominance relation.	The tool facilitates for "Multi Criteria based classification", "Rule Extraction", "Data set partition" using the popular methods of Dominance based RST.
JAMM	http://idss.cs.put.poznan.pl/site/60.html	It is a decision support tool and used for multi criteria based classification problems.	This tool supports for "Minimal Cover/ Complete set of Rules", "Decision rules (classifier)" and used in the applications of finance, medicine, geology, pharmacology etc.

4.1. RST ALGORITHMS

The basic objectives of the most of the packages are to deal with data handling activities (upload/down load), data structure derivation, addition and deletion of object

processes, how to set decision attributes, how to calculate statistical information about data etc. These RST based operations or processes are performing through a set of algorithms, which seems to be the most essential part of the software packages. We discuss some such algorithms, as follows:

- ***Reduction Algorithm***

Reduction algorithms calculate the required *'reducts'* of an input decision table or information system. In addition to calculating number of reducts, it also performs the solutions pertaining to approximate and heuristics like *"genetic"*, *"covering"* and *"Johnson"* algorithms.

- ***Rule induction Algorithm***

Decision rules are framed from such set of calculated and available reducts. In order to calculate rules, user has to go through a procedure to find constraints like accuracy, coverage etc., for the set of decision rules. Further, these rules need to be accompanied with set of coefficients in order to be used, finally to the group of objects.

- ***Discretization Algorithm***

Discretization is the process to reduce the attribute complexity of the input dataset. In this process, the input decision table is allowed to convert into a simplified format using symbolic attributes. By this process, the complexity of the table reduces in preserving the original information concern to the *'disscernibility'* of objects.

- ***Algorithms for Template Generation***

Template generation algorithms provide user to add new attribute(s), to calculate number of templates and to find generalized templates. Table decomposition processes are adopted in this process.

- ***Classification Algorithm***

Classification algorithms provide user to find decision value for objects. In this process, already generated decision rules and/or templates are used.

5. ROUGH SET THEORY IN HEALTH CARE

Clinical dataset in general, consists of information about the current condition of the patients. Mining knowledge from clinical dataset refers to the discovery of hidden valuable knowledge and to develop clinical expert system [23]. The repository systems of datasets are huge with high dimensional attributes. In order to discard redundant information from database, RST processes provide the most feasible solution. The process filters to a minimum number of attributes that would still allow each data record to be distinguished from the others and in RST it is termed as *"reduct"*.

Delivering quality healthcare is one of the most complex businesses in this present era. In the last two decades' healthcare domain has seen major paradigm shift in the way pharmaceuticals, diagnostics and healthcare IT systems have evolved, which in turn benefitted patients in terms of shorter length of stay in the hospitals and reduction in the cost of treatment. Today, we are talking of personalized healthcare and home-based-care, which its base on evidence has based and targeted treatment.

One of the most impactful change which we all can observe is that of usage of social media, You Tube, WhatsApp, etc., for the communication purpose with care givers and patients by the clinicians. Communicating through e-mail has become bygone era's mode of communication. Smart devices such as mobile phones, i-pads and other hand held devices are also used by the clinicians to source the information as well as to spread the information to the patients.

Health care analysis depend on the result of collected data from various sources like, *"clinical data"* (collected from electronic medical records (EHRs) from hospitals), *"patient behavior and sentiment data"* (patient preferences and behaviors), *"retail*

purchases" (*e.g.*, captured data from stores), *"pharmaceutical R&D data"* and *"claims and insurance data"*. As medical diagnosis is a data-centric task which needs minute analysis on symptoms with high accuracy and thus the collected dataset need for classifications as par with applications of data mining.

At the same time, some clinical information systems/ datasets contain high degree of ambiguity. Applications of RST to medical diagnosis have been very successful because RST offer tools which are very efficient in handling ambiguity in data. In Table-III we present some studied contributions about health care, specifically RST in medical or clinical information datasets. The thrust area of our review is further limited on three basic concepts of RST: knowledge analysis, knowledge discovery and classification/attribute reduction.

Out of 52 papers we refer 29 articles, published during 2002 and 2017. Some of the applications under Knowledge analysis of our study are, *"EEG Signal"*, *"Headache/Meningitis", "CVDDatabase", "peritoneal lavage in acute pancreatitis", "medical–image", "urolithiasis patients treated by extra-corporeal shock wave lithotripsy", "factors affecting the occurrence of breast cancer among women treated in US military facilities", "factors affecting the differential diagnosis between viral and bacterial meningitis", "therapeutic experience with acute pancreatitis"* and *"Surgical wound infection")*.

Secondly, applications under Knowldege discovery of our study are, *"Knowledge acquisition in nursing"* and *"Discovery of attribute dependencies in experience with multiple injured patients".*

Lastly, applications under classification/attribute reduction of our study are: *"Classification of histological pictures", "Supporting of therapeutic decisions", "Diagnosis of pneumonia patients", "Diagnosing in progressive encephalopathy", "Automatic detection of speech disorders", "Rough set-based filtration of sound applicable to hearing prostheses", "Modelling cardiac patient set residuals", "Breast cancer detection using electro-potentials"* and *"Attribute reduction in a database for hepatic diseases".*

Table III. RST in Health Care

Application field	**Author-Reference**	**Studied Contribution**	**Limitations/ Strengths**
"Rough Set Classifiers from Gene Expressions and Clinical Datasets" (2002)	Midelfart et al.[30]	According to the authors, genetic content of samples has high dimensional data which need to be reduced intelligently in order to discover useful information from it. RST based feature reduction and classification rule discovery methods applied to discover from genetic samples.	Their approach was applied on gastric-tumors dataset. All the developed-classifiers satisfying six clinical parameters were validated biologically.
"A New Application of Rough Set to ECG recognition" (2003)	Xian-Ming et.al.[42]	Authors used RST to reduce the recognition rules for certain points in ECG. These rules outperform for smooth recognition for ECG.	Authors used 'MIT-BIH' data to verify 'R- wave' recognition. The detection rate is higher than routine recognition method.
"Rough set based clinical decission model" (2004)	Farion et al. [12]	A decision model designed to assist inexperienced physicians. RST based such model supports the diagnosis by distinguishing between 3 disposition categories *'discharge', 'observation/further*	This model seems to be dependent and useful for qualitative data analysis. The model also gives accuracy in decision making comparable to decisions by physicians.

		investigation' and *'consult'*.	
"Extracting Protein-protien intera-ction sentences by applying Rough set data analysis" (2004)	Ginter et al. [14]	Authors used RST to extract decision rules of '*protein names*', '*interaction words*' and their '*mutual positions in sentences*'. To increase the potential interaction words, they modeled with 'spelling' and 'inflection variants'.	Their method uses one thousand eight hundred ninety four sentences of hand-tagged dataset. The method also passes through a Precision-recall break even performance of 79.8% with 'leave-one-out' 'cross-validation'.
"Classification of calculated electroence-phalogram parameters for detection of intraoperative awareness" (2004)	Ningler et al. [33]	This study employs RST and generates classification rules in comparing Crisp and fuzzy discretization values. The generated rules were allowed for further classifications.	Result accuracy rates of approx 90%.
"Rough set approach for attribute reduction and rule generation: A case of patients with suspected breast cancer" (2004)	Hassanien et al. [19]	This study is based on breast cancer dataset that finds reducts and classification rules concern to the dataset.	The model uses 360 samples and with the use of the Simplification Rule algorithm the rules were reduced to 30 from 428.
"Rough set approach to medical diagnosis system" (2005)	Grzegorzllczuk et.al.[16]	This study is based on Heart Disease and applied drug.	Finds accuracy in Decision classification, in generating if-then decision rules.
"Applying rough set theory to medical diagnosing" (2007)	Piotr et al. [36]	Diagnose mitochondrial encephalomyopathies (MEM) in child.	In using RST-support decision making, the time of diagnosis is shorten to 50%.
"On Medical Image Filtering Based on RST" (2008)	Yi Xie [43]	This study uses RST to de-noise image and classical process for noise removal an; in filtering subsets.	The method is flexible; reduces the noise, improves the PSNR(peak signal/ noise ratio).
"A Study on Rough Set Theory for Medical Image Segmentation" (2009)	Senthil kumaran et al. [37]	The main contribution in this study is the application of RST in medical image segmentation.	Reviewed various image segmentation techniques and highlighted ideas for future research.
"A framework for intelligent medical Diagnosis using RST with formal Concept analysis" (2011)	Tripathy et al. [40]	This study minimizes the rules using RST based rule generation algorithm for heart disease based attributes.	As, it is cumbersome to deduce decision from a lengthy set of rules, the study discovers suitable rules to identify the chief characteristics, so as to reduce execution time.
"A Support Vector machine classifier with Rough Set based feature selection for breast cancer diagnosis" (2011)	Chen et al. [5]	The main contribution in this study is to use RS_SVM (hybrid approach of RST and SVM) to classify the dataset and used for breast cancer diagnosis.	They use subset of 5 features and 3 types of training-test partitions (50-50 %age, 70-30 %age and 80-20 %age). Classification accuracy for the first type partition is 99.41 %age and for others 100 %age.
"Application of rough set classifiers for determining hemodialysis adequacy in ESRD patients" (2012)	Chen Y S et al. [6]	This study discovers knowledge from the HAD data set using RST and identifies the urea-reduction ratio for assessing HD-adequacy for ESRD patients and their doctors.	The model is designed to find good relationship between patient and medical in fulfilling the best quality.
"Dynamic context adaptation for diagnosing the heart disease in healthcare environment using optimized RST approach" (2012)	NaliniPriya et al. [32]	This study is based on the diagnosis of heart disease and also known as CAD (coronary artery disease).	Rough set provides a good clarity around 90% over the incomplete data set compared to fuzzy set.

"Color image segmentation using rough set based k-means algorithm" (2012)	Halder et al. [18]	This study uses '*neighborhood relationship*' and '*intensity-information*' with an aim to find an efficient segmentation of image.	In testing some real-time applications, this method lacks the improvements of speed.
"Unsupervised leukocyte image segmentation using rough fuzzy clustering" (2012)	Mohapatra et al.[31]	This method offers to identify '*leukemia*' from input of white blood cell (WBC/Leukocyte). Authors used RST based clustering techniques for the color based segmentation of Leukocyte.	Authors have not found any significance on computational time. The system also not support for "segmentation of blood smear images for overlapping leukocytes".
"Hepatitis disease diagnosis using Rough Set" (2012)	Tomasz et al.[39]	This study takes input of biometric data and predicts hepatitis disease in reducing input-data, with RST.	It increases the prediction accuracy without losing information/ knowledge.
"Rough Set based MRI Medical image segmentation using optimized initial centroids" (2013)	Anupama et al. [2]	This study is meant for analysis of MRI images of Brain (Web Database) using RST based clustering. Authors used '*RST based k-means(RKM)*' and '*fuzzy c-means (FCM)*' algorithms. They optimized the	Authors discuss that, the result is fully dependent on the initial parameters (membership value- FCM and Centroid RKM). It is also observed that, "random initialization of centroid also leads to an
"Design of a Diabetic Diagnosis System Using Rough Sets" (2013)	Margret et al. [28]	Authors used RST to diagnose diabetic conditions. The system compares with the knowledge base to the user given input.	The limitation of the knowledge base meant for only one disease.
"A hybrid decision support system based on rough set and extreme learning machine for diagnosis of hepatitis disease" (2013)	Kaya et al. [22]	To make diagnosis of hepatitis disease, authors developed a hybrid-decision-support system. Their system use RST, ELM (Extreme Learning Machine).	With the usage of RS-ELM, the classification accuracy found to be 100 %age.
"Application of Artificial Neural Networks & RST for the Analysis of Various Medical Problems & Nephritis Disease Diagnosis"	Devashri et al. [8] (2014)	Authors used ANN and RST for diagnosis of nephritis-disease patients.	Authors observed the significance factors of each chemical-test in applying RST and determined the %age accuracy of training during the diagnosis.
"Image Segmentation Using Rough Set Theory: A Review" (2014)	Payel et al.[35]	This paper contributes an analysis and comparison of image-segmentation using RST .	Discussed different methods for RST image- segmentations and highlighted some future research perspectives.
"An Overview of Rough-Hybrid Approaches in Image Processing" (2014)	Aboul [1]	Authors find to integrate RST with '*fuzzy sets*', '*mathematical morphology*', '*neural-networks*', '*genetic algorithms'*, SVMs and '*swarm intelligence*' algorithms in image processing.	Determined the hybrid approaches with RST in the applications of image processing.
"Rough Sets in Medical Informatics Applications"	Hassanien , et al. [19]	Authors used RST in medical image-segmentation, pattern classification, Medical Data-Mining and computer assisted medical decision-making systems.	Discusses the approach of hybrid techniques with RST would improve the performance of image processing tasks.
"Knowledge Mining from Clinical Datasets Using Rough Sets and Backpropagation Neural Network" (2015)	Kindie et al.[23]	Authors combine RST with back propagation neural network to classify clinical dataset. This system used clinical datasets like hepatitis, breast cancer, and heart disease.	RST application find accuracy of 97.3 %age in hepatitis dataset, 98.6 %age in breast cancer dataset and 90.4%age in heart-disease dataset.
"Rough set approach for an efficient	Manimara n et al.	Authors used five attributes from the Wisconsin breast	Authors discuss that, the computation time and resource

medical diagnosis system" (2015)	[27]	cancer dataset. RST is used for relative reduct function for classifications to remove the useless- attributes from the dataset.	usage was exponentially reduced with relative reduce algorithm.
"Disease diagnosis using rough set based feature selection and k-nearest neighbor classifier" (2015)	Femina et al. [13]	This system uses RST based '*feature- selection*' and '*KNN classifier*' for Hepatitis disease diagnosis.	Their method achieves 84.52 % of accuracy.
"Diagnosis of cancer using fuzzy rough set theory" (2016)	Meenachi et al. [29]	Authors use classification method to classify the cancer data with fuzzy RST and Particle Swarm Optimization technique to reduce features.	Efficiency of the classified data is measured through kappa statistics, sensitivity, AUC and F-measure.
"Rough set based rule generation technique in medical diagnosis: with reference to identification of heart disease" (2017)	Jain et al. [21]	Authors use rule generation techniques, LEM2 algorithm and used on heart disease dataset to demonstrate how rules can be generated without any calculation of reduct.	Quick Reduct Algorithm is found to be less time consuming compared to other used algorithms.

6. CONCLUSION

Rough set theory is an effective mathematical tool and applied by researchers successfully in many real-life intelligent decision making systems. This theory performs better in such cases where knowledge base is inconsistent, imprecise or uncertain. We discussed some intuitive issues and placed an exhaustive review in collecting papers across heterogeneous domain of applications. To make the discussion informative, utmost care is taken in highlighting fundamental concepts of this theory, successive variances and some generalisations. Although each extension methods of rough set has its own merit, however concepts of neighbourhood RST and covering based RST are still evolving with new generalisations. *Core*, *reducts*, *feature selection*, *dimensionality reduction* and *approximations* are some of the powerful concepts of RST that many researchers prefer to use for their own applications. We highlighted some existing RST based software systems/tools that would be useful in optimizing time and effort. In general, size of clinical or medical data set seems to be high dimensional in nature. Sometimes clinical knowledge base is also inconsistent, imprecise or uncertain. Hence rough set based feature selection and dimensionality reduction methods are obvious to be get used in healthcare data analytics. We presented a detailed review on this regard. Recently rough set theory is also used in social network analysis and hence future research on rough set based epidemic control, disease control systems, e-health gateway would bring some fruitful results.

REFERENCES

[1] Aboul-Ella Hassanien, Rough set approach for attribute reduction and rule generation: A case of patients with suspected breast cancer, J. The American Society for Info.Sc.&Tech., 55(11), (2004): 954-962.
[2] Anupama, N., Kumar, S. S., & Reddy, E. S., Rough set based MRI medical image segmentation using optimized initial centroids, Int. J. Emerging Technologies in Computational and Applied Sciences, 6(1), (2013): 90-98.
[3] Blanford, J. I., Blanford S., Crane R. G., Mann M. E., Paaijman K. P. Schreiber K. V. and Thomas M. B., Implications of temperature variation for malaria parasite development across Africa, Scientific Reports 3, (2013), Article number: 1300 doi:10.1038/srep01300.
[4] Bouma, M. J., Methodological problems and amendments to demonstrate effects of temperature on the epidemiology of malaria. A new perspective on the highland epidemics in Madagascar, Trans R Soc Trop Med Hyg . 97(2):133-9, (2003): 1972-89.

[5] Chen H.-L., Yang B, Liu J., and Liu D.-Y., A support vector machine classifier with rough set-based feature selection for breast cancer diagnosis, Expert Systems with Applications, 38(7), (2011): 9014-9022.
[6] Chen You-Shyang, Cheng Ching-Hsue, Application of rough set classifiers for determining hemodialysis adequacy in ESRD patients, Springer-Verlag, Knowl Inf Syst., 34, (2013): 453-482.
[7] Craig, M.H., Snow R.W. and le Sueur D., A climate-based distribution model of malaria transmission in sub-saharan Africa, Parasitol Today, (1999): 15:105-111.
[8] Devashri Raich and P.S. Kulkarni, Application of artificial neural networks and rough set theory for the analysis of various medical problems and nephritis disease diagnosis, Advances in Intelligent Systems & Computing, Springer Int.Pub. Switzerland (2013): 83-90.
[9] Dhingra N, Jha P, Sharma VP, Cohen AA, Jotkar RM, Rodriguez PS, Adult and child malaria mortality in India: A nationally representative mortality survey, 376(9754), (2010): 1768–74.
[10] Duane J. Gubler, "Resurgent vector-borne Diseases as a global health problem, Emerging Infectious Diseases, 4(3),(1998): 442-450.
[11] Dubois, D., Prade, H., Putting rough sets and fuzzy sets together, Intelligent decision support, Handbook of applicatons and advances of the rough set theory, Kluwer Acad Publ, Dordrecht, (1992): 203-232.
[12] Femina B, Anto S, Disease diagnosis using rough set based feature selection and K-nearest neighbor classifier, Int. J. Multidisciplinary Research and Development, 2(4), (2015): 664-668.
[13] Farion,K, Michalowski, W. Slowinski, R., Wilk, S., Rubin, S. Rough set methodology in clinical practice: controlled hospital trial of the MET System, Int. Conf. Rough Sets & Current Trends in computing, Lecture Notes in AI, 3066, (2004): 805-814.
[14] Ginter, F., Pahikkala, T., Pyysalo, S., Boberg, J., Jarvinen, J., Salakoski, T., Extracting protein-protein interaction sentences by applying rough set data analysis, Int. Conf. Rough Sets and Current Trends in Computing, Lecture Notes in AI, 3066, (2004): 780-785.
[15] Greco, Salvatore, Matarazzo, Benedetto, Słowiński, Roman, Rough sets theory for multicriteria decision analysis, European Journal of Operational Research, 129 (1), (2001): 1-47.
[16] Grzegorz Iiczuk, Alicja Wakulicz-Dejz, Rough sets approach to medical diagnosis system, Int. Atlantic Web Intelligence conf., (2005): 204-210.
[17] Grzymala-Busse, J., Knowledge acquisition under uncertainty–a rough set approach, J. Intelligent and Robotics Systems, 1, (1988): 3-16.
[18] Halder, A., Dasgupta, A., Color image segmentation using rough set based k-means algorithm, Int. J. Computers and Applications, 57(12): 32-37.
[19] Hassanien A.E., Abraham A, Peters J.F., and Schaefer G., Overview of rough-hybrid approaches in image processing, IEEE Conference on Fuzzy Systems, (2008): 2135-2142.
[20] Herbert, J. P.; Yao, J. T., Game-theoretic rough sets. Fundamenta Informaticae, 108(3-4), (2011):267-286.
[21] Jain P, Agrawal K , Vaishnav D, Rough set based rule generation techniques in medical diagnosis: with reference to identification of heart disease, Int. J. Scientific Research in Mathematical and Statistical Sciences, 4 (3), (2017): 12-18.
[22] Kaya Y and Uyar U, A hybrid decision support system based on rough set and extreme learning machine for diagnosis of hepatitis disease, Applied Soft Computing Journal, 13(8), (2013): 3429-3438.
[23] Kindie Biredagn Nahato, Khanna Nehemiah Harichandran, and Kannan Arputharaj, Knowledge mining from clinical datasets using rough sets & backpropagation neural network, Computational and Mathematical Methods in Medicine, Article ID 460189, (2015): (13), http://dx.doi.org/10.1155/2015/460189.
[24] Kumar A, Valecha N, Jain T, Dash AP, Burden of malaria in India: retrospective and prospective view, American Journal of Tropical Medicine and Hygiene, (2007): 69-78.
[25] Lin T. Y., Neighborhood systems and relational database, Proc. of ACM 16th annual computer science conference, (1988): 725-732.
[26] Lin,T.Y., Granulation & nearest neighborhoods, Roughset approach, granular computing: An emerging paradigm, 70(2001): 125-142.
[27] Manimara A, Chandrasekaran V. M. , Asesh Aishwarya, Rough set approach for an efficient medical diagnosis system, Int. J. Pharmacy and Technology, 7(1), (2015): 8049-8060.
[28] Margret A.S., Clara Madonna L. J., Jeevitha P., Nandhini R.T., Design of a diabetic diagnosis system using rough sets, Cybernetics and Information Technologies, 13(3), (2013): 124-139
[29] Meenachi1 L, Ramakrishnan S, Arunithi M., Karthiga R. Karthika S, Nandhini P, Diagnosis of cancer using fuzzy rough set theory, Int. Research J. of Engg. & Tech.(IRJET), 03(01), (2016): 1203-1208
[30] Midelfart H., Komorowski, H.J., Norsett, K. Yadetie, F., Sandvik, A.K., Laegreid, A, Learning rough set classifiers from gene expressions and clinical data, Fundamenta Informaticae, 53(2), (2002): 155-183.
[31] Mohapatra, S., Patra, D., & Kumar, K., Unsupervised leukocyte image segmentation using rough fuzzy clustering, ISRN Artificial Intelligence., (2012): 1-14. doi:10.5402/2012/923946.

[32] NaliniPriya, G., Kannan A and Ananahakumar P, Dynamic context adaptation for diagnosing the heart disease in healthcare environment using optimized rough set approach. Int. J. on Soft Computing (IJSC), 3(2), (2012): 23-33.
[33] Ningler, M., Stockmanns,G., Schneider, G., Dressler, O., Kochs, E. F., Rough set-based classification of EEG-signals to detect intraope- rative awareness: Comparison of fuzzy and crisp discretization of real value attributes, Proc. of Int. Conf. on Rough Sets and Current Trends in Computing, Lecture Notes in A.I. 3066, (2004): 825-834.
[34] Pawlak Z., Rough Sets, Int. Jour. Inf. Comp.Sc., II, (1982): 341-356.
[35] Payel Roy, Srijan Goswami, Sayan Chakraborty, Ahmad Taher Azar, Nilanjan Dey, Image segmentation using rough set theory: A review, Int. J. of Rough Sets and Data Analysis, 1(2), (2014): 62-74.
[36] Piotr Paszek, Alicja Wakulicz-Deja, Applying rough set theory to medical diagnosing, Intl. Conf. on Rough Sets and Intelligent Systems Paradigms, (2007): 427-435.
[37] Senthilkumaran N. , Rajesh R., A study on rough set theory for medical image segmentation, Int. J. of Recent Trends in Engineering, 2(2), (2009): 236-238.
[38] Skowron, A., Stepaniuk J., Tolerance approximation spaces, Fundamenta Informaticae. 27 (2-3), (1996): 245-253.
[39] Tomasz KANIK, Ing., Hepatitis disease diagnosis using rough set, ICTIC 2012, (2012): 19- 23.
[40] Tripathy B.K., Acharjya D. P. and Cynthya1 V., A framework for intelligent medical diagnosis using rough set with formal concept analysis, Int. J. of A. I. & Applications (IJAIA), 2(2), (2011): 45-66
[41] Yao, Y.Y., Wong, S.K.M., and Lingras, P., A decision-theoretic rough set model, Methodologies for Intelligent Systems, Proc. 5th Int. Symposium on Methodologies for Intelligent Systems, Knoxville, Tennessee, USA, (1990): 25-27.
[42] Xian-Ming Huang, Yan-Hong-Zhang, A new application of rough set to ECG recognition, Proc. 2nd Int. Conf. on Machine Learning and Cybernetics, (2003): 1729-1734.
[43] Yi Xie, On medical image filtering based on rough set theory, 5th Int. Conf. on fuzzy systems and knowledge discovery, IEEE, (2008): 276-280.
[44] Zadeh L. A., Fuzzy Sets, Information & Control, 8, (1965): 338-353
[45] Ziarka, W., Variable precission rough set model, J. Computer & System Sciences, 46(1), (1993): 39-59.
[46] Zhu, F, Wang. F. Y., Some results on covering generalized rough sets, Pattern Recognition and Artificial Intelligence, 15(1), (2002): 6-13.

