EUROPEAN JOURNAL OF PURE AND APPLIED MATHEMATICS
2025, Vol. 18, Issue 2, Article Number 5934
ISSN 1307-5543 – ejpam.com
Published by New York Business Global

Enhancing Classification Performance Through Rough Set Theory Feature Selection: A Comparative Study Across Multiple Datasets

T. Ashika[1], G. Hannah Grace[1,*]

[1] *Department of Mathematics, School of Advanced Sciences, Vellore Institute of Technology Chennai, India*

Abstract. In Machine Learning (ML), handling high-dimensional data with redundant or irrelevant features presents significant challenges. Effective feature selection is essential for enhancing model performance, reducing computational complexity, and improving interpretability. Rough Set Theory (RST) provides a powerful mathematical framework for managing uncertainty, making it a valuable tool for feature selection. This study applies RST-based feature selection to five diverse datasets, aiming to eliminate insignificant attributes. The performance of various ML models, including Logistic Regression (LR), K-Nearest Neighbor (KNN), Support Vector Machine (SVM), Kernel SVM, Naïve Bayes (NB), Decision Tree (DT) and Random Forest (RF), on both the original and RST selected datasets was evaluated. Standard metrics such as accuracy, precision, recall, F1 score and Mean Absolute Error (MAE) are used for evaluation. Results demonstrate that RST effectively selects relevant features without significant information loss. Models trained on RST selected datasets exhibit comparable or improved performance. These findings highlight the potential of RST-based feature selection to enhance ML model performance while reducing computational complexity, making it a valuable approach for various ML applications.

2020 Mathematics Subject Classifications: 68T30, 68T37, 68U35

Key Words and Phrases: Quickreduct, machine learning, rough set theory, feature selection

1. Introduction

A subfield of computer science and artificial intelligence (AI) called "Machine Learning" focuses on using data and algorithms to simulate human learning processes and progressively increase their accuracy. ML techniques have been used in many domains, such as email filtering, computer vision, speech recognition, agriculture, and health, where developing algorithms capable of carrying out the necessary functions would be costly. Although ML has made many revolutions in different fields, it also faces many

*Corresponding author.
DOI: https://doi.org/10.29020/nybg.ejpam.v18i2.5934

Email addresses: ashika.t2023@vitstudent.ac.in (T. Ashika), hannahgrace.g@vit.ac.in (G. H. Grace)

limitations such as lack of data, challenges in obtaining the data, bias in the data, and a vast amount of irrelevant attributes in the data.

Pawlak [1] introduced RST, a technique used in mathematics that can be applied to problems involving imprecision and vagueness in intelligent data mining and analysis for handling huge data. RST provides several benefits, such as the ability to operate without needing prior or supplementary information about the data. It delivers meaningful analysis even when data is incomplete, supports the interpretation of both quantitative and qualitative data, and can model highly nonlinear or discontinuous relationships, offering intricate characterizations of the data. RST excels in uncovering hidden patterns and representing them as decision rules, all of which are grounded in facts supported by a set of examples. RST is a valuable tool for the ML community since it has significant potential in areas including dynamic learning, multi-label classification, imbalanced classification, and big data research [2].
In this paper, five datasets were sourced from Kaggle and the UCI ML repository. The study is organized into two phases. In the first phase, the data was pre-processed, followed by the application of various ML algorithms and an evaluation of their performance. The results of these models were analyzed to assess their classification effectiveness. In the second phase, RST feature selection was applied, involving similar steps of pre-processing and the application of ML algorithms. After evaluating the RST based models, a comparison was made between the results of the original ML models and those utilizing RST based feature selection. This comparative analysis highlights the impact of RST on improving classification performance in various models and provides insight into its potential to improve model accuracy and robustness.

1.1. Related Work

Pawlak [1] was the one who first proposed the theory of uncertainty and RST, and later modified it. The cardiac-related dataset was classified and forecasted using a variety of ML algorithms, such as NB, SVM, DT J48, JRip, Adaboost, KNN, and stochastic gradient descents (SGD), and several others by Mohan *et al.* [3]. Using RST, Mishra *et al.* [4] offered a set of guidelines for solving social science challenges. In order to obtain justice, employees frequently turn to legal authorities, which is harmful to both the employer and the employee. The paper addresses the need for employers in various organizations to provide proper justice. The key characteristics that contribute to this kind of situation in various establishments were identified using RST.
Nayak *et al.* [5] created a model using RST for malaria symptom prediction. In order to identify the accurate symptoms of malaria, the study addresses the application of RST to extract hidden information from a set of imprecise data. Research has indicated that RST is marginally superior to other soft computing techniques and is a valuable tool for handling ambiguous data. Subhalaxmi Das *et al.* [6] has developed a technique for diagnosis of cardiac problem using RST and ML. The study proposes a model to diagnose cardiac disorders by utilizing ML techniques in conjunction with the stacking method. The study uses the Cleveland dataset to diagnose cardiac issues. RST was used

to determine the most significant disease among all cardiac-related disease. The model's performance was assessed using F1 score, recall, accuracy, and precision. Additionally, error functions were computed to evaluate the performance of the models.

Rani Nuraeni and Sugiyarto Surono [7] has proposed RST for dimension reduction on ML algorithm. The use of dimension reduction in ML to increase computational efficiency is covered in this paper. The authors convert high-dimensional data into much lower dimensions without appreciably losing the original characteristics and information of the data using RST. The authors make use of five UCI ML datasets and applied Core and Reduct. The performance of the reduced dataset is evaluated using popular ML techniques like SVM, LR, and KNN. According to the research, core and reduct can reduce computing time by up to 80% without compromising the value of each evaluation model.

A prediction model for building energy consumption based on RST and Deep neural Network (DNN) has been developed by Lei Lei *et al.* [8]. RST is used to minimize unnecessary influencing factors and to identify the essential components of building energy consumption. Next, a DNN receives these essential components. The output of the DNN is the amount of energy consumed by buildings. Initially, data was collected from 100 public buildings in the civil public buildings. Subsequently, DNNs were trained and tested using data gathered from a university lab in Dalian. The study examined both short- and medium-term energy consumption predictions for buildings. The outcomes of the fuzzy, Elman, and backpropagation neural networks were compared with those of the DNN. Compared to previous research works, the combination of RST and DNN turned out to be the most exact.

Lukshmi R.A *et al.* [9] analyzed RST approach for attribute reduction. In order to create an efficient reduct set and formulate the core of the attribute set, the approach focuses on eliminating redundant attributes. Only a subset of the features that maintain the original feature's accuracy are then chosen for further processing using ML models.

Vamsidhar Talasila *et al.* [10] predicted diseases using RST with Recurrent Neural Network (RNN) in Big Data Analytics where RST technique was used to identify the most pertinent features for effective disease detection and medical data classification. The RNN method for disease prediction receives the chosen features as input. The experiments for the proposed method, also known as RST-RNN, are conducted on the UCI ML repository dataset with respect to accuracy, F-measure, sensitivity, and specificity. The outcomes demonstrated that the RST-RNN approach obtained accuracy of 98.57 % for the heart disease dataset.

Touhid Mohammad Hossain *et al.* [11] had discussed a RST approach along with ML in Electrofacies Classification and Subsurface Lithology Interpretation. In order to classify electrofacies, the paper suggests a RST-based White box classification method that generates decision rules. Additionally, the Extra Tree Classifier (ETC) was used to select significant well log features, from which they created five distinct electrofacies. Yılmaz Kayaa and Murat Uyar [12] proposed the use of a hybrid decision support system to identify hepatitis. RST and Extreme Learning Machine (ELM) algorithms form the foundation of the system. The proposed model was trained using the Indian Liver pa-

tient dataset, then evaluated the effectiveness of the suggested classification methods based on sensitivity and specificity to determine the model's accuracy.

A SVM classifier for breast cancer diagnosis using RST-based feature selection was proposed by Hui-Ling Chen *et al.* [13]. Using RST reducts, Asma Trabelsi *et al.*[14] created an ensemble classifier to handle data with evidential attributes. The objective is to use RS reduct to extract the best feature subsets for the ensemble classifier. The Diversity Reduct method (DR), the Accuracy-Diversity Assessment Function method (AD-AF), and the Ensemble Accuracy Assessment Function method (EA-AF) are the three strategies that have been suggested to choose the best reducts. The key differences between the previously mentioned methods have been compared using a statistical test in terms of reduct diversity, ensemble size, classification performance, and approach performance. The outcomes show that the EA-AF strategy generated the best effects.

Ping-Feng Pai *et al.* [15] analysed foreign exchange rates by RST and directed acyclic graph SVM (DAGSVM). A new approach to foreign exchange rate analysis that combines DAGSVM and RST was proposed. They examined the US dollar's exchange rates versus the euro and the Japanese yen using the suggested methodology. The outcome indicates that this method outperforms alternative strategies like Back Propagation Neural Network (BPNN) and Support Vector Regression (SVR). For real-valued classification problems, Essam Debie *et al.*[16] proposed an ensemble of learning classifier system based on reducts. An ensemble approach for RST-based learning classifier systems was proposed to construct ensemble learning systems. By using RST attribute reduction, the method creates a set of reducts. A variety of these reducts is then chosen and utilized for training of ensemble of base classifiers. According to the analysis, single learning classifier system models are outperformed by reduct-based ensemble systems in terms of classification accuracy.

Georg Peters and Simon Poon [17] has discussed a Dominance based RS approach (DRSA) by analyzing IT business values. This paper examine how IT affects a partnership's business value. A decision table with ordinal data can be used to generate rules using DRSA. An investigation into the IT management practices of Australian businesses was carried out by the Department of Communications, Information Technology, and the Arts in Australia, from which the authors have used ordinal data. The paper offers a thorough analysis of the underlying mechanisms that underpin IT beneficial effects on cooperatives business values.

BrightBox is a RST-based technology designed by Andrzej Janusz *et al.* [18] for identifying errors in ML models. A new method for looking into errors in ML model operations is called BrightBox. The method relies on surrogate RST-based models that are inducted from data approximating the decisions made by the monitored Blackbox models. The neighborhoods of instances that go through the diagnostic process are computed using these approximators which consists of historical instances that were handled similarly by RST-based models. The analysis of mistakes reported in these neighborhoods serves as the foundation for the diagnostic procedure. The study also emphasizes how BrightBox technology can be applied to model and data-related problem identification and analysis as well as prediction model analysis. Experiments conducted on real-world

datasets verify that this type of analysis may provide us efficient and accurate insights into the reasons of ML model's poor performance. Alessio Ferone [19] had proposed feature selection determined by the RST composition resulted by feature granulation. This research suggests a RST-based feature selection method that consists of two stages: Granulation and Roughinement. The suggested method reduces computation costs with respect to the Quickreduct (QR) algorithm.

In recent years, several studies have advanced the application of RST in feature selection, addressing various challenges in dynamic and high-dimensional datasets. Zhao *et al.*[20] proposed a triple nested equivalence class (TNEC) approach for group incremental feature selection, which effectively reduces redundancy and enhances efficiency in processing ultra-high-dimensional datasets. Their experimental results demonstrated significant improvements in classification efficiency and selection accuracy across 18 dynamic datasets. Similarly, Almania *et al.* [21] introduced a two-level feature selection approach that combines RST with Binary Particle Swarm Optimization (BPSO) to enhance accuracy and reduce computational overhead in Intrusion Detection Systems (IDS). Their method achieved a notable improvement in classification accuracy while streamlining the feature set.

Furthermore, Alsabilah and Rawat [22] leveraged RST in conjunction with XGBoost to develop a robust network IDS , significantly outperforming existing methods in terms of accuracy and precision. Turaga and Chebrolu [23] addressed the inefficiencies of traditional RST algorithms by proposing a novel attribute reduction algorithm designed for execution on Graphics Processing Units (GPUs), achieving high classification accuracies in reduced execution times. These studies collectively highlight the evolving landscape of RST applications and underscore the importance of integrating RST with modern computational techniques to enhance feature selection processes. The summary of the related work is provided in table 1.

Table 1: Comparison of RST-based and Non-RST Feature Selection Methods from Recent Literature

Method	**References**	**Approach**	**Advantages**	**Limitations**
RST	[1, 5, 6]	Rule-based reduction using core and reducts	Handles uncertainty, generates human-understandable rules, supports imprecise data	Needs discretization, can be computationally intensive for large feature sets
RST vs Core + Reduct	[7]	Compared performance with and without RST-based dimensionality reduction	Reduced computation time by up to 80% with minimal accuracy loss	Focused on only a few ML models (SVM, LR, KNN)
RST + DNN	[8]	RST for dimension reduction before Deep Learning (DL)	Enhanced prediction accuracy by removing noisy/irrelevant features before DNN	Requires tuning both RST reduct and DL model
RST + RNN	[10]	Feature selection via RST, then fed to RNN	Achieved 98.57% accuracy on heart disease dataset	Limited to specific dataset; needs sequential data handling
RST + ETC	[11]	RST to identify features for lithology; ETC to classify	White-box interpretability with competitive classification performance	Focused on geophysical/electrofacies domain
Non-RST Feature Selection	[3, 12]	Used traditional ML or hybrid models (e.g., ELM, SVM) without RST	Simpler implementation, fast training on structured datasets	May not handle ambiguity/uncertainty well; lacks interpretability of RST rules
RST Ensemble Feature Selection	[14]	DR, AD-AF, EA-AF for best reducts in ensemble classifiers	EA-AF achieved highest classification performance using feature diversity metrics	Needs statistical evaluation and multiple reduct generation
RST + DRSA	[17]	Dominance relations used in ordinal decision tables	Suitable for subjective/ordinal data (e.g., IT business values)	More complex than standard RST; limited applicability
RST + TNEC	[20]	Group incremental feature selection for high-dimensional dynamic data	Reduces redundancy, improves classification accuracy on large-scale datasets	Focused on dynamic datasets, complex implementation
RST + BPSO	[21]	Two-level hybrid feature selection for IDS	High accuracy and reduced overhead; robust for cybersecurity	Tailored to IDS; parameter tuning required

1.2. Motivation and Novelty

ML has revolutionized various fields by enabling systems to learn from data and make informed decisions. However, the presence of irrelevant or redundant features in datasets can hinder model performance, leading to overfitting, increased computational costs, and reduced interpretability. Traditional feature selection methods often rely on statistical measures or domain knowledge, which may not always be available or sufficient. This study is motivated by the need for an effective, data-driven approach to feature selection that can handle uncertainty and imprecision inherent in real-world datasets. This research introduces a novel application of RST for feature selection in ML workflows. The novelty lies in integrating RST-based feature selection with ML models across diverse domains, providing a comprehensive evaluation of its impact on model performance.

1.3. Contribution

The contribution of this study is provided below

(i) Application of RST-based feature selection to five diverse datasets, encompassing domains such as Seed, Heart Attack, Obesity, Online Delivery, and Breast Cancer. This extensive evaluation provides insights into the applicability and effectiveness of RST across various contexts.

(ii) Comparative analysis of ML models trained on datasets with and without RST-based feature selection is conducted. The results indicate that RST either enhances or maintains model performance, highlighting its potential as a valuable tool in ML workflows.

(iii) The findings suggest that integrating RST into feature selection processes can lead to more efficient and accurate ML models, with potential applications in various fields.

By addressing the limitations of traditional feature selection methods and demonstrating the efficacy of RST, this work contributes to the advancement of ML techniques and their practical applications. The structure of the paper is as follows: Section 2 presents the materials and methods. The Research findings of the suggested model are covered in Section 3. Section 4 discusses the paper and Section 5 concludes the suggested work.

2. Materials and Methods

This section outlines the materials and methods utilized in this study, including the proposed methodology, description of the datasets, the data preprocessing steps, the application of RST for feature selection, the ML models employed, and the evaluation metrics used to assess model performance.

2.1. Proposed work

The paper proposes two process which includes classification without RST and classification with RST. Five different datasets with varying number of attributes and variables are chosen. Firstly, these datasets were preprocessed and evaluated using various ML classifiers like LR, KNN, SVM, Kernel SVM, NB, DT and RF. The model's performance is then assessed using evaluation metrics, which include MAE, accuracy, precision, recall, F1 score, and confusion matrix. Secondly, the insignificant features of these five datasets are reduced using RST reduct and core. These reduced dataset are evaluated using the ML techniques like LR, KNN, SVM, Kernel SVM, NB, DT and RF and the performance of these models is evaluated and compared with the original dataset to analyse whether the RST selected features improve performance compared to the original dataset. Suitable ML algorithms for each dataset were also identified based on the comparative analysis. The architecture of the suggested work is shown in figure 1.

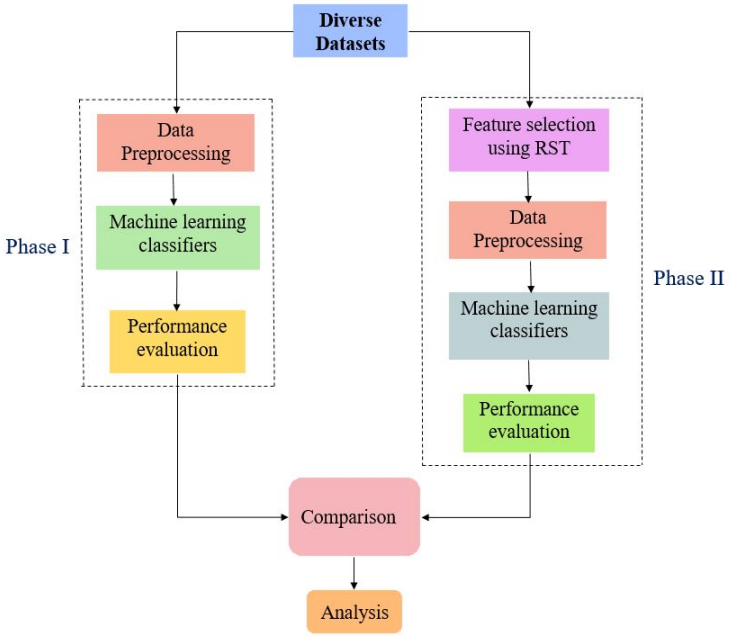

Figure 1: Overall Architecture of the suggested work

2.2. Dataset Description

The study utilized five datasets sourced from Kaggle and the UCI ML Repository, a widely recognized platform for its diverse and comprehensive data. This selection provides a robust foundation for evaluating the effectiveness of classification models both with and without RST based feature selection. The selected datasets include Seeds [24], online food delivery [25], Heart attack [26], Obesity [27] and Breast cancer [28]. The Seeds dataset features 7 conditional attributes and 1 decision attribute to clas-

sify wheat types. The Online Food Delivery dataset, focusing on the Bangalore region, includes 54 conditional attributes and 1 decision attribute, examining the growth in demand for food delivery. The Heart Attack dataset, with 54 conditional attributes and 1 decision attribute, identifies predictors of heart disease and classifies outcomes as positive (disease present) or negative (absent). The Obesity dataset contains 6 conditional attributes and 1 decision attribute, categorizing individuals into weight classifications such as Overweight, Underweight, Normal, or Obese. For Breast Cancer, the Wisconsin dataset includes 8 conditional attributes and 1 decision attribute to classify tumors as benign or malignant. Together, these datasets provide a diverse and structured basis for classification and prediction in the study. Table 2 presents a list of datasets, detailing the number of attributes and instances for each dataset.

Table 2: List of Dataset along with the number of features and instances

S. No	Datasets	Number of features	Conditional Attribute	Decision attribute	Number of Instances
1	Seeds	8	7	1	211
2	Online Food Delivery	55	54	1	387
3	Heart attack	9	8	1	1319
4	Obesity	7	6	1	108
5	Breast Cancer	9	8	1	699

2.3. Data preprocessing

The dataset might include insufficient data, missing data or categorical data. In order to convert this raw dataset into a clean dataset, data preprocessing is performed. Data preprocessing techniques include: handling missing data, splitting the dataset, feature scaling, and encoding categorical data. Missing values were imputed by replacing them with the column average to mitigate potential errors. Categorical features were converted into numerical format using one-hot encoding to create binary vectors for each category. In addition, standardization was applied to scale all features uniformly, typically adjusting values to a range between -3 and +3. Finally, each dataset was divided into training and testing sets using an 80:20 split, ensuring a consistent evaluation framework across the diverse datasets.

2.4. Rough set theory

A mathematical tool called RST is used to deal with uncertain and inconsistent data. It was first proposed by Pawlak [1] at the beginning of the 1980's. RST expresses uncertainty and imprecision through a set's boundary region rather than through partial membership function as Fuzzy Set Theory does. RST consists of lower and upper approximations, a boundary region, and the concepts of indiscernibility, which lead to the identification of reducts and cores. RST allows for a crisp set to be formally approximated using the lower and upper approximations of the original set. To enable effective classification, the most pertinent features are selected using the RST technique. Let

$I = (U, A)$ represent an information system, where U denotes a nonempty set of finite objects that is the universe of discourse. A is non-empty. An associated set of values (V_a) exists for each attribute $a \in A$. For a subset, $X \subseteq U$, X can be approximated using only information contained within P by constructing the P-lower approximation given as P_*X, is the set of all elements of U, which can be certainly classified as elements of X based on the attribute set P. The P- upper approximation of X, denoted as P^*X, which can be possibly classified as elements of X based on the attribute set P. The lower and upper approximation can be defined as shown in equations (1) and (2),

$$P_*(X) = \{x \in U : P(x) \subseteq X\} \tag{1}$$

$$P^*(X) = \{x \in U : P(x) \cap X \neq \emptyset\} \tag{2}$$

The difference of the upper P^*X and the lower approximation P_*X yields the boundary region $PN(X)$ as shown in equation (3).

$$PN(X) = P^*(X) - P_*(X) \tag{3}$$

The accuracy of approximation $\alpha_P(X)$ is shown in equation (4) where the resultant values lies in the range of $0 \leq \alpha_P(X) \leq 1$.

$$\alpha_P(X) = \frac{|(P_*X)|}{|(P^*(X))|} \tag{4}$$

Dependency between the decision attribute D and the condition attribute C, where D depends on C in a degree $k(0 \leq k \leq 1)$, denoted $C \Rightarrow_k D$, if $k = \gamma(C, D)$ are represented as given in equation (5) and (6).

$$\gamma(C, D) = \frac{|(POS_C(D))|}{|U|} \tag{5}$$

$$POS_C(D) = \cup P_*(X) \tag{6}$$

where, the set of all elements of U that can be specifically assigned to blocks of the partition U/D by means of C is known as $POS_C(D)$ also known as the positive region of the partition $\frac{U}{D}$ with respect to C. Definition of the indiscernibility relation $IND(P)$ is shown in equation (7).

$$IND(P) = \{(x, y) \in U^2 | \forall a \in P, a(x) = a(y)\} \tag{7}$$

The minimum subset of attributes known as a reduct that permits the same classification of universe's elements as the entire set of attributes is known as a reduct where if $I(P) = I(P - \{a\})$, then 'a' is dispensable in P; otherwise, 'a' is indispensable in P. Set P is independent if all its attributes are indispensable. Subset P' of P is a reduct of P if P' is independent and $I(P') = I(P)$. Assume P is a subset of A. The combination of all the vital attributes that make up P is its core. Core is the intersection of all reducts as shown in equation (8).

$$Core(P) = \cap Red(P) \tag{8}$$

$Red(P)$ denotes the set of all reduct of P. Numerous applications of RST have shown it to be effective, including data reduction, identifying hidden trends in data, assessing the significance of data, and creating sets of decision rules based on data.

2.4.1. RST QR Algorithm

QR algorithm [29] constructs a reduct by sequentially adding those attributes that most significantly enhance the dependency degree, halting the process when no further improvement is observed. Because it is computationally impractical to evaluate all possible feature combinations, QR adopts a greedy strategy by selecting one attribute at a time for inclusion. Algorithm 1 provides the implementation of the QR approach.

Algorithm 1: QR Algorithm
Input: C : Set of all Conditional attributes D : Set of Decision attribute **Output:** A : Minimal Subset of C that retains the same classification power as the full set of attributes C w. r. t. D
Step 1: Start with empty set A $A \leftarrow \emptyset$ **Step 2:** Iterate until the discernibility measure of A equals to C **Step 3:** Repeat until $\gamma_A(D) = \gamma_C(D)$ *Step 3.1:* Assign $T = A$ *Step 3.2:* For each $x \in C - A$: Compute, $\gamma_{R(A\cup\{x\})}(D)$ *Step 3.3:* If $\gamma_{R(A\cup\{x\})}(D) > \gamma_T(D)$, Update, $T \leftarrow A \cup \{x\}$ *Step 3.4:* Set $A \leftarrow T$ **Step 4:** Return A
where, $\gamma_A(D) \rightarrow$ Discernibility of A w. r. t. D $\gamma_C(D) \rightarrow$ Discernibility of C w. r. t. D $A \rightarrow$ Set of selected attributes $T \rightarrow$ Temporary set to evaluate attribute addition

2.5. Models of ML

For the study, several well-established ML classifiers are employed to evaluate the diverse datasets. LR [30] was chosen for its simplicity in modeling binary outcomes using a logistic function, while KNN [31] offers a non-parametric approach based on euclidean distance to classify data points. SVM [32] was included for its robust ability to find optimal decision boundaries, and its extension, Kernel SVM [33], enables non-linear classification by mapping inputs into high-dimensional feature spaces. Additionally, NB [31] was utilized for its probabilistic framework based on Bayes' Theorem and the assumption of predictor independence, and DT [31] was selected for its clear, interpretable structure driven by entropy measures. To further enhance prediction accuracy and address complex patterns, ensemble methods such as RF[34]. These classifiers were chosen for their complementary strengths and proven effectiveness in tackling varied classification challenges.

2.6. Evaluation metrics

Evaluation metrics like accuracy, precision, recall, F1-score, and MAE rate are used to assess the effectiveness of the suggested approach. The percentage of all subjects

correctly classified is known as accuracy. The percentage of subjects correctly classified as positive out of all positive subjects is known as precision. The recall evaluates the model's ability to recognize positive samples. A harmonic mean of recall and precision is the F1-Score. A measure of errors between paired observations expressing the same phenomenon is called MAE. The Formula of the evaluation metrics and error rate is shown in table 3.

Table 3: Formulae of Evaluation metrics and error function

S. No.	Evaluation metrics	Formulae
1	Accuracy	$\frac{(TP+TN)}{(TP+TN+FP+FN)}$
2	Precision	$\frac{(TP)}{(TP+FP)}$
3	Recall	$\frac{(TP)}{(TP+FN)}$
4	F1 score	$\frac{2 \times Recall \times Precision}{Recall + Precision}$
5	MAE	$\frac{1}{n}\sum_{i=1}^{n} \lvert y_i - x_i \rvert$

3. Experimental results

The study results include the following steps: feature selection using RST with the QR algorithm, classification with ML algorithms using the original features, classification using features selected by RST, and a comparison of the results before and after feature selection.

3.1. Experimental Setup

The experiments were conducted using R programming, specifically utilizing the RoughSets package for feature selection. Model training was performed in Google Colab, a cloud-based Jupyter notebook environment, utilizing PyTorch for ML implementation. The system used for training was equipped with a 12th Gen Intel Core i3-1215U processor (1.20 GHz) and 8 GB RAM, ensuring efficient data processing and optimal performance across the five different datasets.

3.2. Feature selection

In this study, the QR algorithm of RST is utilized to determine the reducts, enabling the selection of the most relevant features while discarding irrelevant ones. The seeds dataset was reduced from 7 attributes to 2 attributes. The online food delivery dataset was reduced from 54 attributes to 4 attributes. The Heart attack dataset was reduced from 8 attributes to 2 attributes. The Obesity dataset was reduced from 6 attributes to 5 attributes. The Breast cancer dataset was reduced from 8 attributes to 4 attributes. The dataset before reduction and after reduction is shown in table 4.

Table 4: List of Dataset along with attribute list before and after reduction

S. No	Datasets	No. of attributes before reduction	No. of attributes after reduction
1	Seeds	7	2
2	Online Food Delivery	54	4
3	Heart attack	8	2
4	Obesity	6	5
5	Breast Cancer	8	4

3.3. Classification with ML Algorithms Using Original Features

The original dataset is pre-processed and then evaluated using various ML classifiers like LR, KNN, SVM, Kernel SVM, NB, DT, and RF. The model's performance is assessed using evaluation metrics. Figure 2 shows that certain algorithms perform best across different datasets: SVM works well for Seeds, while DT and RF are effective for Heart Attack and Obesity. In Online Food Delivery, LR, SVM, Kernel SVM, NB, and RF shows similar performance, and NB yields the best results in Breast Cancer.

Figure 2: ML Algorithm Performance with Original Features across Datasets

3.4. Classification with ML Algorithms Using Features selected by RST

This section discusses the performance evaluation after applying RST feature selection, focusing on the performance of selected features. From figure 3, it is evident that after applying RST-selected features, LR, SVM, and Kernel SVM perform well for the Seeds dataset; RF achieves the best performance for Heart Attack; DT is most effective for Obesity; KNN, Kernel SVM, and DT perform well for Online Food Delivery; and KNN yields the best results for Breast Cancer.

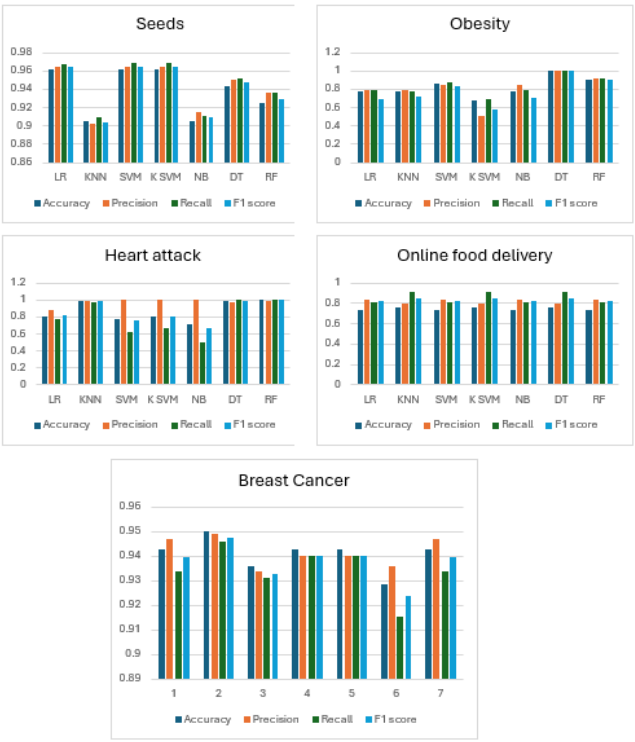

Figure 3: ML Algorithm Performance with RST-Selected Features across different Datasets

3.5. Comparison of Results Before and After Feature Selection using RST

This section discusses the performance changes after applying RST feature selection. By comparing the results before and after feature selection, The study evaluate how RST affects classification accuracy, precision, recall, and other metrics across various ML algorithms. This analysis showcases how reducing the feature set to the most relevant attributes can improve model performance and efficiency.

3.5.1. Seeds Dataset

Table 5, shows that applying RST for feature selection for the Seeds dataset results enhances model performance across metrics, with substantial gains observed in models such as LR, SVM, and Kernel SVM, where accuracy, precision, recall, and F1 scores all increase significantly. While RF shows only a slight improvement, most other models benefit notably from the refined feature set, confirming the effectiveness of RST in boosting classification metrics.

Table 5: Different ML algorithms on Seeds dataset

ML model	Accuracy		Precision		Recall		F1 score		MAE	
	Before *RST*	After *RST*	Before *RST*	After *RST*	Before *RST*	After *RST*	Before *RST*	After *RST*	Before *RST*	After *RST*
LR	0.9245	0.9623	0.9265	0.9649	0.9328	0.968	0.9266	0.9648	0.0943	0.0377
KNN	0.8868	0.9056	0.8844	0.9025	0.8936	0.9094	0.8870	0.9043	0.1509	0.1321
SVM	**0.9434**	**0.9623**	**0.9421**	**0.9649**	**0.9486**	**0.9683**	**0.9440**	**0.9648**	**0.0755**	**0.0377**
Kernel SVM	0.9245	0.9623	0.9233	0.9649	0.9290	0.9683	0.9256	0.9648	0.0943	0.0377
NB	0.8868	0.9057	0.9010	0.9158	0.9010	0.9106	0.8947	0.9101	0.1132	0.1132
DT	0.9245	0.9434	0.9216	0.95	0.9290	0.9524	0.9233	0.9473	0.1132	0.0566
RF	0.9245	0.9245	0.925	0.9365	0.9264	0.9365	0.9241	0.9298	0.1132	0.0755

3.5.2. Heart attack dataset

Table 6 shows that applying RST feature selection for the Heart Attack dataset significantly enhances classification performance for several models. LR improving from 0.7424 to 0.8101, and Kernel SVM, DT, and RF achieving high metrics across the board, highlighting the effectiveness of RST in enhancing model precision and accuracy.

Table 6: Different ML algorithms on Heart Attack dataset

ML model	Accuracy		Precision		Recall		F1 score		MAE	
	Before *RST*	After *RST*	Before *RST*	After *RST*	Before *RST*	After *RST*	Before *RST*	After *RST*	Before *RST*	After *RST*
LR	0.7424	0.8101	0.7471	0.8864	0.8355	0.7697	0.7888	0.8239	0.2576	0.1894
KNN	0.625	0.9811	0.67096	0.9868	0.6842	0.9803	0.6775	0.9835	0.375	0.0189
SVM	0.7841	0.7803	0.8025	1	0.8289	0.6184	0.8155	0.7642	0.2159	0.2197
Kernel SVM	0.6856	0.8106	0.6885	1	0.8289	0.6711	0.7522	0.8031	0.3144	0.1894
NB	0.6894	0.7121	0.9861	1	0.4671	0.5	0.6339	0.6667	0.3106	0.2879
DT	0.9962	0.9886	0.9935	0.9806	1	1	0.9967	0.9902	0.0038	0.0114
RF	**0.9924**	**0.9962**	**1**	**0.9935**	**0.9868**	**1**	**0.9934**	**0.967**	**0.0076**	**0.0038**

3.5.3. Obesity dataset

RST feature selection maintains high classification performance for DT with perfect metrics across all evaluation measures, while RF experiences a slight decline in accuracy from 1 to 0.9091, showcasing the robustness of DT and RF even after feature selection.

Table 7: Different ML algorithms on Obesity dataset

ML model	Accuracy		Precision		Recall		F1 score		MAE	
	Before *RST*	After *RST*	Before *RST*	After *RST*	Before *RST*	After *RST*	Before *RST*	After *RST*	Before *RST*	After *RST*
LR	0.7727	0.7727	0.7917	0.7917	0.7917	0.7917	0.6881	0.6881	0.3182	0.3182
KNN	0.7727	0.7727	0.7875	0.7875	0.7708	0.7708	0.7261	0.7261	0.4545	0.4545
SVM	0.8636	0.8636	0.8542	0.8542	0.8889	0.875	0.8496	0.83095	0.2727	0.2272
Kernel SVM	0.6818	0.6818	0.5063	0.5063	0.6875	0.6875	0.5761	0.5761	0.6364	0.6364
NB	0.7727	0.7727	0.8438	0.8438	0.7917	0.7917	0.7078	0.7078	0.2273	0.2273
DT	**1**	**1**	**1**	**1**	**1**	**1**	**1**	**1**	**0**	**0**
RF	1	0.9091	1	0.9167	1	0.9167	1	0.9	0	0.1818

Table 7 shows that the DT classifier achieved perfect scores both before and after RST-based feature reduction for the Obesity dataset. Given the small dataset size, such uniform performance strongly indicates overfitting, where the model memorizes patterns including noise rather than generalizing effectively. In contrast, the RF model, which also achieved 1.0 before RST, dropped slightly to 0.9091 after RST, reflecting its ensemble robustness and sensitivity to the removal of marginally informative features. Compared to SVM (0.8636) and LR (0.7727), RF maintains stronger generalization, making it a more reliable choice for small datasets under feature reduction.

3.5.4. Online food delivery dataset

RST feature selection leads to improved performance in classifiers like KNN, Kernel SVM, and DT, especially enhancing recall and F1 scores across these models. Table 8 describes the online food delivery dataset, where some algorithms show improvement, while others yield comparable results.

Table 8: Different ML algorithms on Online food Delivery dataset

ML model	Accuracy		Precision		Recall		F1 score		MAE	
	Before *RST*	After *RST*	Before *RST*	After *RST*	Before *RST*	After *RST*	Before *RST*	After *RST*	Before *RST*	After *RST*
LR	0.7526	0.7308	0.7526	0.8305	1	0.8167	0.8588	0.8235	0.2474	0.2692
KNN	**0.7216**	**0.7564**	**0.7614**	**0.7971**	**0.917**	**0.9167**	**0.8323**	**0.8527**	**0.2784**	**0.2436**
SVM	0.7526	0.7308	0.7526	0.8305	1	0.8167	0.8588	0.8235	0.2474	0.2692
Kernel SVM	0.7526	0.7564	0.7526	0.7971	1	0.9167	0.8588	0.8527	0.2474	0.2436
NB	0.7526	0.7308	0.7526	0.8305	1	0.8167	0.8588	0.8235	0.2474	0.2692
DT	0.7423	0.7564	0.75	0.7971	0.9863	0.9167	0.8521	0.8527	0.2577	0.2436
RF	0.7526	0.7308	0.7526	0.8305	1	0.8167	0.8588	0.8235	0.2474	0.2692

3.5.5. Breast cancer dataset

RST feature selection results in consistent performance for most classifiers, with KNN showing a slight reduction in precision but maintaining high recall and F1 scores, while RF and NB demonstrate stable accuracy, highlighting the utility of RST in retaining essential classification performance. Table 9 describes the breast cancer dataset, where RST yields comparable results.

Table 9: Different ML algorithms on Breast cancer dataset

ML Model	Accuracy		Precision		Recall		F1 score		MAE	
	Before RST	After RST	Before RST	After RST	Before RST	After RST	Before RST	After RST	Before RST	After RST
LR	0.9571	0.9429	0.9551	0.9467	0.9551	0.9337	0.9551	0.9393	0.0857	0.1143
KNN	0.9643	0.95	0.9640	0.94896	0.96096	0.94598	0.9624	0.9474	0.0714	0.1
SVM	0.9572	0.9357	0.9551	0.9339	0.9551	0.9310	0.9551	0.9324	0.0857	0.1286
Kernel SVM	0.9571	0.9429	0.9551	0.9401	0.9551	0.9401	0.9551	0.9401	0.0857	0.1143
NB	0.9643	0.9429	0.9594	0.9401	0.9674	0.9401	0.9629	0.9401	0.0714	0.1143
DT	0.9357	0.9286	0.9461	0.9356	0.9214	0.9155	0.9309	0.9235	0.1286	0.1429
RF	**0.9286**	**0.9429**	**0.9313**	**0.9467**	**0.9187**	**0.9337**	**0.9241**	**0.9393**	**0.1429**	**0.1143**

In conclusion, the comparison of results before and after applying RST feature selection highlights its significant impact on enhancing classification performance across various datasets. RST consistently improves or maintains high performance for multiple ML models. While some models, such as LR and Kernel SVM, show considerable improvements, others like RF exhibit slight variations in performance, yet still retain their robustness. Overall, RST proves to be an effective method for reducing dimensionality while preserving or enhancing essential classification capabilities, thus improving model efficiency and performance across diverse datasets.

3.6. Statistical validation

In this study, the impact of RST based feature selection on the performance of classification models across five datasets was evaluated. To statistically validate the observed improvements, paired t-tests on performance metrics was conducted, comparing model results before and after RST-based feature selection. The p-values for each performance metric across the five datasets are summarized in table 10.

Table 10: P-values from Paired t-Tests Comparing Performance Metrics Before and After RST-Based Feature Selection

Metric	**Seeds**	**Heart Attack**	**Obesity**	**Online Food Delivery**	**Breast Cancer**
Accuracy	0.1536	0.3559	0.5902	0.0524	0.00468
Precision	0.0455	0.3559	0.0001	0.0632	0.00122
Recall	0.8392	0.2814	0.0042	0.0489	0.00254
F1 Score	0.2754	0.2743	0.0827	0.0522	0.00248

Based on the results presented in table 10, the paired t-tests confirm that precision and recall show statistically significant improvements in several datasets, particularly obesity, online food delivery, and breast cancer. This indicates that RST-based feature selection enhances the model's ability to accurately identify and classify positive instances. While accuracy did not significantly improve in most cases, marginal gains were observed in the online food delivery and breast cancer datasets. Similarly, F1-score showed borderline significance in some scenarios, suggesting a balanced enhancement in both precision and recall. By evaluating a range of performance metrics, comprehensive assessment of model behaviour is ensured. Notably, improvements in precision and recall are especially critical in medical and sensitive applications, where minimizing false positives and false negatives is essential. The inclusion of statistical validation addresses the credibility of the findings. These results support the conclusion that RST-based feature selection is an effective and valuable technique for enhancing classification performance across diverse datasets, reinforcing its potential for broader application in ML tasks.

4. Discussion

Although RST may result in marginally lower classification metrics in certain cases, its value extends beyond predictive accuracy. As a feature selection method, RST excels in enhancing model interpretability, reducing computational complexity, and minimizing overfitting. These advantages are particularly important in fields like healthcare, where both transparency and efficiency are crucial. The results of this study indicate that RST-based feature selection consistently preserves or improves classification performance across various datasets. For instance, in the high-dimensional Online Food Delivery dataset (54 features), RST reduced the feature set to just 4 attributes, improving KNN accuracy from 0.7216 to 0.7564. In the low-dimensional Seeds dataset, RST

reduced the features from 7 to 2, increasing SVM accuracy from 0.9421 to 0.9649.
Larger datasets, such as the Breast Cancer dataset (699 instances), also showed improvements (e.g., RF accuracy increased from 0.9286 to 0.9429), demonstrating that RST scales effectively with dataset size. Model-specific analysis revealed that algorithms sensitive to feature space, such as LR and SVM, benefitted the most from RST, as seen in the Seeds dataset, where LR improved from 0.9245 to 0.9623. On the other hand, ensemble models like RF, which are robust to redundant features, showed smaller improvements or slight performance degradation, such as in the Obesity dataset (RF accuracy dropped from 1.0 to 0.9091). RST's impact also varied depending on the dataset characteristics. In balanced datasets like Seeds, improvements were consistent across classifiers, while in imbalanced datasets like Heart Attack, the effects were more varied. For example, KNN in Heart Attack improved significantly from 0.625 to 0.9811, likely due to RST's focus on features associated with the majority class.
Despite these benefits, RST has limitations, including the need for discretization of continuous attributes, which can introduce quantization bias or reduce information precision. Additionally, RST can be sensitive to noisy or overlapping data. The QR algorithm, while computationally efficient, may not always identify the globally optimal reduct. Future research could address these issues by integrating advanced discretization techniques, noise handling methods, and metaheuristic-based feature selection strategies to enhance RST's robustness and scalability.

5. Conclusion

The study presents a two-phase evaluation framework to assess the impact of RST on classification performance. In the first phase, standard ML models are applied to raw datasets. In the second phase, RST-based feature selection was implemented using the QR algorithm, and classifier performance was re-evaluated on the reduced feature sets. Overall, RST generally enhanced or maintained classification performance across the five datasets examined. Notable improvements were observed for models such as LR, SVM, and KNN, particularly in datasets where redundant or noisy features were effectively eliminated. However, slight performance declines occurred in certain cases, for example, RF on the Obesity dataset (accuracy declined from 1.0 to 0.9091) and SVM on the Breast Cancer dataset (from 0.9572 to 0.9357). These declines may be attributed to the removal of minor yet informative features during the reduction process. Such outcomes underscore that while RST enhances model interpretability and computational efficiency, its effectiveness can vary depending on dataset-specific factors such as sample size, class balance, and feature dependencies.
To address the limitations of the QR algorithm and improve generalizability, future research could explore hybrid approaches. Integrating RST with metaheuristic optimization techniques such as Genetic Algorithms may enhance reduct selection by evolving multiple candidate subsets in parallel potentially identifying more discriminative feature sets for complex, high-dimensional datasets. Additionally, combining RST with DL could capitalize on RST's interpretability and DL's representational power. For in-

stance, RST-reduced features could be used as inputs to feedforward or convolutional neural networks, reducing training time and mitigating overfitting by focusing on the most relevant predictors.

References

[1] Zdzisław Pawlak. Rough sets. *International journal of computer & information sciences*, 11:341–356, 1982.

[2] Rafael Bello and Rafael Falcon. Rough sets in machine learning: a review. *Thriving Rough Sets: 10th Anniversary-Honoring Professor Zdzisław Pawlak's Life and Legacy & 35 Years of Rough Sets*, pages 87–118, 2017.

[3] Senthilkumar Mohan, Chandrasegar Thirumalai, and Gautam Srivastava. Effective heart disease prediction using hybrid machine learning techniques. *IEEE access*, 7:81542–81554, 2019.

[4] Sujogya Mishra, SP Mohanty, and SK Pradhan. Reasons for employees need justice from legal bodies: A rough set approach. In *2016 3rd International Conference on Computing for Sustainable Global Development (INDIACom)*, pages 2018–2022. IEEE, 2016.

[5] Subrata Kumar Nayak, Sateesh Kumar Pradhan, Sujogya Mishra, Sipali Pradhan, and PK Pattnaik. Rough set technique to predict symptoms for malaria. In *2021 8th International Conference on Computing for Sustainable Global Development (INDIACom)*, pages 312–317. IEEE, 2021.

[6] Subhalaxmi Das, Sateesh Kumar Pradhan, Sujogya Mishra, Sipali Pradhan, and PK Pattnaik. Diagnosis of cardiac problem using rough set theory and machine learning. *Indian Journal of Computer Science and Engineering*, 13(4):1112–1131, 2022.

[7] Rani Nuraeni and Sugiyarto Surono. Rough set theory for dimension reduction on machine learning algorithm. *Jurnal Fourier*, 10(1):29–37, 2021.

[8] Lei Lei, Wei Chen, Bing Wu, Chao Chen, and Wei Liu. A building energy consumption prediction model based on rough set theory and deep learning algorithms. *Energy and Buildings*, 240:110886, 2021.

[9] RA Lukshmi, PV Geetha, and P Venkatesan. Rough set theory approach for attribute reduction. *Int. J. Autom. Artif. Intell*, 1(3):70–80, 2013.

[10] Vamsidhar Talasila, Kotakonda Madhubabu, Meghana Chakravarthy Mahadasyam, Naga Jyothi Atchala, and Lakshmi Sowjanya Kande. The prediction of diseases using rough set theory with recurrent neural network in big data analytics. *International Journal of Intelligent Engineering & Systems*, 13(5), 2020.

[11] Touhid Mohammad Hossain, Junzo Watada, Izzatdin A Aziz, and Maman Hermana. Machine learning in electrofacies classification and subsurface lithology interpretation: A rough set theory approach. *Applied Sciences*, 10(17):5940, 2020.

[12] Yılmaz Kaya and Murat Uyar. A hybrid decision support system based on rough set and extreme learning machine for diagnosis of hepatitis disease. *Applied Soft Computing*, 13(8):3429–3438, 2013.

[13] Hui-Ling Chen, Bo Yang, Jie Liu, and Da-You Liu. A support vector machine classifier with rough set-based feature selection for breast cancer diagnosis. *Expert systems with applications*, 38(7):9014–9022, 2011.

[14] Asma Trabelsi, Zied Elouedi, and Eric Lefevre. An ensemble classifier through rough set reducts for handling data with evidential attributes. *Information Sciences*, 635:414–429, 2023.

[15] Ping-Feng Pai, Shi-Yu Chen, Chao-Wei Huang, and Ya-Hsin Chang. Analyzing foreign exchange rates by rough set theory and directed acyclic graph support vector machines. *Expert systems with applications*, 37(8):5993–5998, 2010.

[16] Essam Debie, Kamran Shafi, Chris Lokan, and Kathryn Merrick. Reduct based ensemble of learning classifier system for real-valued classification problems. In *2013 IEEE Symposium on Computational Intelligence and Ensemble Learning (CIEL)*, pages 66–73. IEEE, 2013.

[17] Georg Peters and Simon Poon. Analyzing it business values–a dominance based rough sets approach perspective. *Expert Systems with Applications*, 38(9):11120–11128, 2011.

[18] Andrzej Janusz, Andżelika Zalewska, Łukasz Wawrowski, Piotr Biczyk, Jan Ludziejewski, Marek Sikora, and Dominik Ślezak. Brightbox—a rough set based technology for diagnosing mistakes of machine learning models. *Applied Soft Computing*, 141:110285, 2023.

[19] Alessio Ferone. Feature selection based on composition of rough sets induced by feature granulation. *International Journal of Approximate Reasoning*, 101:276–292, 2018.

[20] Jie Zhao, Dai-yang Wu, Yong-xin Zhou, Jia-ming Liang, WenHong Wei, and Yun Li. Rough set theory-based group incremental approach to feature selection. *Information Sciences*, 675:120733, 2024.

[21] Moaad Almania, Anazida Zainal, Fuad A Ghaleb, Ahmad Alnawasrah, and Mahmoud Al Qerom. Two-level feature selection for enhanced accuracy and reduced computational overhead in intrusion detection systems using rough set theory and binary particle swarm optimization. *Journal of Robotics and Control (JRC)*, 6(1):262–271, 2025.

[22] Nasser Alsabilah and Danda B Rawat. Joint rough set theory and xgboost based learning for network intrusion detection system. *IEEE Internet of Things Journal*, 2025.

[23] VK Hanuman Turaga and Srilatha Chebrolu. Efficient and fast algorithm for attribute reduction of large dimensional data using rough set theory on graphics processing unit. *Arabian Journal for Science and Engineering*, 50(2):1209–1231, 2025.

[24] Kaggle. Seeds dataset. `https://www.kaggle.com/datasets/rwzhang/seeds-dataset`.

[25] Kaggle. Online food delivery preferences (bangalore region). `https://www.kaggle.com/datasets/sujithmandala/obesity-classification-dataset`.

[26] Kaggle. Heart disease classification dataset. `https://www.kaggle.com/datasets/`

benroshan/online-food-delivery-preferencesbangalore-region/code.
[27] Kaggle. Obesity classification dataset. https://www.kaggle.com/datasets/benroshan/online-food-delivery-preferencesbangalore-region/data.
[28] UCI. Breast cancer wisconsin original dataset. https://archive.ics.uci.edu/dataset/15/breast+cancer+wisconsin+original.
[29] Alessio Ferone and Antonio Maratea. Adaptive quick reduct for feature drift detection. *Algorithms*, 14(2):58, 2021.
[30] Vladimir Nasteski. An overview of the supervised machine learning methods. *Horizons. b*, 4(51-62):56, 2017.
[31] Batta Mahesh et al. Machine learning algorithms-a review. *International Journal of Science and Research (IJSR).[Internet]*, 9(1):381–386, 2020.
[32] Yichuan Tang. Deep learning using support vector machines. *CoRR, abs/1306.0239*, 2(1), 2013.
[33] Arti Patle and Deepak Singh Chouhan. Svm kernel functions for classification. In *2013 International conference on advances in technology and engineering (ICATE)*, pages 1–9. IEEE, 2013.
[34] Leo Breiman. Random forests. *Machine learning*, 45:5–32, 2001.