International Research Journal of Multidisciplinary Scope (IRJMS), 2024; 5(2): 732-744

Original Article | ISSN (O): 2582-631X DOI: 10.47857/irjms.2024.v05i02.0631

Coupling of Rough Set Theory and Predictive Power of SVM Towards Mining of Missing Data

Surendra Nath Bhagat[1*], Premansu Sekhar Rath[1], Anirban Mitra[2]

[1]GIET University, Gunupur, Odisha, India. [2]Amity University, Kolkata, West Bengal, India. *Corresponding Author's Email: surendra.bhagat@giet.edu

Abstract

Rough set theory offers a novel approach to identifying structural correlations amidst imprecise or noisy data, particularly applicable to variables with diverse values. It presents a promising avenue for handling fuzzy, conflicting, and uncertain data, with recent models incorporating various fuzzy generalizations. This technique stands out as a popular solution within artificial intelligence, particularly in data analysis and processing tasks. In the medical domain, where missing data poses a significant challenge, leveraging rough set theory alongside machine learning algorithms for disease prediction is common. This paper proposes a model that effectively predicts missing values using rough set theory, addressing the prevalent issue of incomplete data. By providing a systematic approach and robust algorithm, the model demonstrates the adaptability and potential of rough set theory in contemporary data analysis scenarios. Classification of the predicted data set using supervised Learning Model (SVM) results in accuracy of 82.1% while the F1 score is 82.6%. Through validation with real-life medical datasets using supervised classification techniques, the paper underscores the accuracy and applicability of the proposed algorithm, offering a valuable tool for researchers and practitioners grappling with the complexities of modern data analysis.

Keywords: Knowledge Extraction, Missing Value Prediction, Rough Set, Support Vector Machine.

Introduction

In a world of advancing technology and rapidly increasing data, the search for valuable perceptions from uncertain or noisy information has become extremely important. This search lines up well with Rough Set Theory, a unique mathematical technique that excels at dealing with uncertain and inconsistent data. Professor Pawlak first proposed the rough set theory in 1982 (1) as an important numerical technique that offers methods for using mathematics to find unidentified patterns in information. It may be used for information reduction, decision rule creation, pattern extraction (templates, association rules), feature selection, feature extraction, and pattern extraction along with recognizing full or partial data dependencies, removing redundant data, providing a method for handling null values, missing data, dynamic data, and other issues (2, 3). It may be assumed as a powerful tool for realizing hidden patterns in data, especially when the data's attributes are distinct values. Its superior ability to handle unclear, inconsistent, and ambiguous information makes it really useful for handling real-world data challenges Rough Set Theory, garnering attention in artificial intelligence, amalgamates concepts from probability, fuzzy set, and evidence theories to analyze uncertain data. Researchers propose various methodologies adaptable for different applications, aiming to understand its fundamental concepts and unique features. The objective of this research is to inspect the essential concepts and working principle of rough set theory, it also tried to find out the exclusivity of this theory. This work proposes a novel model to predict missing values in dataset by using the concepts of rough set. The work also proposes an algorithm for this prediction. The design proposes an iteration process that makes intelligent prediction about what the missing values should be based on special relationships and decision rules.

(Received 18th February 2024; Accepted 23rd April 2024; Published 30th April 2024)

This technique helps in the dataset's completion. Also, the present work uses the real-life medical dataset. The dataset comprises 1032 patient data. The proposed algorithm is used for predicting missing values. This work also presents the detailed result analysis of dataset having missing value and the dataset without missing value.

The next part of the paper is organized to clarify major features of Rough Set Theory and how it is applied. Section 2 provides an overview of Rough Set Theory, discussing its key notions and how it can be useful when dealing with unclear data. Section 3 examines how Rough Set Theory is applied to different fields. Section 4 discusses our new Rough Set Theory-based methodology for anticipating missing values. Section 5 gives the experimental validation of the proposed algorithm using supervised machine learning techniques. Finally, we summarize our observations and conclusions in Section 6. Overall, this paper illustrates how powerful Rough Set Theory can be when dealing with tricky data situations. By coming up with a new model and a way to predict missing values, and by applying these ideas to graphs, we're adding to the growing knowledge about how to handle complex data problems in the world of data science and artificial intelligence. In the current work , using a real life data set with missing values, we have developed a robust and efficient system for predicting missing values such that a supervised machine learning model can effectively learn from the data towards realizing an automated solution.

Overview of Rough Set

A brand-new branch of mathematics is rough set theory which follows probability theory, fuzzy set theory, and proof theory and used as an instrument for dealing with uncertain information as a framework of data processing and analysis. The membership function in fuzzy set theory is crucial. But on other side, it is uncertain how the membership function will be chosen. As a result, in some ways, fuzzy set theory is in some respects an ambiguous mathematical instrument for handling ambiguous circumstances. To describe the imprecise concepts, rough set theory establishes two definite boundary lines. As a result Rough set theory is at that moment a mathematical tool for resolving ambiguous problems. The concept of rough set is definable with the help of topological operation, interiors and closure. Let U be a limited collection of things, and there is a given binary relation R ⊆ U × U. The set U is referred as Universe, whereas R is known as indiscernibility relation. The Indiscernibility Relation IND(B) is defined as a binary relation on U defined for a and y, where x,y ∈ U as follows:

$(x, y) \in IND(B) if\ and\ only\ if\ \rho(x, a) = \rho(y, a) for\ all\ a \in B$ [1]

To simplify it we take R as an sameness relation and (U,R) as an approximation space. The most important objective is to describe the set X (where X ⊆ U) with respect to R. R(x) stands for the R's equivalence class as determined by the element X. A collection of object which can be classified with full certainty as members of set X is defined as lower approximation to R with regard to a given set X can be represented by

$R_*(X) = \{ x: R(x) \subseteq X\}$ [2]

The collection of object that may possibly be classified as member of Set X is defined as Upper approximation to R with regard to a given set X can be represented by

$R^*(X) = \{x: R(x) \cap X \neq \emptyset\}$ [3]

The Boundary region with regard to R of the set X is comprised of object that cannot be classified with certainty to be neither inside X, nor outside X, is represented by

$BNR(X) = \{R^*(X) - R_*(X)\}$ [4]

So now we can conclude that If and only if X is considered rough with regard to R if a set X has a non-empty boundary region. It's exciting to compare how fuzzy sets, classical sets and rough sets are defined. A fundamental idea that may be used naturally to defined the classical set. In order to define fuzzy sets, one uses the complex numerical combination of numbers, functions, and structures known as the fuzzy membership function. Approximations define rough sets. As a result, this definition necessitates complex mathematical ideas.

Review of Literature

Rough set theory is crucial for decision-making in handling uncertain or imprecise information. It finds applications in various domains including image processing, pattern identification, data mining, social

media, and medical informatics. Often used alongside fuzzy logic, genetic algorithms, and neural networks, it extracts patterns from databases effectively.

Application of Rough Set in Health Care Science

Healthcare science's precision challenges require complex analysis. Rough set theory, tailored for the medical sector, adapts well. Studies show its importance: Ali R et al. (4) propose H2RM for diabetes prediction; Chen HL et al. (5) introduce RS SVM for breast cancer; Maji P et al. (6) use rough-fuzzy techniques for medical images; Wang W et al. (7) merge rough set theory with CART in R2-CART; Tsumoto et al. (8) develop PRIMEROSE-REX for knowledge extraction; Huang et al. (9) combine FCM clustering with set theory for segmentation; Rajesh T et al. (10) use rough set theory for attribute mining and brain tumor recognition; Senthil Kumar S et al. (11) categorize signals using Pan Tomkins and Wavelet transform; Yekkala I et al. (12) predict heart infection; Santra D et al. (13) design a rule-based expert system; Xie Y et al. (14) propose a rough set-based image filtering method; Hassanien AE et al. (15) categorize breast cancer with 98% accuracy; Rajeash T et al. (16) categorize MRI brain pictures using Rough Set Theory and Feed Forward Neural Network. Figure 1 illustrates its significance in breast cancer, heart disease, diabetes, and brain cancer until mid-2022.

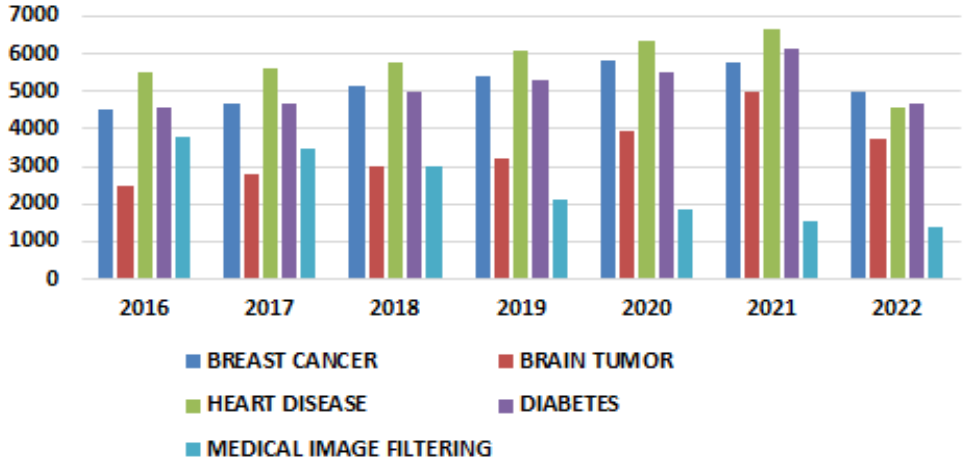

Figure 1: Implementation of Rough Set in Health Care

Rough set in Data Mining

Rough set theory, applied across diverse fields, addresses decision-making ambiguity by Slimani et al. (17), redefines approximation notions using topological concepts in Kusiak A et al. (18), forecasts semiconductor trends in Wang GY et al. (19), introduces the RIDAS system advanced pattern recognition systems in Chan CC et al. (20), and develops quasi-incremental models with LERS in Beaubouef et al. (21). It also enhances topological connectionsin Guan JW et al (22), challenges correlation methodologies in Griffin et al. (23), integrates rough sets into Tcl scripting language with RSLTcl in Lingras et al (24), performs resilient clustering in Chen H et al. (25), enables dynamic updates with UAGOAAS in Zhu W et al. (26), establishes relationships using exact-reduct in w Grzymala-Busse et al.(27), navigates data complexities in Reddy GT et al.(28), measures partitioning efficacy in Stepaniuk et al (29), and identifies essential medical datasets related to type1 diabetes in Dutta S et al. (30).

This summary collectively underscores the comprehensive effectiveness of rough set theory in diverse data mining domains, exemplified by the noticeable expansion in annual studies illustrated in Figure 2, up to mid-2022.

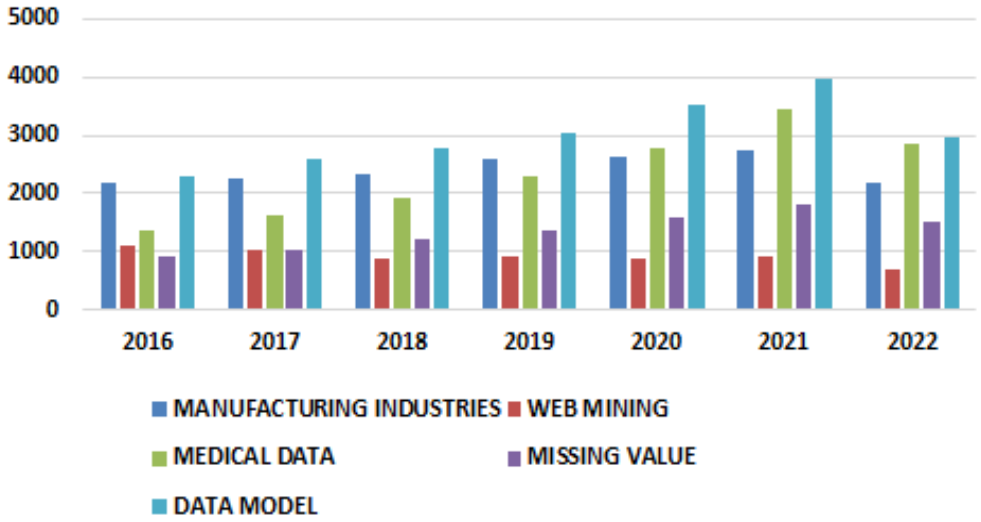

Figure 2: Implementation of Rough Set in Data Mining

Rough set in Social Network

Social media platforms like Facebook, Twitter, and LinkedIn provide valuable data on user interactions, facilitating audience segmentation and sentiment analysis for marketers. Research on sentiment analysis on Twitter, particularly with fuzzy and rough set classifiers, shows promise. Kundu S et al. (31) propose a spam categorization model using rough set theory features. (32) addresses social network categorization and clustering challenges with rough set theory. Fan et al. (33) introduce a method for community relationships, and (34) presents a fuzzy-rough community detection algorithm. Khachfeh RA et al. (35) use rough set-based indiscernibility relations to study social networks. Yang L et al. (36) analyze relational structures in social networking data. Islam MA et al. (37) explore link prediction using rough set theory on Facebook. Jader RF et al. (38) propose a rough set-based approach for Digital Rights management. Gaeta A et al. (39) used Rough Set Theory to analyze phenomena related to Information Disorder. Figure 3 depicts the increasing use of rough set theory in social network research.

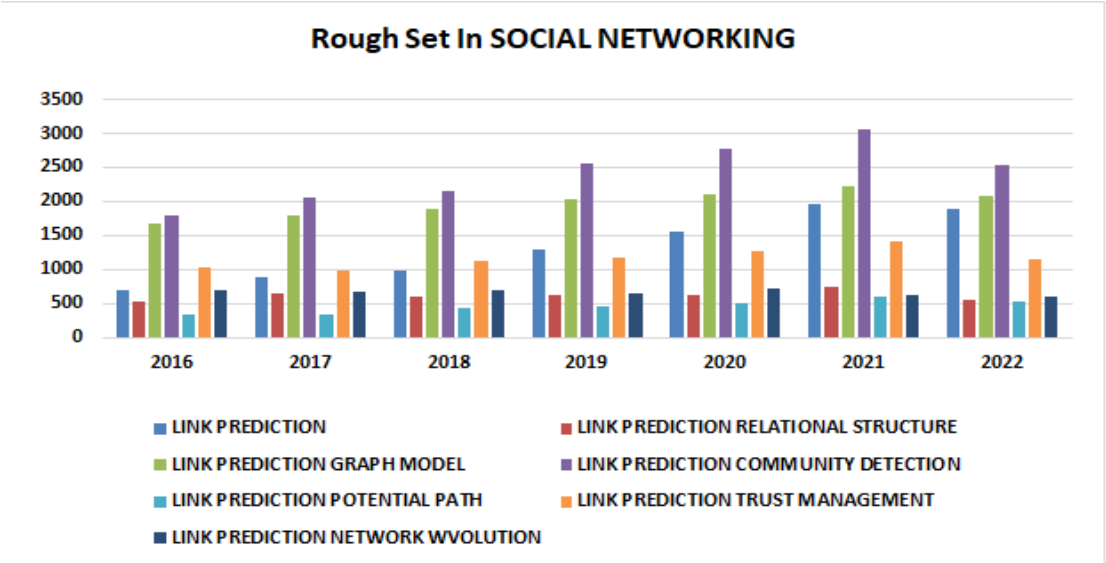

Figure 3: Implementation of Rough Set in Social Network

Motivation

Based on the review of relevant literature it is observed that no such research have taken this approach of using SVM with data generated by RST. One fundamental aspect of combining RST with SVM is the treatment of missing data. RST provides a framework for handling uncertainty and incomplete information, which can be seamlessly integrated into the SVM algorithm.

Problem Statement

The current work aims to address the following problem:

Given any real-life data set with missing values, can we develop a robust and efficient system for predicting missing values such that a supervised machine learning model can effectively learn from the data towards realizing an automated solution.

Methodology

The proposed model as shown in Figure 4 begins by getting raw data. This raw data, known as primary data, is the foundation for later analysis. First, it is checked if the data is having missing values, paying close attention to missing values in specific parts of the data table. If there are missing values strong rough set methods are used to predict the missing values. This process finishes by filling the missing value and completing the table.

Once successfully missing values are predicted, the processed table becomes a structured rough set representation. This sets the stage for creating a complete dataset on which decision-making tools can be applied to extract knowledge. In complex datasets, getting knowledge and representing it is important to find hidden connections and new ideas. This knowledge extraction method reveals important connections between things, even when they're delicately linked. Ultimately, the model helps to extract knowledge by comparing the previous and the present predicted missing values. This way of extracting knowledge reveals important connections between different things. This helps us understand what data is relevant. Then, we carefully get it ready and mix it together, bringing together different types of data in a complete way. These results are very important, as they are ready to help make accurate predictions and informed decisions in various areas of operation.

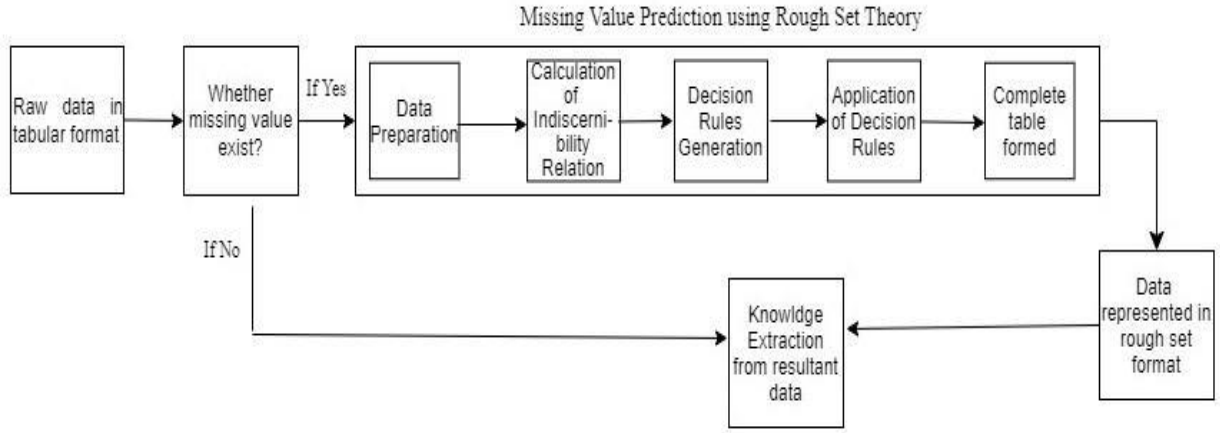

Figure 4. Proposed model

Proposed Algorithm

Algorithm for prediction of missing values using rough set theory:

Algorithm: To predict all missing values in a table using rough set theory, you can follow a step-by-step procedure. Here's an algorithmic approach to predict all missing values in a table using rough set theory:

Input: Table with missing values

Output: Table with predicted missing values

1. Data Preparation

Identify the decision attribute(s) for which missing values need to be predicted.

Split the table into two sets: a set with complete records (training set) and a set with missing values (test set).

2. Indiscernibility Relation

Apply the indiscernibility relation to the training set. Group records with similar attribute values together. Create equivalence classes based on the indiscernibility relation.

3. Decision Rules Generation

For each equivalence class, determine the most frequent values of the decision attribute(s) in the training set. Create decision rules that map the attribute values of the equivalence class to the predicted decision attribute values.

4. Apply Decision Rules

Iterate through each record in the test set with missing values.

Identify the equivalence class (es) to which the record belongs based on its attribute values.

Use the corresponding decision rule(s) to predict the missing value(s) for the decision attribute(s).

5. Predict Missing Values Iteratively

Repeat steps 3 and 4 for each missing value in the test set.

After each prediction, update the attribute value in the test set with the predicted value.

Continue the iteration until all missing values are predicted.

6.Output

Return the table with all missing values replaced by the predicted values.

Explanation

The algorithm for predicting missing values utilizing rough set theory, is a methodical approach that starts with a table having missing values. We have to identify the decision attribute (D) for which missing values need to be predicted i.e:

Find D such that $\exists i,j$ where $M_{i,j} = NA$ [5]

Where, D: represents the decision attribute.

$M_{i\,j}$ represents the value of the i-th object for the j-th attribute.

$\exists i,j$ state that there is an existence of at least one object (i) and one attribute (j) where the value is missing ("NA").

In the next step, decision attributes are recognized, and the table is separated into two sets: a training set containing complete records (Ucomplete) and a test set containing the records having missing values (Umissing). Logically these two sets can be expressed as:

$U_{complete} = \{u \in U \mid \forall j\ (M_{uj} \neq NA)\}$ [6]

$U_{missing} = \{u \in U \mid \exists j\ (M_{uj} = NA)\}$ [7]

Where, U represents the universe, i.e the set of objects.

$\forall j$ states that, for all attributes Aj, the value is not missing for the object u in U complete.

$\exists j$ states that there exists at least one attribute Aj for which the value is missing for the object u in U missing.

Next, the training set is subjected to the indiscernibility relation ($\sim$) which is defined as:

$[a] \sim [b] \Leftrightarrow [a]\sim = [b]\sim$ [8]

Where $[a] \sim [b]$ means that a and b cannot be distinguished based on the attributes in X. It means that both a and b share the same type of value with respect to these attributes.

The indiscernibility relation groups records with similar attribute values into equivalence classes. These classes represent distinct patterns in the data. The equivalence class can be represented as:

$$[a]Ix = \{b \in U \mid a\ Ix\ b\}$$ [9]

Where: [a]Ix is an equivalence class which consist of all elements which are indiscernible

w.r.t the attribute X.

U is universe of object with asset of attribute.

Ix represents the equivalence classes.

a and b is an element where $a,b \in U$.

a Ix b representing that element a and b are indiscernible w.r.t the attribute X.

Once equivalence classes are formed, decision rules are produced. For every class, the algorithm recognizes the most frequent values of the decision attribute(s) within that class. These recurrent values help us for prediction of missing values. The decision rules generation can be expressed as:

IF {X1,X2,…,Xm} THEN D =Y [10]

Where, X1,X2,…,Xm are conditions based on attributes A1,A2,…,An.

D is the decision attribute.

Y is the decision class

The conditions X1,X2,…,Xm are determined by the discernibility relation ($\sim$). The conditions specify the characteristics that object in the same decision class share, considering the indiscernibility relation. The

decision attribute D is then determined based on these conditions.

Moving to the application stage, the algorithm recurrently processes every record in the test set having missing values. It allocates the record to its corresponding equivalence class on the basis of its attribute values. The decision rules generated earlier are then employed to predict the missing values for the decision attribute(s) of the record.

This prediction process is repeated for all missing values present in the test set. After every prediction, the predicted value is replaced in the test set. This iterative process continues until all missing values in the test set are predicted using the decision rules. The output of the algorithm is a transformed table where all missing values have been replaced with the predicted values, effectively leveraging the principles of rough set theory to enhance the completeness of the data. This technique provides a systematic method for using data relationships and patterns to produce intelligent guesses for missing values.

Example

Suppose we have a dataset of potential car buyers. The table has attributes like "Age," "Income," "Gender," "Education," and "Purchased" (our decision attribute). The goal is to predict missing values in the "Purchased" attribute.

Consider that we have a dataset of possible car buyers (Table 1). This table has 5 attributes such as "Age," "Income," "Gender," "Education," and "Purchased", out of which "Age," "Income," "Gender," and "Education," are conditional attributes and "Purchased" is a decisional attribute. Our goal is to predict the missing value in the decisional attribute "Purchased" by using the above algorithm. The table is as given below.

Table 1. Dataset with missing value

Conditional Attributes				
Age	**Income**	**Gender**	**Education**	**Decisional Attribute Purchased**
25	50000	M	Grad	Yes
30	60000	F	Undergrad	No
NULL	70000	M	Grad	NULL
40	80000	M	Grad	No
NULL	90000	F	Undergrad	NULL
22	55000	F	Grad	Yes
35	65000	M	Undergrad	Yes
NULL	75000	F	Grad	NULL
28	85000	M	Grad	Yes
45	95000	M	Undergrad	NULL

Now, we try to explain the above algorithm with the help of above table:

Step 1: Data Preparation

- First we select the decision attribute that is the attribute whose missing value is to be predicted, in the above example the decision attribute is: "Purchased."
- Then we divide the above dataset in to two segments named as training set (which consists of complete records) and test sets (which consists of record which have missing value for attribute “Purchased”)

Step 2: Indiscernibility Relation

- By applying indiscernibility relation on the training set of the dataset we group the records with similar value of "Age," "Income," "Gender," and "Education" together and different equivalence classes are created.

Step 3: Decision Rules Generation

- The most frequent "Purchased" value is determined from each equivalence classes of the training set
- For example, in the above dataset if we have an equivalence class which consists of the following

attributes: "Age" in between 25 and 35, "Income" is less than 70000, "Gender" as "Male," and "Education" as "Grad," the most frequent "Purchased" value is "Yes."

Step 4: Apply Decision Rules

- Now we apply the decision rule through iterate to the test set (which consists of record which have missing value for attribute "Purchased"). For example, consider a record from the above data sets with "Age" as NULL, "Income" as 70000, "Gender" as "Male," and "Education" as "Grad." This record belongs to the equivalence class mentioned earlier. As because the most frequent value of decision attribute "Purchased" in this class is "Yes," we predict "Yes" for this record.

Step 5: Predict Missing Values Iteratively

- The above process (step 4) is repeated for all records in the test set who have missing value for the decision attribute "Purchased" and for every records that belong to an equivalence class with a clear majority decision, predict that value.
- For example, consider if another record from the above dataset has "Age" as null, "Income" as 75000, "Gender" as "Female," and "Education" as "Grad," it belongs to an equivalence class with no clear majority. In this case, we cannot confidently predict the "Purchased" value.

Step 6: Output

- After running the algorithm, the table will be populated with the missing "Purchased" values from the test set based on the projected decisions from the decision rules.
- Table 2 shows the table might look after the algorithm:

Table 2: Dataset with predicted missing value

Conditional Attributes				
Age	**Income**	**Gender**	**Education**	**Decisional Attribute Purchased**
5	50000	M	Grad	Yes
30	60000	F	Undergrad	No
NULL	70000	M	Grad	Yes
40	80000	M	Grad	No
NULL	90000	F	Undergrad	Yes
22	55000	F	Grad	Yes
35	65000	M	Undergrad	Yes
NULL	75000	F	Grad	Yes
28	85000	M	Grad	Yes
45	95000	M	Undergrad	No

Experimental Validation

In this section, we utilized the proposed algorithm to fill in missing values in a medical history dataset. A standard dataset was obtained, and, for experimentation, random instances were removed using a Montecarlo simulation approach. Three versions of the dataset were generated. Classification performance was evaluated for the dataset with missing values and for the dataset where missing values were replaced by our proposed algorithm. A comparative analysis was conducted based on the obtained results.

Data Description

The current work uses data from the work by Islam et. al (37) the dataset was gathered from a survey collected from Enam Medical College, Savar, Dhaka, Bangladesh. Different medical data like blood pressure, hypertension, RBC, WBC , age etc. were a included in the dataset. The dataset consists of 29 such attributes of 1032 patients.

Supervised Learning

In supervised learning Support Vector Machines (SVM) excel, particularly in classifying labeled data. SVM aims to establish an optimal hyperplane for effective separation between various classes within a dataset. Operating on labeled training data, each

data point is tagged with a specific class label, and SVM identifies crucial support vectors for defining the decision boundary. Its effectiveness lies in navigating high-dimensional feature spaces and handling non-linear relationships through kernel functions (38). In this study, SVM Classification was applied to datasets containing missing values, both in their original state and after modification by our algorithm. A comprehensive performance evaluation and comparative analysis were conducted to assess the model's efficacy in handling missing values. The findings contribute insights into the adaptability and performance of our proposed missing value prediction model within the context of missing data scenarios.

Evaluation Metrics

In this experiment, a supervised classifier was used where the training and testing data were distributed in a 63:37 ratio. The F1 scores have been used to show the classifier's performance. The F1 score can be defined as:

F1 score = 2 * Precision * Recall / (Precision + Recall).

Precision and recall can also be represented as:

Precision = True Positive / (True Positive + False Positive) and,

Recall = True Positive / (True Positive + False Negative).

In this context, "True Positive" denotes the classifier accurately identifying the positive class as positive in its prediction. Similarly, "True Negative" refers to the classifier correctly classifying a negative class as not positive during its prediction. Conversely, if the classifier wrongly predicts a negative class as positive, it is termed "False Positive." On the other hand, when the classifier incorrectly labels a positive class as not positive in its prediction, it is referred to as a "False Negative."

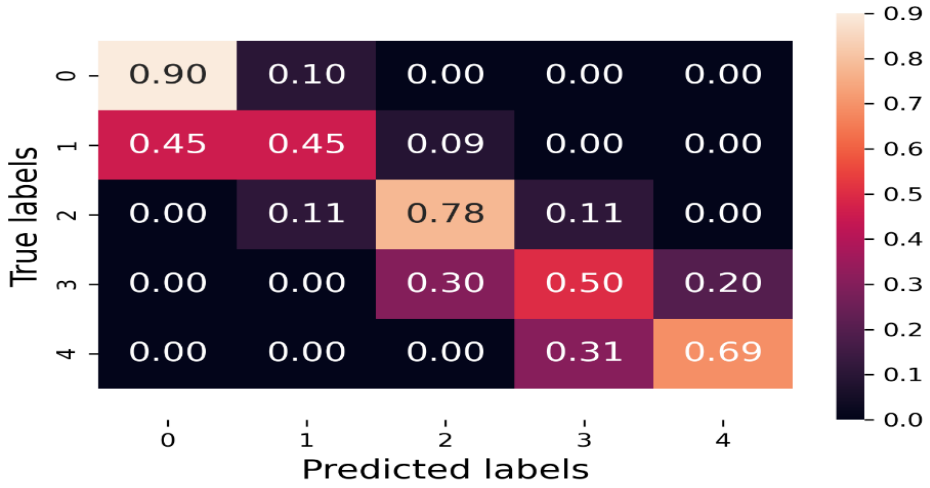

Figure 5: Confusion Matrix for 1st version

Results and Discussion

In the first version of dataset with randomly simulated missing values, It is observed that the SVM classifier identifies stage 0 of chronic kidney disease symptoms most accurately. This is crucial for an automated healthcare system as early detection and diagnosis may result in effective and time treatment. However, remaining stages of chronic kidney disease are not that effectively detected. There is a good amount of misclassification in other cases as shown in Figure 5. The overall accuracy of the system is 69.8% with F1 score of 69.1%.

In this second version of dataset with randomly missing values, the classifier performs with better accuracy and F1 scores of 73% and 72.4% respectively. In this case also it is observed that the early stage detection accuracy is high, especially for stages 0 and 1. However, in other cases as shown in Figure 6, the misclassification rate is also high, as observed for stage 4 where 70% of the data is wrongly detected to diagnose that that patient is in stage 3. This may lead to faulty treatment undermining the gravity of the patients health condition.

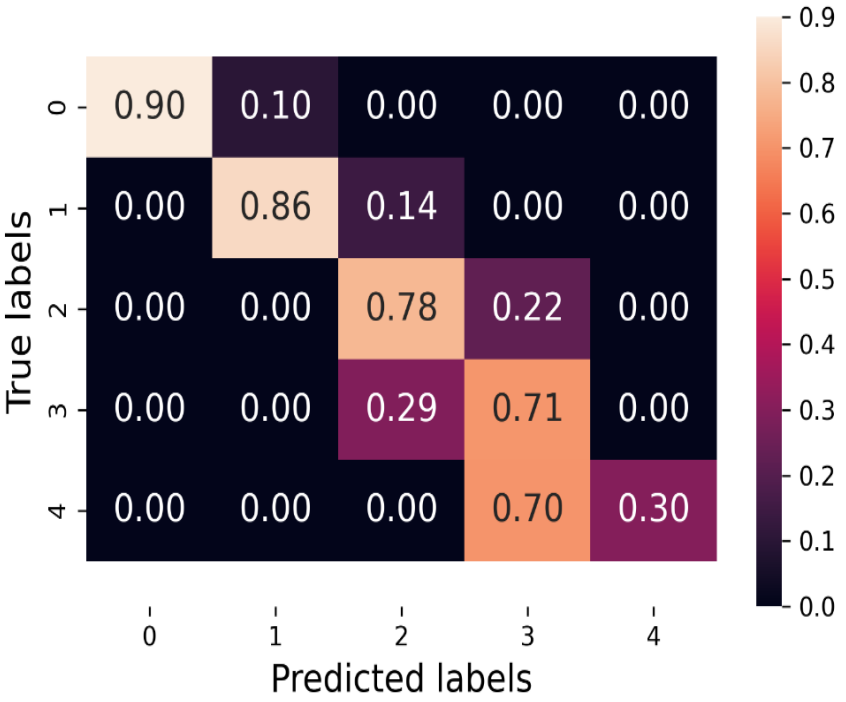

Figure 6: Confusion Matrix for 2nd Instance

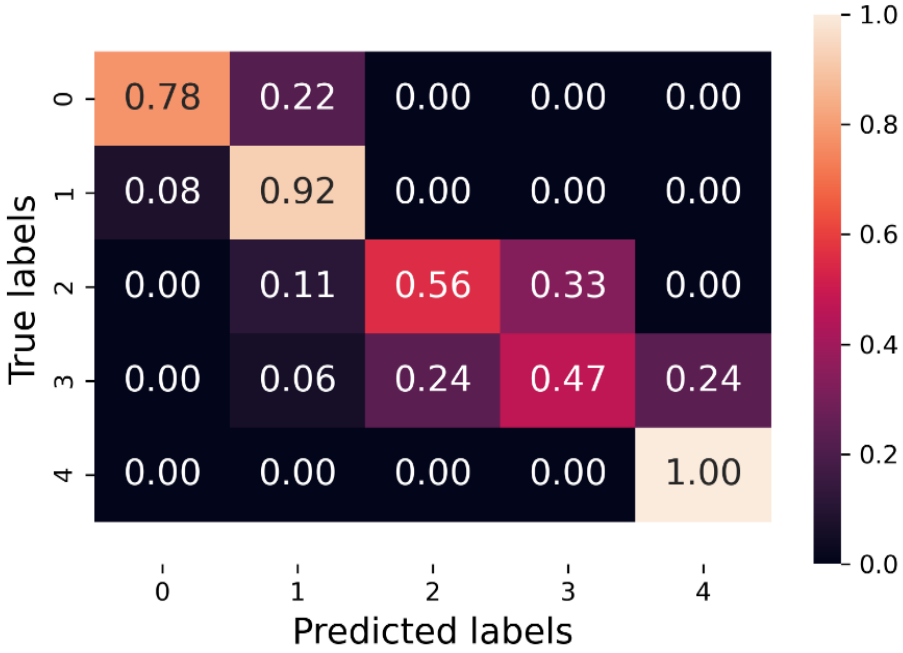

Figure 7: Confusion Matrix for 3rd version

In comparison to the prior datasets, for this third version of dataset it is observed that stage 0 has lower detection accuracy while stage 1 of chronic disease detection is higher. Also, the last stage of chronic disease is detected with 100% accuracy. However, for this dataset there is substantial mis-classification for the stages 2 and 3 of the disease shown in Figure 7. The overall accuracy obtained by the system is 71.4% whereas the F1 score is 70.7%.
On applying our proposed algorithm to the datasets with missing values, we use the predicted values to generate a complete dataset. On this dataset, it is observed that all the stages of chronic kidney disease are detected with a good amount of efficiency with minimum mis-classification as shown in figure 8. The highest misclassification is noted for the last stage of the disease, i.e., stage 4 where 27% of stage 4 data is classified as stage 3. The accuracy achieved by the system is 82.1% while the F1 score is 82.6% which is much higher than that for the datasets with missing values.

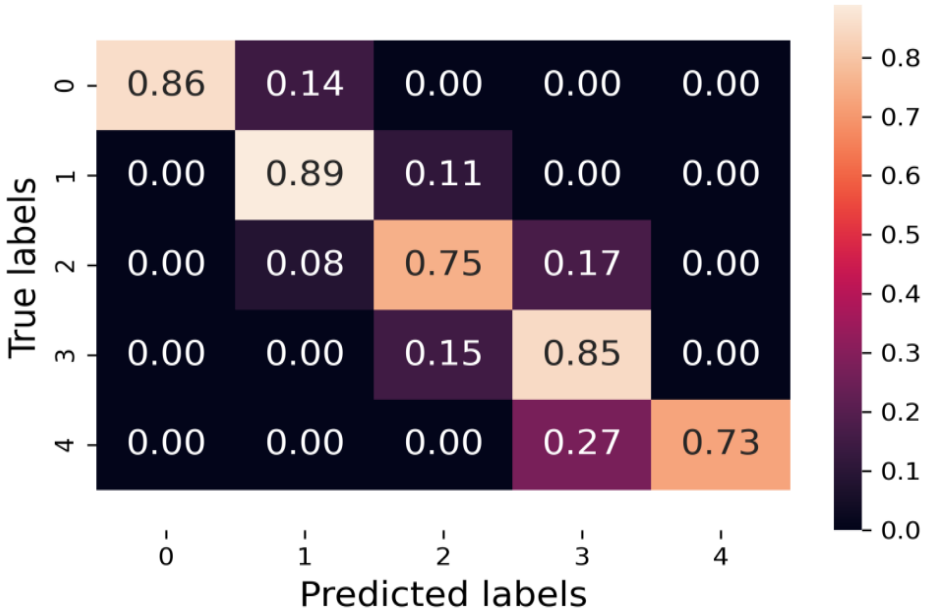

Figure 8: Confusion Matrix for dataset with predicted values

Thus, it can be concluded that the proposed algorithm results in the prediction of missing values such that the dataset is effectively usable for the development of an automated chronic kidney disease identification system.

Conclusion

The study demonstrates the effectiveness of rough set theory in handling complex medical data in health care science. It shows remarkable performance for pattern detection in large datasets in data mining and has a deep impact on social networks with complex relations. The research ascertains the application of rough set theory in predicting missing values, addressing challenges in health care science, data mining, and social networks with unspecified, imprecise, and partial data. The authors present a methodical procedure, including an algorithm, to predict missing values, offering a useful framework for real-world applications. The study highlights the groundbreaking potential of rough set theory in data analysis and prediction, showcasing its flexibility and applicability across various domains of real-world applications in healthcare science, data mining, and social networks. The study employs a dataset from Enam Medical College, Savar, Dhaka, Bangladesh, with information from 1032 patients. The algorithm predicts missing values, enhancing the dataset for an automated chronic kidney disease identification system. The research provides a detailed analysis of the dataset, including scenarios with and without missing values. This study extends beyond theoretical principles, offering practical insights into applying rough set

theory to real-world problems. The classification performance of a machine learning algorithm has been used to highlight the significance of missing value prediction via rough set theory. The work can be extended in the future to effectively make informed decisions with respect to more complex data and related challenges by leveraging techniques such as cross validation, etc.

Abbreviation

SVM: Support Vector Machine

MRI: Magnetic resonance imaging

UAGOAAS: Updating Approximations based on Granules w.r.t. Objects and Attributes are Added Simultaneous

RIDAS: rough set based intelligent data analysis system

PSONN: particle swam optimization neural network

CART: Classification And Regression Tree

KDD: Knowledge Discovery Database

Acknowledgement

Nil

Author Contributions

These authors contributed equally to this work.

Conflict of Interest

The authors declare that they have no competing of interest.

Ethics Approval

Not Applicable

Funding

Nil

References

1. Pawlak Z. Rough sets. International journal of computer & information sciences. 1982 Oct;11:341-56.
2. Yeh CC, Chi DJ, Hsu MF. A hybrid approach of DEA, rough set and support vector machines for business failure prediction. Expert Systems with Applications. 2010 Mar 1;37(2):1535-41.
3. Tay FE, Shen L. Economic and financial prediction using rough sets model. European Journal of Operational Research. 2002 Sep 16;141(3):641-59.
4. Ali R, Hussain J, Siddiqi MH, Hussain M, Lee S. H2RM: a hybrid rough set reasoning model for prediction and management of diabetes mellitus. Sensors. 2015 Jul 3;15(7):15921-51.
5. Chen HL, Yang B, Liu J, Liu DY. A support vector machine classifier with rough set-based feature selection for breast cancer diagnosis. Expert systems with applications. 2011 Jul 1;38(7):9014-22.
6. Maji P. Advances in rough set based hybrid approaches for medical image analysis. In Rough Sets: International Joint Conference, IJCRS 2017, Olsztyn, Poland, July 3–7, 2017, Proceedings, Part I 2017 (pp. 25-33). Springer International Publishing.
7. Wang W, Gao W, Wang C, Li J. An improved algorithm for CART based on the rough set theory. In2013 Fourth Global Congress on Intelligent Systems 2013 Dec 3 (pp. 11-15). IEEE.
8. Tsumoto S. Automated extraction of medical expert system rules from clinical databases based on rough set theory. Information sciences. 1998 Dec 1;112(1-4):67-84.
9. Huang H, Meng F, Zhou S, Jiang F, Manogaran G. Brain image segmentation based on FCM clustering algorithm and rough set. IEEE Access. 2019 Jan 15;7:12386-96.
10. Rajesh T, Malar RS, Geetha MR. Brain tumor detection using optimisation classification based on rough set theory. Cluster computing. 2019 Nov;22(Suppl 6):13853-9.
11. Senthil Kumar S, Hannah Inbarani H. Cardiac arrhythmia classification using multi-granulation rough set approaches. International Journal of Machine Learning and Cybernetics. 2018 Apr;9:651-66.
12. Yekkala I, Dixit S. Prediction of heart disease using random forest and rough set based feature selection. International Journal of Big Data and Analytics in Healthcare (IJBDAH). 2018 Jan 1;3(1):1-2.
13. Santra D, Basu SK, Mandal JK, Goswami S. Rough set based lattice structure for knowledge representation in medical expert systems: Low back pain management case study. Expert Systems with Applications. 2020 May 1;145:113084.
14. Xie Y. On medical image filtering based on rough set theory. In2008 Fifth International Conference on Fuzzy Systems and Knowledge Discovery 2008 Oct 18 (Vol. 5, pp. 276-280). IEEE.
15. Hassanien AE, Ali JM. Rough set approach for classification of breast cancer mammogram images. In International Workshop on Fuzzy Logic and Applications 2003 Oct 9 (pp. 224-231). Berlin, Heidelberg: Springer Berlin Heidelberg.
16. Rajesh T, Malar RS. Rough set theory and feed forward neural network based brain tumor detection in magnetic resonance images. In International Conference on Advanced Nanomaterials & Emerging Engineering Technologies 2013 Jul 24 (pp. 240-244). IEEE.
17. Slimani T. Application of rough set theory in data mining. arXiv preprint arXiv:1311.4121. 2013 Nov 17.
18. Kusiak A. Rough set theory: a data mining tool for semiconductor manufacturing. IEEE transactions on electronics packaging manufacturing. 2001 Jan;24(1):44-50.
19. Wang GY, Zheng Z, Zhang Y. RIDAS-a rough set based

intelligent data analysis system. In Proceedings. International Conference on Machine Learning and Cybernetics 2002 Nov 4 (Vol. 2, pp. 646-649). IEEE.
20. Chan CC. A rough set approach to attribute generalization in data mining. Information Sciences. 1998 Jun 1;107(1-4):169-76.
21. Beaubouef T, Ladner R, Petry F. Rough set spatial data modeling for data mining. International Journal of Intelligent Systems. 2004 Jul;19(7):567-84.
22. Guan JW, Bell DA, Liu DY. The rough set approach to association rule mining. In Third IEEE International Conference on Data Mining 2003 Nov 22 (pp. 529-532). IEEE.
23. Griffin G, Chen Z. Rough set extension of Tcl for data mining. Knowledge-Based Systems. 1998 Nov 12;11(3-4):249-53.
24. Lingras P. Rough set clustering for web mining. In2002 IEEE World Congress on Computational Intelligence. 2002 IEEE International Conference on Fuzzy Systems. FUZZ-IEEE'02. Proceedings (Cat. No. 02CH37291) 2002 May 12 (Vol. 2, pp. 1039-1044). IEEE.
25. Chen H, Li T, Luo C, Horng SJ, Wang G. A decision-theoretic rough set approach for dynamic data mining. IEEE Transactions on fuzzy Systems. 2015 Jan 6;23(6):1958-70.
26. Zhu W, Wang FY. A new type of covering rough set. In2006 3rd international IEEE conference intelligent systems 2006 Sep 4 (pp. 444-449). IEEE.
27. W. Grzymala-Busse J. Rough set strategies to data with missing attribute values. In Foundations and novel approaches in data mining 2005 Nov 22 (pp. 197-212). Berlin, Heidelberg: Springer Berlin Heidelberg.
28. Reddy GT, Reddy MP, Lakshmanna K, Rajput DS, Kaluri R, Srivastava G. Hybrid genetic algorithm and a fuzzy logic classifier for heart disease diagnosis. Evolutionary Intelligence. 2020 Jun;13:185-96.
29. Stepaniuk J. Rough set data mining of diabetes data. In International Symposium on Methodologies for Intelligent Systems 1999 Jun 8 (pp. 457-465). Berlin, Heidelberg: Springer Berlin Heidelberg.
30. Dutta S, Ghatak S, Dey R, Das AK, Ghosh S. Attribute selection for improving spam classification in online social networks: a rough set theory-based approach. Social Network Analysis and Mining. 2018 Dec;8:1-6.
31. Kundu S, Pal SK. Double bounded rough set, tension measure, and social link prediction. IEEE Transactions on Computational Social Systems. 2018 Aug 19;5(3):841-53.
32. Kundu S, Pal SK. Fuzzy-rough community in social networks. Pattern Recognition Letters. 2015 Dec 1;67:145-52.
33. Fan TF, Liau CJ. Rough set-based concept mining from social networks. In2016 IEEE International Conference on Fuzzy Systems (FUZZ-IEEE) 2016 Jul 24 (pp. 663-670). IEEE.
34. Fan TF. Rough set analysis of relational structures. Information Sciences. 2013 Feb 1;221:230-44.
35. Khachfeh RA, Elkabani I. Using rough sets in homophily based link prediction in online social networks. In2014 World Congress on Computer Applications and Information Systems (WCCAIS) 2014 Jan 17 (pp. 1-6). IEEE.
36. Yang L, Zhang Z, Pu J. Rough set and trust assessment-based potential paths analysis and mining for multimedia social networks. International Journal of Digital Content Technology and its Applications. 2012 Dec 1;6(22):640.
37. Islam MA, Akter S, Hossen MS, Keya SA, Tisha SA, Hossain S. Risk factor prediction of chronic kidney disease based on machine learning algorithms. In2020 3rd international conference on intelligent sustainable systems (ICISS) 2020 Dec 3 (pp. 952-957). IEEE.
38. Jader RF, Aminifar S, Abd MH. Diabetes detection system by mixing supervised and unsupervised algorithms. Journal of Studies in Science and Engineering. 2022 Sep 13;2(3):52-65.
39. Gaeta A, Loia V, Lomasto L, Orciuoli F. A novel approach based on rough set theory for analyzing information disorder. Applied Intelligence. 2023 Jun;53(12):15993-6014.