Data Reduction Techniques: A Comparative Study

Ahmed Reyd AlKarawi
Department of Computer Science Faculty of CS
and Mathematics University of Kufa
ALNajaf, Iraq
ahmedr.alshabani@student.uokufa.edu.iq

Kadhim B. S. AlJanabi
Department of Computer Science Faculty of CS
and Mathematics University of Kufa
ALNajaf, Iraq
kadhim.aljanabi@uokufa.edu.iq

DOI : http://dx.doi.org/10.31642/JoKMC/2018/090201
Received Mar. 16, 2021. Accepted for publication Dec. 1, 2021

Abstract— *Data preprocessing in general and data reduction in specific represent the main steps in data mining techniques and algorithms since data in real world due to its vastness, the analysis will take a long time to complete .Almost all mining techniques including classification, clustering, association and others have high time and space complexities due to the huge amount of data and the algorithm behavior itself. That is the reason why data reduction represent an important phase in* Knowledge Discovery in Databases *(KDD) process. Many researchers introduced important solutions in this field. The study in this paper represents a comparative study for about 22 research papers in data reduction fields that covers different data reduction techniques such as dimensionality reduction, numerisoty reduction, sampling, clustering data cube aggregation and other techniques. From the conducted study, it can be concluded that the appropriate technique that can be used in data reduction is highly dependent on the data type, the dataset size, the application goal, the availability of noise and outliers and the compromise between the reduced data and the knowledge required from the analysis.*

Keywords: Data Mining, Data Preprocessing, Data Reduction, Dimensionality Reduction

I. INTRODUCTION

Data mining (DM) is a branch of computer science and statistics that focuses on finding patterns in large databases of data. For the most part, the goal of the data mining process is to pull out relevant information from a large collection and organize it in a way that may be used in the future.

Data mining supplies us with relevant information that queries and reports are not able to adequately deliver. There is no clear way to save the information that is gleaned by data mining methods in a database, while database applications can only display the data that is already stored there. As a result, data mining may best be characterized as the discovery of information in databases [1].

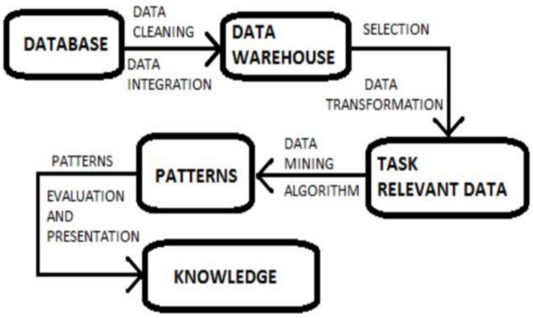

Fig. 1. Steps in KDD Process [1]

Data mining is a method based on statistical analysis of large datasets. using this method, massive volumes of previously undiscovered data may be extracted. Data mining is used by the banking and insurance sectors to identify fraud, provide consumers with credit or insurance options, and better understand their needs.

A. Why Data Mining?

As we all know, data grows exponentially from terabytes to petabytes. The biggest difficulty was lack of knowledge. We are awash with data but devoid in wisdom. Massive data warehouses have cost a lot of money, but the results have been disappointing (return on investment). They are unable to satisfy demand due to a shortage of manpower and equipment. Data mining aims to automate case classification using collected data patterns. Various algorithms have been developed to aid decision making. These algorithms extract data and reveal knowledge patterns Data mining is called KDD [2].

B. Data Mining Applications

Data Mining applications cover a wide range of applications including: Healthcare, market basket analysis, education system, manufacturing engineering, CRM, language research, Medical Science, Web Education, Credit Scoring, Intrusion Detection in the Network, Malicious Executable files, Sports data Mining, The Intelligence Agencies, Internal Revenue Service, E-commerce, Digital Library Retrieves, prediction in engineering applications and many others [3].

C. Data Mining Techniques

1) Data classification
2) Prediction data mining technique
3) Data clustering technique
4) Outlier analysis technique
5) Association rule mining (ARM) [1].

II. DATA PRE-PROCESSING

The goal of pre-processing data is to make it easier and more efficient to utilize raw data in later processing procedures by simplifying and improving its efficiency.

Before using a data mining approach, it is necessary to do data pretreatment, which is one of the most important concerns in the well-known Knowledge Discovery from Data process, as illustrated in figure 1.2. Data with inconsistencies and redundancies cannot be used to begin a data mining process since the data is likely to be poor. Data generating rates and sizes in business, industry, academia, and research are all rapidly increasing. Analyzing large volumes of data necessitates increasingly complex tools [4]. The ability to analyze data that would otherwise be impossible without data pretreatment makes the most sophisticated algorithms obsolete, since models trained on faulty data may actually impair your analysis and give you "junk" findings. Preprocessing data is thus more vital than ever. you may end up with data that is out of range or contains an inaccurate feature, such as a family income below zero or a picture from a collection of "zoo animals" that is really a tree, depending on your data collection methods and sources. There may be omissions or gaps in your data set. It's fairly uncommon to see misspelled words and other inaccuracies in text data (such as URLs and symbols) , etc.

Data that has been preprocessed and cleaned adequately will lead to more accurate results in the subsequent processing stages. Even though the term "data-driven decision making" is commonly bandied around, incorrect data may nonetheless lead to poor conclusions.[5]

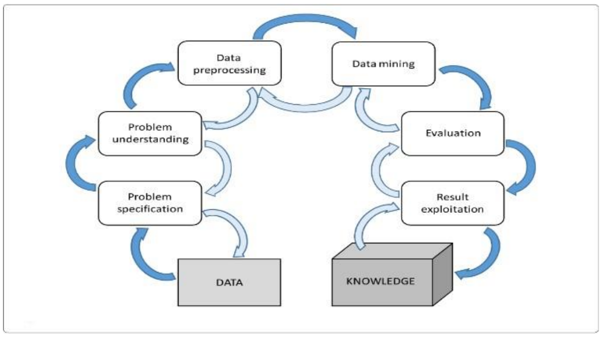

Fig. 2. KDD Process [5]

A. Strategies in Data Preprocessing

[1] Data preprocessing consists of many phases, techniques and algorithms as shown in Fig. 3

1) Data Reduction
2) Data Cleaning
3) Data Construction
4) Data Integration
5) Data Format Change [6]

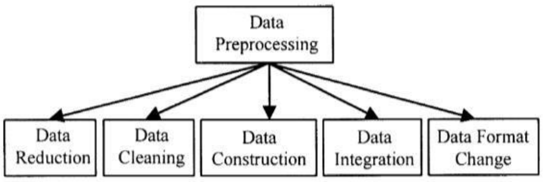

Fig. 3. Data Preprocessing Phases [6]

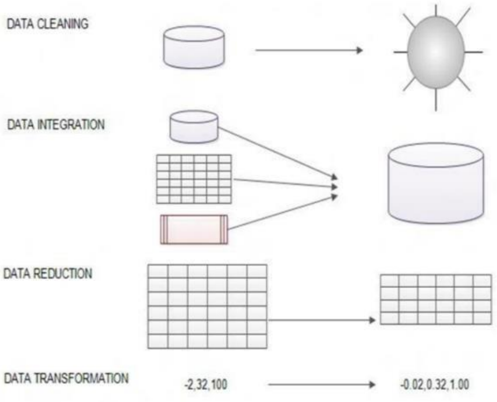

Fig . 4 . Data Preprocessing Semantics [7]

B. Why do we need to preprocess data?

1) Make our dataset more accurate. We remove any values that are wrong or aren't there because of errors or the human aspect.
2) Boost consistency. Due to inaccurate or redundant data, the findings might be inaccurate..
3) Make the database more complete. If there are any missing attributes, we may add them.Smooth the data. This way we make it easier to use and interpret.

C. Why Data Preprocessing is Important?

Preprocessing data is critical since when data isn't clean, duplicates and poor quality data are there, and the results

aren't any good either. The quality data must be used to

make quality judgments.Data quality may be measured in terms of correctness, ompleteness, consistency, timeliness, believability, and interpretability via the use of data processing techniques.[7]

The following are the most common types of data Processing and their applications:

1) Transaction Processing
2) Distributed Processing
3) Real – time Processing
4) Batch Processing
5) Multiprocessing

III. DATA REDUCTION

In data prepossessing strategies, Data Reduction (DR) is one of the approaches that is used to deal with large amounts of dimensional data. When using this approach, the primary aim is to minimize the size of the original data while maintaining the integrity of the original information. By using certain specialized approaches, reduction may be accomplished in either vertical columns of characteristics or horizontal rows of instances, as desired. [8]

Prior to the implementation of instance-based machine learning algorithms in the Big Data era, data reduction has become very important. Reducing the amount of datasets while maintaining representative data is the goal of data reduction. A trade-off between learning precision and the pace at which data is reduced is inherent in the current algorithms.In recent years, owing to the rapid development in data volume in both the industrial and scientific realms, Big Data has attracted the interest of a wide range of applications due to the huge potential value it holds. However, the processing capacity of popular machine learning programs is suffering as a result of this expansion.[7],[8]. For instance-based learning algorithms, keeping a high number of instances results in a big memory footprint and a poor execution time. Data reduction may be used as a preprocessing step in machine learning to address the issue [9]. Data reduction is the process of removing from the original datasets the instances that aren't relevant, such as noise, redundant data, or data that isn't closely connected. With decreased computational and storage costs, it is predicted to improve the performance of instance-based learning methods. [10].

When it comes to data analysis, you have to go deep to uncover even the most minute trends and patterns. It's a time-consuming rocedure since every option must be thoroughly investigated in order to unearth any helpful information. Data reduction aids in reducing the amount of data while preserving the integrity of the data at the same time. Because of this, data management and analysis are simplified, resulting in improved efficiency and lower costs. As long as you're using the same data, you may save a lot of resources by reducing the amount of data you're processing.

DA DATA REDUCTION TECHNIQUES

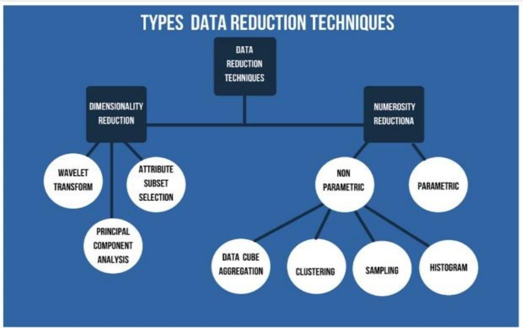

Fig .5. Types of Data Reduction Techniques [11]

A.Dimensionality Reduction

Atechnique called "Dimensional Reduction" reduces the number of dimensions that data may be seen from. As the number of features rises, the sparsity of the qualities or features in the data collection increases. These algorithms rely heavily on the sparsity of the data. Data may be more easily seen and manipulated when it has fewer dimensions. [11]

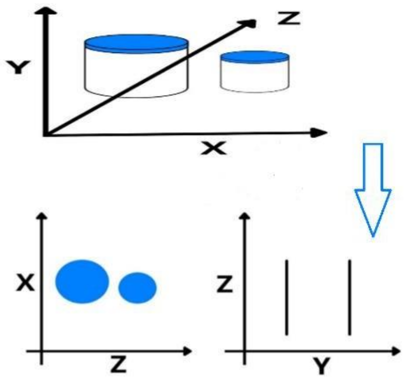

Fig .6. Dimensionality Reduction [12]

There are three types of Dimensionality reduction.

1) Wavelet Transform
2) Principal Component Analysis
3) Attribute Subset Selection

B. Numerosity Reduction

With this technique, you may reduce the amount of data you're working with by switching to a smaller data representation. Numerosity reduction is divided into two categories: parametric and non-parametric(e.g. Histograms). [11]

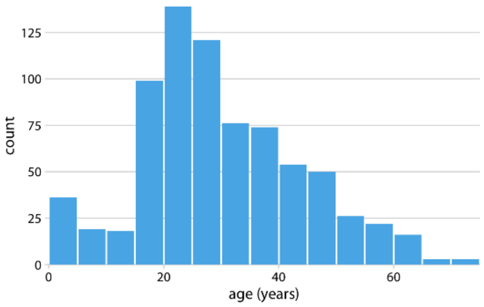

Fig .7. Histogram Technique [13]

C. Clustering as Data Reduction Technique

As a result of Clustering, the data set is replaced by a cluster representation, where the data is separated into clusters based on the cluster's similarities and dissimilarities to the rest of the data set. The closer two cluster members look to one another, the more alike they are. As the distance among two data objects grows, so does the cluster's quality

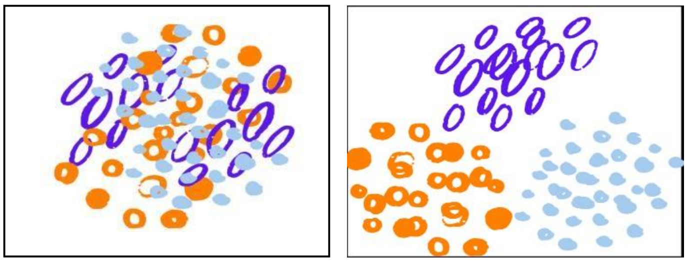

Fig . 8. Clustering Technique

D. Sampling

Sampling may reduce a huge data set data collection into a representation of the original by dividing it into smaller sample data sets. Methods for reducing the amount of sample data are classified into four broad categories.

1) Simple Random Sample Without Replacement of sizes
2) Simple Random Sample with Replacement of sizes
3) Cluster Sample
4) Stratified Sample

E. Data Cube Aggregation

In order to reflect the original data set, Data Cube Aggregation employs aggregation at several layers of a data cube to achieve data reduction. Accumulation in the form of data cubes allows for more efficient data storage, which in turn allows for smaller datasets to be assembled more quickly

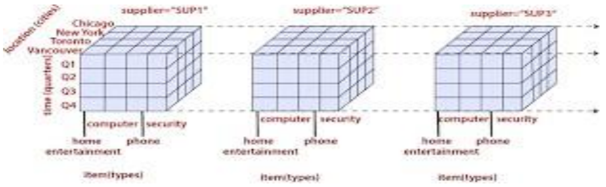

Fig . 9 . Data Cube Aggregation[14]

F. Data Compression

In order to save space, it alters, encodes, or converts the data's structure. Data compression is the procedure of reducing the size of data by eliminating redundant information and storing it in binary form. Lossless compression refers to data that can be successfully recovered from its compressed form, whereas Lossy compression refers to data that cannot be restored from its compressed form.

IV. Aims and Objectives

1) Conducting a comparative study for the most popular data reduction techniques and algorithms including: dim reduction, sampling, clustering
2) Identifying the advantages and drawbacks of the most common research papers in the field related to the criteria such as accuracy, time and space complexity, scalability and others

3) Providing the researchers with required recommendations when to use and when not to use each data reduction technique depending on data size, data type and others

V. Conducted Survey

Several papers on will be reviewed in this section.data reduction techniques are presented.

- Ibrahim M. El-Hasnony, Hazem M. El Bakry And Ahmed A. Saleh, 2015 [15] FRFS, RSAR, PCA, CFS, and gain ratio have been offered in data reduction, and healthcare has been used as a numerical UCI machine learning repository. Various data reduction approaches were examined to determine the main contribution of research and it is advantages and drawbacks to classification efficiency, as well as their efficacy. Gain ratio (GR), rough set, correlation feature selection CFS, principal components analysis (PCA), and fuzzy rough feature selection are the suggested dimension reduction techniques. It was examined in terms of classification accuracy for the C4.5, fuzzy rough nearest neighbor, MLP, NNGE, Fuzzy nearest neighbor, sequential minimum optimization (SMO), classification through clustering, NB-tree, and naive Bayes (NB) algorithms. It was also tested in terms of classification accuracy for C4.5. FRFS outperformed the other methods in minimizing medical data, according to the findings. Compared to RSAR, PCA, or gain ratio, CFS performed well

- R.K. Bania , 2014 [8] Using PCA, ICA, LDA, and Feature Selection (FS). the FS algorithms. There are three techniques based on subset assessment criteria: filter, wrapper, and embedding. The hybrid approach combines the filter and wrapper methods and is used in many other papers.Filter technique utilizes distinct features This strategy is speedier and aims to provide best outcomes, while This method's main flaw is that it overlooks the learning algorithm's interaction. Wrapper technique uses dependent assessmentcriteria, i.e. The key benefit of this strategy is that it outperforms filters. It is faster than a filter model sinceit must regularly re-run or invoke the inductionmethod.The benefit of this strategy is that it has interaction with the learning model and is less computationally costly than the other ways. This approach is less prone to overfiiting because to itslarger data involvement capacity.
- Ramona Georgescu, et al , 2010 [16] for data reduction, they used PCA, PLS, SRM, and OMP. In the first scenario, they use PCA and PLS, whereas in the second, they use SRM and Orthogonal Matching Pursuit (OMP). they assess their classification performance on publically accessible datasets using Support Vector Machines (SVM) and Proximal Support Vector Machines (PSVM) (PSVM). The experiments use the WDBC and Ionosphere datasets from the UC-Irvine Machine Learning Repository. PCA was deemed worthwhile. Based on these findings, data reduction looks to be a viable method for reducing storage or transmission resource needs while maintaining performance.
- Reham M. Alamro , And Abdou S. Youssef, 2018 [17] used (SUFSS), Correlation-based Feature Subset Selection (CFSS), Probabilistic Significance (PS)-based Feature Selection, Chi-squared Statistic (CHI), Clustering Variation (CV), Consistency within Feature Selection technique, and Simple Random Sample without. For example, all nominal values except class labels were converted to numeric, all tuples with missing or inconsistent class labels were removed, and all missing features were replaced with zeros. they assessed each classifier's accuracy, training time, and testing time before and after each reduction approach. they averaged the accuracy, training time, and testing time of each classifier over all ten datasets before and after each reduction approach, and then Filtering The classifiers that pass the speed-and-accuracy criterion are kept (and counted).
- Shailesh Singh Panwar , And . Dr. Y. P. Raiwani , 2014 [18] The NSLKDD dataset is chosen as the experiment object (CFS, Gain Ratio, Info Gain, OneR, Wrapper, Symmetrical). It is now the most extensively used dataset for network forensic data reduction assessment. To estimate the performance of the suggested approach on the NSL KDD Cup-99 dataset, we employed a 10% subset of the original NSL KDD Cup-99 data set. The NSL KDD Cup data sets included continuous, discrete, and symbolic attributes with varied resolution and ranges. Most pattern classification algorithms cannot handle such data. So pre-processing was needed! 2 Steps of pre-processing The first stage mapping numerical-valued attributes from symbolic ones, while the second involved scaling.
- Rusul Kadhim and Kadhim B.S. Al Janabi , 2018 [19] Strategy that begins with a subset of features that is completely empty and then adds important features to the subset, it is cal (sequential forward selection and sequential backward selection). It is possible to add more features to the subset without lowering the quality of the subset in sequential forward selection, but it is not possible to remove features from the subset in sequential backward selection since the removed feature cannot be picked again. A quadratic time complexity in terms of data dimensionality characterizes most heuristic tactics such as greedy sequential search, best-first search, and evolutionary algorithms, but they do not guarantee the discovery of the optimum features for this.
- D. M. Deepak Raj , And R. Mohanasundaram , 2020 , [20] working on cancer microarray datasets to get rid of the dimensionality curse, The studies used 10 high-dimensional cancer microarray datasets. Experiment data were preprocessed. The leukemia, GCM, lung, and DLBCL datasets had null values. The features with above 35% missing values were omitted. RELIEF's main idea is to iteratively alter feature weights depending on their ability to differentiate between

neighboring patterns. This is achieved by dividing the nearest neighbor patterns, one of the same class (NH|) and another of distinct classes (NM|). We used tenfold cross-validation to test classification precision (CV). A tenfold CV estimates the generalization error for stable classifiers like KNN. Using the tenfold CV, the model predicted each sample in the dataset, and the accuracy was incorporated in the final measurement. Every dataset was divided 70:30, with 70 percent utilized to develop the model and the rest predicted. The average precision was the tenth iteration's final measurement. A fivefold CV was used to obtain the optimal parameters for classifiers with tuning parameters (like the SVM) and then utilized for modeling. On has used the literature to simplify the comparison.

- Noviyanti T M Sagala , 2019 , [21] , Based on 858 records with 32 features and 4 target variables indicating the kind of medical test; Hinselmann, Two qualities (STDs) were found to be unhelpful for data mining. Data reduction may be achieved in two ways: reducing data dimensionality or reducing data distribution. Both methods were used in this study. Feature selection reduced data dimensionality while data discretization reduced data dispersion. Data discretization reduces a huge domain of numeric values to a subset of categorical values, which may enhance learning. Smaller data variations usually lead to more exact predictive models and greater prediction rates . The library's “discretization” only applied to a few properties. According to [14], age was converted into four discrete values: 21, 21-29, 30-65, and >65. These variables were then translated into four intervals using MDLP. The training and testing datasets utilized the reducedattribute set. Random Forest and Correlation-based Filter Selection may be used to reduce dimensionality. The optimal Random Forest attribute set was selected bythe number of thresholds. The top 5 were 5, 10, 15, 20, and 25.
- Saba Bashir, et al, 2019 [22] , UCI data sets were downloaded to begin the suggested technique. Preprocessing and data discretization in the form of data cleaning, data transformation, data reduction, binning, and select attributes follow verification of the datasets. This is the last stage before running the analysis. The primary strategy for feature selection is used after all of these strategies have been applied to the downloaded dataset. After that, techniques such as Decision Tree(DT), Logistic Regression(LR) ,Suport Vector Machine(SVM), Nave Bayes (NB), and Random Forest are used to the data... After using several algorithms and methodologies

- Syed Javeed Pasha , And E.Syed Mohamed , 2020 [23] The experiment uses the UCI ML thyroid dataset. Ross Quinlan from Sydney's Garavan Institute provides it. The dataset contains 3772 records, 2800 training (data) and 972 test records. The dataset contains 29 characteristics, with the final column predicting the illness. 29 characteristics, 22 category, 7 numerical “Use of thyroid or antithyroid medication”, “history of thyroid surgery”, “complaint of malaise”, and “psychological symptoms” are 11 clinical features. The dataset also includes six test results: TSH, T3, T4, T4U, FTI, and TBG. The thyroid dataset is severely asymmetrical, with four classifications and numerous missing values. The dataset has 3772 occurrences, 3481 (92.3%) are negative (normal), 194 (5.1%) are compensated-hypothyroid (hyperthyroid), 95 and then preprocess data. The column mean replaces missing values in the dataset. Unbalanced data allocations in each class differ greatly. The dataset must be balanced. For example, the suggested model employs an ensemble method (RF) and a gain ratio approach to determine the most accurate characteristics that contribute to illness prediction

- Mary Monir Saeid , Zaki B. Nossair , and Mohamed Ali Saleh , 2020 [24] , To categorize microarray cancer data, DWT and their enhanced GA are offered as new data reduction methods and feature selection techniques. The proposed classification method's main purpose is to distinguish among normal and pathological microarray data for different cancer conditions. The number of samples in the gene expression dataset is denoted by the number of rows (N). The number of columns (M) denotes each sample's gene count. The fundamentalissue with microarray datasets is the enormous number of characteristics (M) compared to the limited number of samples (N). The enhanced GA then selects the most significant features, reducing unnecessary or noisy features to improve classifier performance. Compared to seven data reduction approaches, it outperforms them in classification accuracy, size reduction rate, and runtime. The experimental findings show that the suggested technique has a lower computing cost and a greater classification accuracy than the other algorithms when the reduction size is the same

- Xiaoyan Sun, Lian Liu,And Shaofeng Yang, 2017 [25] , used numrical data from UCI Repository and then applay wrapper DR ,filter DR by K-means KNN , its advantages to preserve the distribution of instances, speed up computation, and maintain a greater classification accuracy They have high processing costs and trade off size reduction rate and learning accuracy.The algorthems have showed excellent outcomes following data reduction techniques.

- Chih-Wen Chen, et al , 2020 [26] , There are three techniques used to choose features: filter, wrapper, and embedding. Two methods are used to pick ensemble features. First, two various kinds of feature selection techniques are merged, and then three dissimilar kinds of feature selection methods are combined.To decrease a training set of M-dimensional features to a smaller set, each form of feature selection is applied. It is a given that various ways of reducing feature sets will provide different results. For the aggregation of distinct reduced feature sets containing different chosen features, union and intersection techniques may be used

- Elgin Christo et al, 2020 [27], An evaluation of the proposed framework was performed on seven datasets comprising Wisconsin Diagnostic Breast Cancer, Hepatitis, Pima Indian Diabetes, CHD and SHD, the Vertebral Column and Hepatocellular Carcinoma, and the subsystems of preprocessing, FS and instance, and classification.

- Xiaohui Lin, et al , 2019 [28] , refers an Interaction Gain - Recursive Feature Elimination (IG-RFE) approach that combines feature and class label relevance with feature interaction. Symmetrical uncertainty is used to assess feature-class label relevance. On each feature's average normalized interaction gain is calculated. , as well as the class label. Less essential characteristics are repeatedly deleted using symmetrical uncertainty and normalized interaction gain. On eleven available datasets, IGRFE was compared to seven efficient feature selection methods: MIFS, mRMR, CMIM, ReliefF, FCBF, PGVNS, and SVM-RFE. The findings demonstratedthat IG-RFE had higher accuracy, sensitivity, specificity, and stability. Thus, combining feature individual discriminative capacity and feature interaction might improve feature important evaluation in biological data analysis

- Weiru Chen, et al , 2018 [29] , Data clustering in Big Data is used to categorize similar data sets and subsequently uncover their patterns. The schema underpinning this definition is: Find K groups from a sample of N items based on their commonalities. Those belonging to the same group should be similar, while objects belonging to separate groups should be distinct. Data clustering may be based on object properties. However, data clustering is a subjective process impacted by factors such as “the beholder's eye” and the need for prior information. Noise might impact the outcome.

- Praveen Kumar Reddy, et al, 2020 [30] On the CTG dataset, this study examines the impact of feature engineering and dimensionality reduction on ML algorithm performance. First, CTG dataset is subjected to feature engineering, normalization, and numeric data conversion.It uses min-max standardscaler normalization to standardize the input dataset before testing ML algorithms like Decision Tree (DT), Naive Bayes (NB), Random Forest (RF) and Suport Vector Machine (SVM). Precision, Recall, F1-Score, Accuracy, Sensitivity, and Specicity are used to assess the classifiers' performance. On a normalized dataset, LDA extracts the most important characteristics. The resulting dataset is used to test ML techniques. In addition to the measures listed above, The normalized dataset is then PCA'd. The resulting dimensionally summary dataset is then used to test the ML algorithms
- Jinya Su, et al, 2017, [31] investigate traditional dimension reduction techniques for HIS classification. Support Vector Machines (SVM) are used to classify data using mutual information, low redundancy, and maximum relevance (SVM). The methods are tested on a genuine HSI dataset. PCA has the best performance in minimizing the number of features or spectral bands. The suggested technique outperforms the standard SVM on a short training dataset, making it appropriate for real-time applications or when training data is scarce. Itcan also perform as well as standard SVM on huge datasets but takes considerably less time to compute.
- Jovani Taveira De Souza , et al 2019 [32] outlined a four-step plan: Pre-processing, data mining, database selection, and post-processing are all steps in the data mining process. The first step was to pick the databases to be studied. After that, a classification procedure based on AS and PCA algorithms classified all database characteristics.

- Xiaowei Zhao, et al , 2019 [33] In this work has been proposed that a Principal Component Analysis (PCA) and Linear discriminant analysis (LDA) combined model for dimensionality reduction, known as joint principal component and discriminant analysis (JPCDA), which avoids the small sample size problem of LDA while also extracting the most discriminant information possible for the classification task, is evaluated using six benchmark data sets.

- Yunsheng Song, Jiye Liang , Jing Lu, and Xingwang Zhao, 2017 [34] proposed reducing the training set for kNN regression, suggest a method (DISKR). After removing outliers that have an influence on the regressor's performance, we next arrange the remaining instances based on the differences in output among them and their closest neighbors. In the end, only the cases that have had a negligible impact on the training error are kept.

- Priyanka Saha, Srabani Patikar, and Sarmistha Neogy 2020 [35] the results of multiple machine learning algorithms on the feature subset of a correlation-sequential forward selection based hybrid feature selection technique. The filter technique uses correlation to choose features, whereas the wrapper approach uses sequential forward selection. Making healthcare forecasts more correctly and quickly is vital since it directly affects human health. Filter Wrapper based hybrid feature selection It uses correlation-based and sequential-forward feature selection methods. These two ideas are used to offer a superior feature subset for machine learnin g models.

TABLE.1. COMPARISONS BETWEEN DIFFERENT DATA REDUCTION TECHNIQUES

NO	***authors name***	***Year of publication***	***Technique used***	***Data used***	***Data type***	***Algorithms used***	***advantages***	***disadvantages***	***evaluation***
1	Ibrahim M. El-Hasnony Hazem M. El Bakry Ahmed A. Saleh	2015	FRFS, RSAR, PCA, CFS, and gain ratio	healthcare	numrical	Fuzzy rough nearest neighbor Fuzzy nearest neighbor NNGE SMO naïve Bayes	construction of simple and more intelligible models beneficial for analyzing focused data and boosting data mining efficacy. data..	Many sources, such as mobile apps, capturing devices, and sensors, are used to obtain the medical data, and these sources might have a variety of issues, such as missing values duplicate features and noise .	To reduce medical data more effectively than any other method, researchers found that FRFS performed better than the other methods tested. Compared to RSAR, PCA, or gain ratio, CFS performed well.
2	R.K. Bania	*2014*	1. (*PCA*) 2.(*ICA*) 3. (*LDA*) of Features selection (FS)	Dataset	Numerical & categorica	1.Filter 2. wrapper 3. embedded (FS) Isomap locally linear embedding (*LLE*)	avoiding over-fitting and reducing computational complexity As a second benefit, it is more resilient in the face of noise, redundant features, or other distractions, and gives a higher level of accuracy. Seeing all of these benefits convinced me to investigate the FS approach for DR further.	The generated characteristics do not retain actual meaning and need sophisticated calculations. A subset evaluation function selects a reduced selection of initial features in FS.	Feature selection (FS) is the process of removing attributes from data without impacting the core content.
3	1.Ramona Georgescu 2. Christian R. Berger 3. PeterWillett 4.Mohammad Azam 5. Sudipto Ghosh	2007	PCA, PLS SRM OMP	Breast Cancer dataset from Wisconsin Diagnostic		*Support Vector Machine (*SVM*)* PSVM	provide a significant reduction in signal processing load with acceptable loss in performance.	even with dimensionality reduction, the dataset was difficult to categorize.	PCA was deemed worthwhile. Based on these findings, data compression looks to be a valuable method for reducing storage and transmission resource needs while maintaining performance.

NO	*authors name*	*Year of publication*	*Technique used*	*Data used*	*Data type*	*Algorithms used*	*advantages*	*disadvantages*	*evaluation*
4	1. Reham M. Alamro 2. Abdou S. Youssef	2018	Feature selection (FS) Instance selection (IS)	Inter-university Consortium for Political and Social Research (ICPSR)	quantitative, and the class labels are nominal	SVM J48C NB *Principal Component Analysis (PCA)-* *Information Gain (Entropy)- (IGFS)* *Gain Ratio- (GRFS):* *Symmetrical Uncertainty- (SUFS* *Chi-squared Statistic (CHI)-*	Combining Information-Gain Feature Selection (IGFS) with Simple Random Sample without Replacement (RSWR), we accomplish our target	Any method that led to a decrease in accuracy was dropped from further consideration.	The best feature selection is IGFS, the best instance selection is SRSWOR, and the best training and testing time reduction is integrated reduction. In terms of classifiers, SVM and J48C had the fastest training, while NB and SVM had the fastest testing..
5	1. Shailesh Singh Panwar 2. Dr. Y. P. Raiwani	2014	Cfs Gain Ratio Info Gain OneR Wrapper Symmetrical Uncert	network forensic domain	continuous, discrete, symbolic	J48 Naïve Bayes	Reduction edundancy reduction of complexity	There is no way to repair the harm done to the original data once it has been altered.	Of all the data reduction strategies, the OneR attribute assessment comes out on top.
6	Dr. Kadhim B.S. Al Janabi	2018	CFS (IG) GR Symmetrical Uncertainty Chi-S One-R Relief-F CSE Classifier Subset Evaluator Wrapper Subset Evaluator	Najaf Education Directorate in Iraq	Text	Decision Trees (DT) k-Nearest Neighbor(KNN) Support Vector Machines(SVM)	the feature selection method makes it easier to see and interpret data, which minimizes mining time and storage requirements	Although the wrapper makes them more costly, they do not take into account how features are related to one another.	The wrapper uses a predefined classification algorithm to evaluate the subsets of selected features. According to the way in which the features are evaluated, Feature selection methods are categorized into single and subset evaluation
7	D. M. Deepak Raj R. Mohanasundaram	2020	Relief techniqu	cancer microarray datasets	Microarray	1. support vector machine(SVM) 2. k-nearest neighbor (KNN) 3. Naive Bayes (SVM)	high classification accuracy, outstanding noise resistance, and goodstability across a variety of classification methods	Degradation in the performance of Multi SURF's capacity to identify important features from datasets due to a high amount of duplicate features	handle outliers and noise better, and performed is superior in terms of classification error on most test datasets..

NO	***authors name***	***Year of publication***	***Technique used***	***Data used***	***Data type***	***Algorithms used***	***advantages***	***disadvantages***	***evaluation***
8	Noviyanti T M Sagala	2019	Correlation-based Filter (CFS Random Forest	Department of Information System, Krida Wacana Christian University – Indonesia	numeric	support vector machine (SVM) Naïve Bayes (NB) k-nearest neighbor (KNN)	By reducing the range of possible numeric values to just a few discrete categorical ones, data discretization may aid in learning.	The large of number attrbuts on medical tests, caused performance proplem	The less significant qualities included in the classification process, the better. The result defies SVM. A cervical cancer prediction model is known to be built using NB-CFS on biopsy/cytology or NB-RF on Hinselmann/Schiller tests. It may also generate the highest risk variables for cervical cancer while reducing categorization time.
9	1.Saba Bashir 2.Zain ikander Khan 3. Farhan Hassan Khan 4.Aitzaz Anjum 5.Khurram Bashir	2019	MRMR/FS	Medical data FROM UCI Dataset	numeric or alphabetic digital	1.Decision Tree 2.Logistic Regression 3.Random Forest 4. Naïve Bayes 5. Logistic Regression (SVM)	In both Logistic Regression (LR) and Nave Bayes (NB), the suggested study has enhanced accuracy significantly above rule mining.	Resolving discrepancies in the medical records of a patient's illness diagnosis includes removing duplicate records and correcting missing data.	Recommends Logistic Regression as the best feature selection method for heart disease prediction compared to other techniques.
10	Syed Javeed Pasha E.Syed Mohamed	2020	Ensemble Gain Ratio Feature Selection (EGFS)	Actual Dataset From UCI repository of ML	categorical and numerical.	1.K-Nearest-Neighbor (KNN) 2. Logistic Regression (LR) 3. Naïve Bayes (NB)	It helps in the diagnosis by eliminating needless testing, saving time and money, and reducing the financial burden on patients.	The Synthetic Minority Over-sampling Technique employs oversampling in circumstances when the dataset is uneven (SMOTE)	In the absence of EGFS, KNN's accuracy was 91.82 percent, but it was 96.45 percent with EGFS, and its accuracy for LR was 98.01%, while its accuracy for NB was 90.99 percent and its accuracy for LR was 98.04% with EGFS.

NO	*authors name*	*Year of publication*	*Technique used*	*Data used*	*Data type*	*Algorithms used*	*advantages*	*disadvantages*	*evaluation*
11	Mary Monir Saeid Zaki B. Nossair Mohamed Ali Saleh	2020	discrete wavelet transform (DWT) modified genetic algorithm (GA)	Medical data	microarray datasets	1. K-Nearest-Neighbor (KNN) 2. support vector machine (SVM)	DWT is used to minimize the amount of characteristics in microarray data and remove features that are not relevant.	As a general rule, each microarray dataset has a varied number of samples, as well as distinct genes.	show that the proposed method beats the other previous cancer classification algorithms in the majority of instances studied
12	1. Xiaoyan Sun 2. Lian Liu 3. Shaofeng Yang	2017	wrapper DR filter DR	UCI Repository	numrical	K-means KNN	to preserve the distribution of instances, speed up computation, and maintain a greater classification accuracy	They have high processing costs and trade off size reduction rate and learning accuracy.	The algorthems have showed excellent outcomes following data reduction techniques.
13	1. Chih-Wen Chen 2. Yi-Hong Tsai 3. Fang-Rong Chang 4. Wei-Chao Lin	2020	1 Filter techniques. 2. Wrapper techniques 3. Embedded techniques With Feature selection (FS)	medical datasets from the UCI Machine Learning Repository	categorical, numerical, and mixed data types	1. PCA 2. genetic algorithm (GA) 3. C4.5 DT 4. SVM	The union approach of filter (PCA) and wrapper (GA) methods provides superior classification accuracy and feature reduction rate...	amount of feature dimensions, number of data samples, class unbalanced data may all effect the outcome.	The findings demonstrate that using the union and multi-intersection techniques to choose ensemble features allows the SVM classifier to perform better than single feature selection strategies.
14	1. V. R. Elgin Christo 2. H. Khanna Nehemiah 3. J. Brighty 4. Arputharaj Kannan	2020	1.Feature Selection 2.Instance Selection	Datasets Diagnostic 1.Breast cancer 2. Pima Indian Diabetes (PID) 3. Hepatitis 4. CHD 5. SHD 6. Ver tebral Column 7. Hepatocellular Carcinoma (HCC) From State of Wisconsin	Numerical Real	1. co-operative co-evolution 2. the random forest classifier	Feature selection (FS) and instance selection (IS) are seen as separate subproblems in the co-operative co-evolution method.	no classifier interaction, hence,it may result in low performance	WDBC, Hepatitis, PID, CHD, SHD, vertebral, and HCC datasets from UCI repository were tested and found to be 97.1 percent accurate, 82.3 percent, 93.4 percent, 96.8%, 91.4 percent, and 72 percent

NO	***authors name***	***Year of publication***	***Technique used***	***Data used***	***Data type***	***Algorithms used***	***advantages***	***disadvantages***	***evaluation***
15	1. Xiaohui Lin 2. Chao Li 3. Weijie Ren 4. Xiao Luo 5. Yanpeng Qi	2019	1. IG-RFE based on symmetrical uncertainty 2. interaction gain	Dataset 1. WDBC 2. Breast cancer 3. Leukemia 4. Lymphoma	numrical	1. IG-RFE 2. MIFS 3. Mrmr 4. CMIM 5. ReliefF 6. FCBF 7. PGVNS 8. SVM-RFE	integrating feature individual discriminative ability and the interaction among features could better evaluate feature importance in biological data analysis.	Interaction between features in data analysis may cause the loss of important information and alter the findings.	IG-RFE was shown to be more accurate than MIFS, mRMR, CMIM, ReliefF, FCBF, PGVNS, and SVM-RFE in most circumstances when it comes to measuring features.
16	1. Weiru Chen 2. Jared Oliverio 3. Jin Ho Kim 4. Jiayue Shen	2018	data clustering	Dataset	Neumrical And catigrical	1. Hierarchical Clustering 2. Centroid based Clustering 3. Distribution-based Clustering 4. Density-based Clustering	Hierarchical Clustering produces clear and well-sorted results. Centroid Expectation Maximization works very well in practice DENSITY tolerant to outliers, no need to define cluster size	Hierarchical Clustering has O(2 n) complexity, making it too sluggish for huge data sets. Slow convergence of centroid The “Curse of Dimensionality” prevents correct selection of DENSITY.”	*Clustering techniques for massive data mining. Data clustering may improve mining productivity and accuracy.*
17	1. PRAVEEN KUMAR REDDY 2. KURUVA LAKSHMANNA 3. RAJESH KALURI 4. GAUTAM SRIVASTAVA	2020	1. (LDA) 2. (PCA)	(CTG) dataset	University of California and Irvine Machine Learning Repository	1. Decision Tree (DT) 2. Support Vector Machine (SVM) 3. Naive Bayes (NB)	In contrast, classifiers using PCA outperform those using LDA and those without dimensionality reduction on the IDS dataset. .	Because of the smaller dataset, both PCA and LDA had a detrimental impact on the findings.	Classifiers perform better using PCA than LDA. Also, without dimensionality reduction, Decision Tree and Random Forest classifiers outperform PCA and LDA classifiers.
18	1.Jinya Su 2.Dewei Yi 3.Cunjia Liu 4.Lei Guo 5.Wen-Hua Chen	2017	1.feature selection 2.extraction techniques	Hyperspectral images HSI dataset.		Support Vector Machine (SVM	Less computational time (especially in testing) is accomplished, allowing for real-time HSI classification .	Classification of hyperspectral images (HSI) with little training data is examined.	The suggested technique outperforms the standard SVM on tiny training datasets, making it suited for real-time applications or situations when training data is scarce. It can also perform as well as standard SVM on huge datasets but takes considerably less time to compute.

NO	*authors name*	*Year of publication*	*Technique used*	*Data used*	*Data type*	*Algorithms used*	*advantages*	*disadvantages*	*evaluation*
19	1. JOVANI TAVEIRA DE SOUZA 2. ANTONIO CARLOS DE FRANCISCO 3.DAYANA CARLA DE MACEDO	2019	1.attribute selection AS 2.principal component analysis PCA	Healthcare (microarray data)	biomedical data repository	1.NB 2.J48 3.3-NN 4.5-NN 5. SVM 6. 1-NN	(_flter and wrapper), the classification accuracy was significantly increased compared to the three databases studied. Using this strategy also reduced the amount of attributes. The WA had the greatest success rate, while requiring a lot of computation time. .	data processing challenges caused by the large number of genes and the lack of sample-specific information.	In certain circumstances, the PCA approach had a hit rate of over 80%, making it a realistic reduction method. The top algorithms were NB, J48, 3-NN, and 5-NN. The AS technique
20	1.Xiaowei Zhao 2.Jun Guo, Feiping Nie 3.Ling Chen 4. Zhihui Li 5. Huaxiang Zhang	2019	Joint Principal Component and Discriminant Analysis (JPCDA)	images		K-Nerst Niehgiboor KNN	The short sample size issue may be avoided using the LDA technique. .	Due to the curse of dimensionality, high-dimensional images	LDA's small sample issue is solved by this approach, which uses PCA to extract the most important information from LDA's data set.
21	1. Yunsheng Song 2. Jiye Liang 3. Jing Lu 4. Xingwang Zhao	2017	Regression	Numerical ,float	KEEL Repository	k-Nearest Neighbor algorithm(kNN)	it's a way to save space and a strategy for coping with enormous amounts of data	Slower performance and higher memory demands might be the outcome of huge data sets.	Obtain a same level of prediction accuracy, but with a lower instance storage ratio.

22	1.Priyanka Saha 2.Srabani Patikar 3.Sarmistha Neogy	2020	1.Filter 2.Wrapper	Healthcare dataset	UCI machine learning repository	1. KNN 2. Decision Tree 3.Random Forest	Feature Selection improves accuracy and eliminates feature that are unnecessary.	analysis and prediction of a massive amount of data	KNN and Decision tree improve accuracy. Discarding strongly linked information decreases model ambiguity. Selected characteristics aid in categorization and prediction..

VI. CONCLUSIONS

In this study, different data reduction techniques were surveyed and the related research papers were studied. The different data reduction techniques can be classified as follows:

Dimensionality reduction, numerosity reduction, sampling, data cube aggregation, data clustering, data compression and others. From the conducted study, it can be concluded that the appropriate technique that can be used in data reduction is highly dependent on the data type, the dataset size, the application goal, the existence of noise and outliers and the compromise between the reduced data and the knowledge required from the analysis. Dimensionality reduction can be highly used when the number of features(attributes) are large as in text mining for example and there exists big variations in the effect of each attribute on the overall analysis. Reducing the number of distinct values in each attribute is effective when replacing the attribute content with nominal or categorical data. This will be effective when using different mining techniques such as Decision trees, clustering and other techniques. Clustering on the other hand is efficient when the data set attributes are highly dependent where they can be turned into clusters. The dataset in this technique is converted into number of clusters with their centroids and the objects related to each cluster. On the other hand, data cube aggregation is an efficient data reduction technique when summarized and highly summarized data are the heart of the analysis process instead of the detailed data .

REFERENCES

[1] Agarwal, Shivam. "Data mining: Data mining concepts and techniques." 2013 international conference on machine intelligence and research advancement. IEEE, 2013.

[2] Padhy, Neelamadhab, Dr Mishra, and Rasmita Panigrahi. "The survey of data mining applications and feature scope." arXiv preprint arXiv:1211.5723 (2012).

[3] Hartama, Dedy, Agus Perdana Windarto, and Anjar Wanto. "The Application of Data Mining in Determining Patterns of Interest of High School Graduates." Journal of Physics: Conference Series. Vol. 1339. No. 1. IOP Publishing, 2019.

[4] Gürbüz, Feyza, Lale Özbakir, and Hüseyin Yapici. "Data mining and preprocessing application on component reports of an airline company in Turkey." Expert Systems with Applications 38.6 (2011): 6618-6626.

[5] García, Salvador, et al. "Big data preprocessing: methods and prospects." Big Data Analytics 1.1 (2016): 1-22.

[6] Cano, José Ramón, Francisco Herrera, and Manuel Lozano. "Using evolutionary algorithms as instance selection for data reduction in KDD: An experimental study." IEEE transactions on evolutionary omputation 7.6 (2003): 561-575.

[7] Singhal, Swasti, and Monika Jena. "A study on WEKAtool for data preprocessing, classification and clustering." International Journal of Innovative technology and exploring engineering (IJItee) 2.6 (2013): 250-253

[8] Bania, R. K. "Survey on feature selection for data reduction." International Journal of Computer Applications 94.18 (2014).

[9] Benjelloun, Fatima-Zahra, Ayoub Ait Lahcen, and Samir Belfkih. "An overview of big data opportunities, applications and tools." 2015 Intelligent Systems and Computer Vision (ISCV) (2015): 1-6.

[10] Chen, CL Philip, and Chun-Yang Zhang. "Data-intensive applications, challenges, techniques and technologies: A survey on Big Data." Information sciences 275 (2014): 314-347.

[11] Bokaba, Tebogo, Wesley Doorsamy, and Babu Sena Paul. "Comparative study of machine learning classifiers for modelling road traffic accidents." Applied Sciences 12.2 (2022): 828.

[12] https://www.linkedin.com/pulse/what-dimensionality-reduction-algorithm-ml-how-we-bhattacharjee

[13] https://clauswilke.com/dataviz/histograms-density-plots.html.

[14] https://www.javatpoint.com/data-warehouse-what-is-data-cube.

[15] El-Hasnony, Ibrahim M., Hazem M. El Bakry, and Ahmed A. Saleh. "Comparative study among data reduction techniques over classification accuracy." International Journal of Computer Applications 122.2 (2015).

[16] Georgescu, Ramona, et al. "Comparison of data reduction techniques based on the performance of SVM-type classifiers." 2010 IEEE Aerospace Conference. IEEE, 2010.

[17] Alamro, Reham, and Abdou Youssef. "Impact of data reduction techniques on classification." 2018 International Conference on Computational Science and Computational Intelligence (CSCI). IEEE, 2018

[18] Panwar, Shailesh Singh, and Y. P. Raiwani. "Data reduction techniques to analyze NSL-KDD Dataset." Int. J. Comput. Eng. Technol 5.10 (2014): 21-31.

[19] Al Janabi, K. B., and Rusul Kadhim. "Data reduction techniques: a comparative study for attribute selection methods." International Journal of Advanced Computer Science and Technology 8.1 (2018): 1-13.

[20] Raj, D.M., Deepak, and R. Mohanasundaram. "An efficient filter-based feature selection model to identify significant features from high-dimensional microarray data." Arabian Journal for Science and Engineering 45.4 (2020): 2619-2630.

[21] Sagala, Noviyanti TM. "A Comparative Study of Data Mining Methods to Diagnose Cervical Cancer." Journal of Physics: Conference Series. Vol. 1255. No. 1. IOP Publishing, 2019.

[22] Bashir, Saba, et al. "Improving heart disease prediction using feature selection approaches." 2019 16th international bhurban conference on applied sciences and technology (IBCAST). IEEE, 2019.

[23] Pasha, Syed Javeed, and E. Syed Mohamed. "Ensemble gain ratio feature selection (EGFS) model with machine learning and data mining algorithms for disease risk prediction." 2020 International Conference on Inventive Computation Technologies (ICICT). IEEE, 2020.

[24] Saeid, Mary Monir, Zaki B. Nossair, and Mohamed Ali Saleh. "A microarray cancer classification technique based on discrete wavelet transform for data reduction and genetic algorithm for feature selection." 2020 4th International Conference on Trends in Electronics and Informatics (ICOEI)(48184). IEEE, 2020.

[25] Sun, Xiaoyan, et al. "Fast data reduction with granulation-based instances importance labeling." IEEE Access 7 (2018): 33587-33597.

[26] Chen, Chih-Wen, et al. "Ensemble feature selection in medical datasets: Combining filter, wrapper, and embedded feature selection results." Expert Systems 37.5 (2020): e12553

[27] Christo, VR Elgin, et al. "Feature selection and instance selection from clinical datasets using co-operative co-evolution and classification using random forest." IETE Journal of Research (2020): 1-14.

[28] Lin, Xiaohui, et al. "A new feature selection method based on symmetrical uncertainty and interaction gain." Computational biology and chemistry 83 (2019): 107149.

[29] Chen, Weiru, et al. "The modeling and simulation of data clustering algorithms in data mining with big data." Journal of Industrial Integration and Management 4.01 (2019): 1850017.

[30] Reddy, G. Thippa, et al. "Analysis of dimensionality reduction techniques on big data." IEEE Access 8 (2020): 54776-54788.

[31] Su, Jinya, et al. "Dimension reduction aided hyperspectral image classification with a small-sized training dataset: experimental omparisons." Sensors 17.12 (2017): 2726.

[32] De Souza, Jovani Taveira, Antonio Carlos De Francisco, and Dayana Carla De Macedo. "Dimensionality reduction in gene expression data sets." IEEE Access 7 (2019): 61136-61144.

[33] Zhao, Xiaowei, et al. "Joint principal component and discriminant analysis for dimensionality reduction." IEEE transactions on neural networks and learning systems 31.2 (2019): 433-444.

[34] Song, Yunsheng, et al. "An efficient instance selection algorithm for k nearest neighbor regression." Neurocomputing 251 (2017): 26-34.

[35] Saha, Priyanka, Srabani Patikar, and Sarmistha Neogy."A Correlation-Sequential Forward Selection Based Feature Selection Method for Healthcare Data Analysis." 2020 IEEE International Conference on Computing, Power and Communication Technologies (GUCON). IEEE, 2020.