International Journal of Engineering Trends and Technology
ISSN: 2231–5381 / https://doi.org/10.14445/22315381/IJETT-V73I7P132

Volume 73 Issue 7, 423-442, July 2025

Original Article

A Quality Improvement Framework using Conjoint Analysis with Minimum Redundancy Maximum Relevance for Big Data

Sindhu S[1], Veni S[2]

[1,2]*Department of Computer Science, Karpagam Academy of Higher Education, Coimbatore, Tamil Nadu, India.*

[1]*Corresponding Author: sssindhu784@gmail.com*

Received: 12 March 2025 Revised: 25 June 2025 Accepted: 01 July 2025 Published: 30 July 2025

Abstract - *A substantial volume of data in the form of information is purposefully changing in this era of digital information. The volume of digital data is increasing rapidly every second due to the usage of the Internet through various gadgets for all our everyday activities. However, big data plays a substantial role in information retrieval, which helps in predicting future trends. Due to the characteristics of heterogeneity and huge size, processing and analysing the big data to ascertain useful insights is becoming a challenging task. The selection of quality criteria for processing always determines the quality of the outcomes. Using conventional data mining-based preprocessing techniques alone may not provide an effective result for big data, as it faces decisive challenges. Consequently, a suitable feature selection model is required to enhance the quality. This paper presents a framework for selecting an important feature subset that represents the entire dataset with increased quality. The model utilizes conjoint analysis with a minimum redundancy maximum relevance algorithm for selecting significant attributes and a q-gram-based filtering approach for removing redundant and irrelevant instances. According to the analysis, the suggested model improves data quality and yields superior outcomes using fewer variables and instances. Compared to other big data models already in use, the model uses the Spark framework to produce better outcomes, holding a maximum speed-up rate of 89.50 and a maximum increased accuracy rate of 34.72%.*

Keywords - *Big data, Conjoint analysis, Data quality, Dimensionality reduction, Feature selection.*

1. Introduction

Advancements in data mining have reached a significant technical milestone that facilitates data-driven decision-making across a range of industries. In a number of applications, including consumer preferences [1], healthcare disease diagnostics [2], education [3], social media data [4], biometrics [5], e-commerce [6], and the financial industry [7], as well as forecasting future trends that are valuable and essential. Data are growing exponentially every second, owing to the rapid increase in the use of digital data through the Internet. This has led to the emergence of the big data age [8]. Unlike traditional techniques for information extraction, large data sets present a number of difficulties for knowledge extraction; therefore, standard data mining techniques may not be appropriate. The three main characteristics of big data are its tremendous velocity, large volume, and wide variety [9, 10]. Machine Learning (ML) techniques are used to effectively extract information from such large amounts of data, as conventional data mining methods may not be suitable for these data types [11, 12]. Recently, big data has become a game changer in almost every industry, particularly when it comes to making important decisions [13]. Data analysts, statisticians, and experts often rely on various tools to organize and process historical data to forecast and make accurate decisions across various applications, data analysts, statisticians, and experts often rely on a variety of tools [14]. However, appropriately transforming, integrating, operating, and handling data when dealing with large volumes is undoubtedly challenging. Big data must be managed carefully because the information that can be gleaned from it is extremely useful for a number of applications. The accuracy of the generated information is significantly influenced by the quality of the input data that must be processed [15].

A primary challenge in big-data analysis is data acquisition. When processing substantial volumes of data, information is sourced from multiple origins, resulting in collected data manifesting in various formats depending on the underlying data sources. Dealing with such massive amounts of data presents a secondary challenge. Additionally, it is important to handle missing values effectively because the obtained data may not be comprehensive. The gathered data must be effectively integrated from many sources in an orderly manner. Redundancy is the primary issue in data integration.

The Internet may cause data to be stored in several locations, and the process of retrieving it may result in duplicates, thereby enlarging datasets [16]. In addition to these difficulties, the quality of data has a significant impact on the outcomes of knowledge extraction. Owing to the increasing Number of features, integrated big data tends to have higher dimensionality. In contrast, it may also include redundant and unnecessary information, because not all qualities are relevant to specific research [17]. Similarly, there will be a large number of instances, some of which may be irrelevant because they are outliers or duplicate occurrences [18].

Therefore, a crucial part of processing large amounts of data involves selecting significant features and appropriate instances. This generates understandable data that dramatically improves the model's performance in a short amount of time. Nevertheless, a number of difficulties arise when dealing with feature selection in large datasets, particularly regarding the nature of the features themselves and data from multiple sources. In contrast to regular datasets, big data often exhibits an inherent structure among its attributes. However, the existing feature selection techniques are limited to generic data types. Unlike conventional attribute values, many qualities are connected and associated. Another difficulty arises from multi-view sources, in which many instances represent different features that serve as pivot points within a high-dimensional space [19].

Recent research has focused on either feature selection or instance filtering alone, but there is a lack of comprehensive studies that address both feature and instance reduction in a scalable and effective way. Moreover, the inherent attribute structures and multisource complexity of big data exceed the capabilities of most existing techniques. Consequently, there is a critical need to develop a detailed model for improving data quality that uses scalable frameworks to simultaneously optimize both feature and instance selection for high-dimensional large data. Given the numerous challenges associated with enhancing the quality of large datasets, a proper quality improvement model using appropriate algorithms is urgently needed [20].

As a result, this study proposes a strategy for enhancing data quality by choosing relevant instances and an important feature subset that defines the complete dataset. The model employs a q-gram-based filtering technique to select relevant instances and conjoint analysis using the minimal Redundancy Maximum Relevance (mRMR) method to determine the important feature subset. The model selects essential features and instances both horizontally and vertically, eliminating extraneous attributes and instances that reduce the learning algorithm's performance. Numerous evaluations were conducted to demonstrate the effectiveness of the model in terms of accuracy and execution time. Using a distributed big data architecture, the proposed framework uniquely combines attribute selection and instance filtering compared with existing methods. The innovative part of this study is its two-stage procedure, which addresses data quality concerns simultaneously in a horizontal (feature-wise) and vertical (instance-wise) manner. The analysis of the findings shows that the proposed model enhances the data quality and produces better results with a smaller set of variables and instances. When applied to data using the Spark framework, the model achieved its full potential in terms of speed and accuracy.

The structure of the paper is as follows. Section 2 reviews the important models currently available in the literature related to the intended research. Section 3 describes the proposed strategy for improving the data quality that is suitable for large data. The preprocessing tasks that must be completed before using the algorithms are presented in Section 4. Section 5 describes the map phase, which applies the proposed conjoint analysis to attribute importance and CA-mRMR. Section 6 describes the reduced phase, which uses a q-gram filter to exclude irrelevant occurrences. Section 7 presents an experimental analysis, and Section 8 presents the results. Section 9 discusses the theoretical and practical implications as well as challenges in data science and future work, and Section 10 offers the conclusion and final directions for the intended research.

2. Related Works

Recent research highlights that data quality is fundamental for efficient analytics and machine learning, especially in big data. According to a study [21] that offers a comprehensive overview of important data quality parameters, including completeness, consistency, timeliness, and accuracy, low-quality data might undermine downstream analysis even when algorithms are strong. Similarly, results have shown that data quality affects organizational decision-making and model performance, highlighting how antecedents such as source dependability and preprocessing approaches influence data quality [22]. It was also noted that in data-driven contexts, higher-quality data allows for more accurate, timely, and competitive insights, which in turn increases the strategic value of business intelligence tools [23]. Each of these studies adds weight to the growing body of evidence that preprocessing procedures should include data quality evaluation and improvement as a means to increase robustness and ensure meaningful results in high-dimensional applications.

Researchers in the literature have produced several publications on dimensionality reduction and data quality enhancement. A conceptual framework for researching the connection between big data quality and knowledge management, which affects decision-making quality, was proposed to enable a thorough investigation [24]. This research demonstrates that management information systems rely heavily on high-quality big data for making essential decisions. According to [25], there are six ways to measure

data quality: comprehensiveness, uniqueness, timeliness, validity, precision, and consistency. To improve the quality of big data created for the telecommunications sector, a paradigm based on process patterns was proposed [26].

By eliminating unnecessary features, which increases accuracy, feature selection is regarded as one of the most focused approaches for enhancing data quality. A feature selection technique for text classification from big datasets was developed by combining ant colony optimisation with an artificial neural network [27]. With the k-Nearest Neighbour (KNN) algorithm at its core, a hybrid feature selection model was suggested that works well in cloud settings. It employs the firefly distance metric in addition to the Euclidean distance metric to find the nearest neighbours [28, 29]. There has been research on the advantages and disadvantages of classical feature selection techniques as well as streaming feature selection models that are useful for decision-making in organisations [17]. A technique known as the Parallel, Forward–Backwards with Pruning (PFBP) model was proposed to find important features. To lower the communication cost, the model divides the data into rows and columns and efficiently uses a greedy search [30].

Ensemble-based mRMR (EmRMR) is a novel feature selection approach that was presented for reusing the dimensions in a big dataset [31]. In addition to verifying the highest relevance of the feature associated with the class attribute and the minimal redundancy with other candidate attributes, the model calculates the mutual information for selecting the significant attributes with higher values. A number of modifications were performed using this model as the foundation to enhance the system's performance. To address the shortcomings of mRMR when used with massive data, an EmRMR model was created as an extension of the mRMR model that uses a variety of optimisation strategies at different phases [32]. This considerably decreased the conventional mRMR algorithm's computational complexity.

Another variant, known as the MR-mRMR method, uses the mRMR algorithm inside the MapReduce architecture [33]. To find relevant features, this model, however, uses the RELIEF method rather than Mutual Information (MI) [34]. According to this approach, speed, accuracy, and simplicity are greatly improved by the smaller attribute set [35]. The elimination of outliers is crucial in addition to feature selection through dimensionality reduction. After a study on banking data, a methodology was proposed to eliminate data inconsistencies and enhance the data quality [36]. A deep neural network model was used to make predictions in a data-mining model proposed to identify noise and evaluate the data quality. Although the model's effectiveness was examined using banking data, the authors ensured the model's ability to work with data from other industries [37]. In addition to the KNN technique, four other reduction types were proposed and compared: heuristic-based data cleaning, rule aggregation, rule synthesis using q-gram-based filtering, and ensemble clustering [38]. The Number of features and instances was decreased using this technique based on their importance.

To improve the performance of high-dimensional biological datasets, researchers have recently investigated hybrid feature selection methods. A hybrid strategy was suggested that combined Binary Portia Spider Optimization with the fastest mRMR (FmRMR) to improve convergence speed and classification accuracy [39]. A technique combining improved mRMR (ImRMR) with Binary Differential Evolution has been proposed for gene selection, balancing redundancy and importance in the chosen traits [40]. In clinical datasets, the ReliefF-mRMR method has demonstrated strong efficacy in identifying various diseases [41]. Despite their advantages, these techniques often suffer from limited generalizability across diverse datasets, high computational costs, and considerable need for parameter adjustment. These recent developments support the rationale behind the proposed CA-mRMR approach, which addresses these limitations by combining the complementary strengths of the classical and metaheuristic selection procedures.

Owing to their robustness, interpretability, and ability to handle high-dimensional data, well-known ML algorithms such as the Support Vector Machine (SVM), Naïve Bayes (NB), KNN, Logistic Regression (LR), Random Forest (RF), and Decision Tree (DT) are frequently used in big data classification tasks. SVMs are particularly effective in complex, nonlinear classification scenarios, especially when applied to large-scale data such as clinical and gene expression datasets [42]. NB and KNN are commonly applied because of their simplicity and efficiency in managing large datasets, especially when model transparency is essential [43]. LR is appropriate for binary classification tasks because it provides probabilistic outputs along with solid baseline performance [44]. Ensemble methods such as RF and traditional DTs are widely used because of their ability to model feature interactions and reduce overfitting through bootstrap aggregation [45]. Many recent investigations in big data and related domains have relied on these models for validation and performance evaluation [40, 41].

3. Proposed Framework for Data Quality Improvement

The overall design of the data quality improvement framework using conjoint analysis with minimum redundancy, maximum relevance, and q-gram-based filtering for big data is shown in Figure 1. This framework aims to provide quality data for effective processing, thereby producing accurate results specifically for big data. However, it is very difficult to find a single approach that can improve the quality of data by overcoming the challenges encountered by big data. The proposed framework has three phases in which the first phase focuses on data extraction and cleaning, the second is the map phase that implements Conjoint

Analysis with the Minimum Redundancy Maximum Relevance model (CA-mRMR), and the third is the reduce phase that implements the q-gram-based filtering approach for detecting redundant and irrelevant instances. Each phase and the corresponding algorithms used in the framework are explained in the following sections.

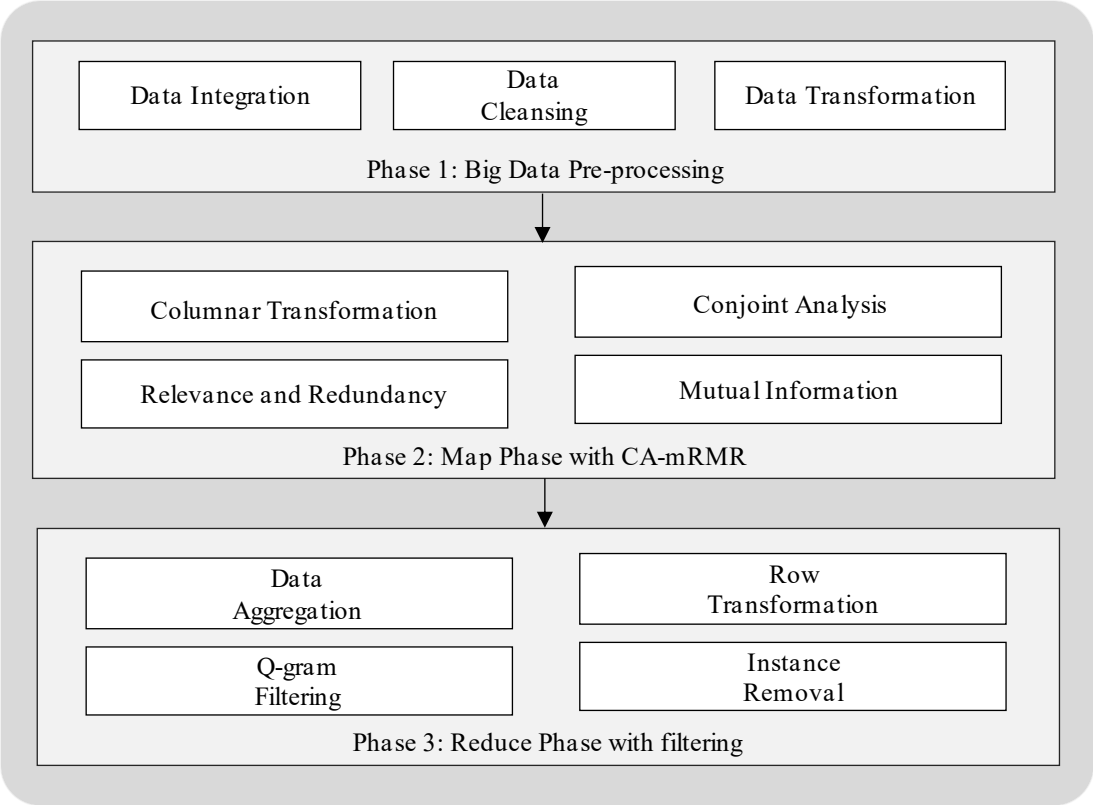

Fig. 1 Overall design of the proposed data quality improvement framework

4. Big Data Preprocessing Phase

The performance of the mining or knowledge extraction system depends not only on the underlying logical model but also on the input data quality given to the system [46]. Generally, to extract knowledge from given data, it is essential to prepare the raw data in a particular form that can be processed. Thus, the preprocessing phase of big data aims to extract the data related to a particular event and convert it to a suitable form for further processing. This phase is inevitable for big data because it incorporates enormous amounts of data from multiple sources. This phase includes data integration, cleaning, and transformation.

4.1. Big Data Integration

This process is the initial step, which deals with collecting data from various sources in various formats. Thus, combining the data extracted from various sources is necessary to form a unified data view. When it comes to big data, data integration is performed by first gathering information from social media platforms, the Internet of Things, and transactional data. Although it is extracted from various sources, the integration process results in a single target form. In the case of data warehouses, the Extraction, Transformation, and Loading (ETL process) is used to consolidate the data. With evolving technologies, the ETL process has evolved to be used with big data [47].

4.2. Data Cleansing

This process aims to remove incomplete, irrelevant, and incorrect data from the datasets to ensure the quality of the data. This step is substantial, as the error and incomplete data lead the user to make immoral conclusions and decisions. In the case of business applications, poor data may lead to a loss of money, whereas in the most critical applications, such as medical data, poor decisions can lead to loss of life. Several methods exist for cleaning big data that vary from simple models, such as deduplication and irrelevant data removal, filling missing data, to complex ML models for parsing, mining association rules, and mining outliers [48]. Here, with a huge volume of data, if the Number of incomplete records is minimal, then the records can be filtered out. The other steps, removing redundant data and reducing the dimensions, were carried out in the next phases.

4.3. Data Transformation

Data transformation involves transforming data from one form to another for easy processing. In general, data from the source is mapped to the specified targeted form. This targeted

form can be used to obtain effective results. This can be achieved by scripting or using ETL tools. Although the method is substantial, it is time-consuming and a slow process requiring higher human resources, especially for unstructured data. Once the preprocessed big data has been retrieved, the map phase is used to select the important features that, through enhanced predictive power, represent the complete dataset.

5. Map Phase using CAFS-mRMR

The preprocessed data are normalized, and the normalized data are partitioned based on the Number of mappers. For each partition, the mapper applies a columnar transformation before applying enhanced conjoint analysis with the mRMR algorithm. However, in general, all ML algorithms are row-oriented or instance-oriented, in which they apply and evaluate the data in a row, which requires a higher cost for random access. Conversely, feature selection algorithms are processed in a columnar manner for efficiency, in which each column represents a specific instance with various attributes. Thus, the main aim of applying columnar transformation is to improve the overall performance of the process. The row-oriented and column-oriented representations of features in the contiguous memory location are presented in Figure 2, in which a) represents the row-oriented representation with four records having three features each, and b) represents the columnar representation of the same data.

Instance 1			Instance 2			Instance 3			Instance 4		
f11	*f12*	*f13*	*f21*	*f22*	*f23*	*f31*	*f32*	*f33*	*f41*	*f42*	*f43*

a) Row Oriented Representation

Feature 1				Feature 2				Feature 3			
f11	*f21*	*f31*	*f41*	*f12*	*f22*	*f32*	*f42*	*f13*	*f23*	*f33*	*f43*

b) Columnar Oriented Representation

Fig. 2 Different feature representation

For the transformed values, the data were then passed to the mapper, where the proposed conjoint analysis with the mRMR approach was applied. The result is a subset of features that strongly represent the entire dataset, which helps in effective prediction. The worker nodes in the map phase apply the proposed CA-mRMR algorithm to the input partitions and provide the key-value pair as the output, which, in turn, serves as an input for the reduce phase. The result of the map phase is a set of significant features after applying the CA-mRMR algorithm, as given in Equation 1:

$$FP_i = < f_{i1}, f_{i2}, \ldots, f_{in} > \quad (1)$$

Here, FP_i represents the subset of features selected from partition i, and f_{in} represents the binary value that is 1 for the selected feature and 0 for the non-selected features.

To provide more clarity on the feature selection process, conjoint analysis and the CA-mRMR algorithm are explained in detail below.

5.1. Conjoint Analysis for Attribute Significance

The Number of representation features to be chosen from the complete collection can be determined by evaluating the attribute importance, and is then given as an input for the mRMR algorithm. Conjoint analysis aims to identify the importance of features based on which the model selects the substantial attributes in the given dataset [49]. In the proposed model, conjoint analysis was performed to identify and remove irrelevant attributes based on the attribute significance score, which is the approximate k-value that specifies the Number of relevant attributes to be selected for the mRMR model. This analysis identifies the significance of an attribute with respect to the total dataset by using the utility range or preference range. However, the computed attribute's significance always specifies the relative value concerning the other attributes, and thus, the sum of the utility values of all the attributes is always 1.

Initially, the utility of each attribute value was calculated by determining the mean partworths of a particular attribute value. However, the partworths for the attribute at various levels can be estimated using various models, such as dummy variable regression or the LOGIT model. Upon identifying the part and attribute utility values, the utility range of a property can be determined by calculating the difference between its maximum and minimum utility values, as shown in Equation 2.

$$Rng_{Attr_i} = max(u_{attr_i}) - min(u_{attr_i}) \quad (2)$$

The relevance of an attribute can be calculated as the ratio of the utility range of that attribute to the overall utility range of all attributes, as shown in Equation 3.

$$Attr\_Sing_i = \frac{Rng\_Attr_i}{\sum_{i=1}^{n} Rng\_Attr_i} \quad (3)$$

Attributes with very low scores were filtered based on the above significance score. The remaining attributes and their utility values are retained as inputs for the mRMR process. This step effectively reduces dimensionality while preserving the predictive relevance.

5.2. CA-mRMR Model

Generally, mRMR is a model used for feature selection in which the input is the set of records, and the output is the ranked attribute based on its significance. The features are ranked based on the relevance of the attribute to that of the class variable, thereby penalizing the redundancy between the attributes. Thus, it aims to maximize the dependency between the attribute and target variables and minimize the redundancy between the dependent variables. To compute the relevancy

between feature f and target class c, weighted mutual information is computed, in which the attribute significance obtained through conjoint analysis serves as the weights for the attributes. The formula for computing the weighted mutual information between attribute A and target class C is given in Equation 4:

$$wMI(A,C) = \sum_{a\epsilon A}\sum_{c\epsilon C} w(a)p(a,c)log\left(\frac{p(a,c)}{p(a)p(c)}\right) (4)$$

Here, *w(a)* is the utility value of attribute a. The main aim is to maximize the attribute relevance of the target class and minimize the redundancy between the two attributes. These can be achieved using the maximization and minimization constraints given in Equations 5 and 6, respectively:

$$Max\ Rel(A,C) = \frac{1}{|A|}\sum_{a\epsilon A} wMI(a,C) \quad (5)$$

$$Min\ Red(A) = \frac{1}{|A|}\sum_{a_i,a_j\epsilon A} wMI(a_i,a_j) \quad (6)$$

The CA-mRMR algorithm first selects the attribute with the highest relevance score. Then, in each iteration, the redundancy between the newly selected attribute and all the remaining candidates is calculated. The attribute that maximizes (relevance redundancy) was added to the final selected set. This process is repeated until k features (determined from conjoint analysis) are selected.

In the case of the CA-mRMR algorithm, the relevance score for all attributes is computed using weighted mutual information and stored in memory. The feature with the highest relevance mutual information score compared to that of the target class variable was selected.

Then, the selected attribute acts as a reference for computing redundancy, in which mutual information is computed with the selected attribute and that of each unselected attribute. The redundancy values were accumulated at each iteration. This process continues for all features selected using conjoint analysis. The algorithm for the proposed CA-mRMR algorithm is as follows.

```
Input: a set of features from the dataset D
Output: Significantly selected features
begin map_phase()
final_set = (); candidate_deature={}; total_range_utility = 0;
//Conjoint analysis
for each feature f do
    //Apply Logit model to find the part worths
    partworths[f]=Logit();
    //Attribute-value utility computation
    for attribute value i from 1 to n do
        utility_attr_i = avg(part_worths_i)
    end for
    range_attr_i(f) = (max(utility_attr_i)-
    min(utility_attr_i))
    total_range_utility = total_range_utility +
    range_attr_i(f)
end for
for each feature f do
    //Significance score computation
    sig_score(f) = range_attr_i(f) / total_range_utility
    if sig_score(f) > threshold then
      candidate_feature = candidate_feature ∪ f
      //Number of features to be selected
      nfeature=nfeature+1
    end if
end for
//Feature selection with weighted mutual information
for candidate_feature f in D do
    relevance[f] = wMI(f, class);
     accumulatedRed[f] = 0;
end for
selected = getMaxRelevance(relevance);
final_set.add(selected);
candidate_feature.remove(selected);
while final_set.size() < nfetaure do
    for feature f in candidate do
        rel = relevance[f] //relevancy
        //redundancy computation.
        red = wMI(f, selected);
        //Score with max. relevance & min.
        redundancy
        CA_mRMR = rel - red;
        if CA_mRMR is max then
            selected = f;
        end if
    end for
    final_set.add(selected);
    candidate_feature.remove(last_selected);
end while
end procedure
```

To optimize the algorithm, the computed values were stored in the cache for further use. This minimized the Number of computations required. In addition, to minimize random memory access, columnar transformation was initially performed before applying the algorithm. These simple modifications significantly influence the performance of the model without altering the final results. It was also implemented in Apache Spark using MapReduce. The input data is split into partitions and set as inputs for the Hadoop Distributed File System (HDFS) worker nodes at the map phase. These worker nodes process the files and produce results in a key-value pair that contains the selected set of features.

6. Reduce Phase using Q-Gram Filter

The training dataset and output from the map phase were subsequently sent through the master node to the reduce phase, where the inputs were merged and sorted. These were then

further analysed to identify and eliminate redundant instances, resulting in a clean, reduced dataset. A data integration process was performed using the binary vector results from the different worker nodes to determine the important features. The averages of these binary vectors were calculated to identify the most significant features. Equation 7 provides the formula used to compute the average of the binary vectors obtained from the mapping phase.

$$SA_i = \frac{1}{k}\sum_j f_{ij}, \forall j = 1, 2, .. k \quad (7)$$

Here, j indicates the Number of worker nodes in the map phase that varies from 1, 2, …, k, and i represents the Number of features in the datasets that range from 1, 2, …, n. Thus, the attributes with maximum values are selected.

Because the inputs obtained from the mapper nodes are feature- or column-oriented, the input dataset partitions must be converted to instance-oriented or row-oriented, in which each row depicts an instance from the dataset. A row representation of the dataset with four instances and three attributes is shown in Figure 2. This step significantly increases the performance of the system because it reduces the unnecessary time required to perform a random search.

6.1. Q-Gram-based Filtering Approach

The training sets given to the map phase are also presented as inputs to the reduction phase. However, the dimensionally reduced dataset was used, in which the attributes that were not selected in the previous step were removed from the dataset for further processing. A Q-gram-based filtering approach was used to eliminate irrelevant instances [38].

Thus, all attribute values in each instance are concatenated to form a single string for which the Q-gram approach is applied to find duplicate records. Specifically, Q-gram Alignment based on Suffix Arrays (QUASAR), a simple enhanced q-tuple filtering model that verifies the pattern match, is applied to the data to identify the redundancy [50].

It identifies various occurrences of a specific q-pattern S by performing local approximate matches in the given data D. The algorithm applies edit distance, in which patterns S and d_i can have at most k values. It also uses a q-gram index to specify the Number of occurrences for each substring. The matched patterns are stored in a matching block and sent to the next phase for further analysis. The use of the q-gram and matching blocks reduces the overhead of the underlying system.

The matched patterns are effectively analyzed, based on which the decision on irrelevant and redundant instances can be made. Thus, if pattern S has a complete match with the instances in database D, this implies that the instances are duplicates or redundant and can be removed. In addition, irrelevant instances that do not have any matches are considered outliers and are discarded, which also enhances the performance of the system. The conditions for relevancy, redundancy, and irrelevancy are given in Equation 8.

$$fltr(instance) = \begin{cases} s = d_i, if\ redundant \\ s \cap d_i = \emptyset, if\ irrelevant \\ s \cap d_i \neq \emptyset, if\ relevant \end{cases} \quad (8)$$

The algorithm for the proposed q-gram-based filtering is given below.

```
Algorithm: Q-gram_filtering
Input: set of partitions and features
Output: Reduced dataset
Begin reduce_phase()
//Integration of results from mapper nodes.
for each binary vector b do
    //attribute value utility computation
    for attribute value i from 1 to n do
        selecti=selecti + <bi>
    end for
end for
for each attribute i do selecti = selecti /b
   if selecti is maximum, then
        Select the attribute
   end if
end for
//Q-gram-based filtering
for each instance i in the dataset d do
    patterni = concat(all_attibute values)
end for
for each patterni do
    if patterni ==patternd then
        patterni.remove() //Redundant instance
    elseif no element in patterni match patternd then
        patterni.remove() //Irrelevant instance
    elseif subset of patternd matches with subset of
    patternd then
        patterni.select() //Relevant instances
   end if
end for
end procedure
```

Figure 3 shows the overall workflow design of the proposed framework for enhancing the quality of big data.

7. Experimental Analysis

An evaluation of the proposed approach concerning the quality improvement of the data to be processed is discussed. The testing environment used for the proposed m-MRMR model and the outcomes acquired for different inputs were examined and compared with the existing models.

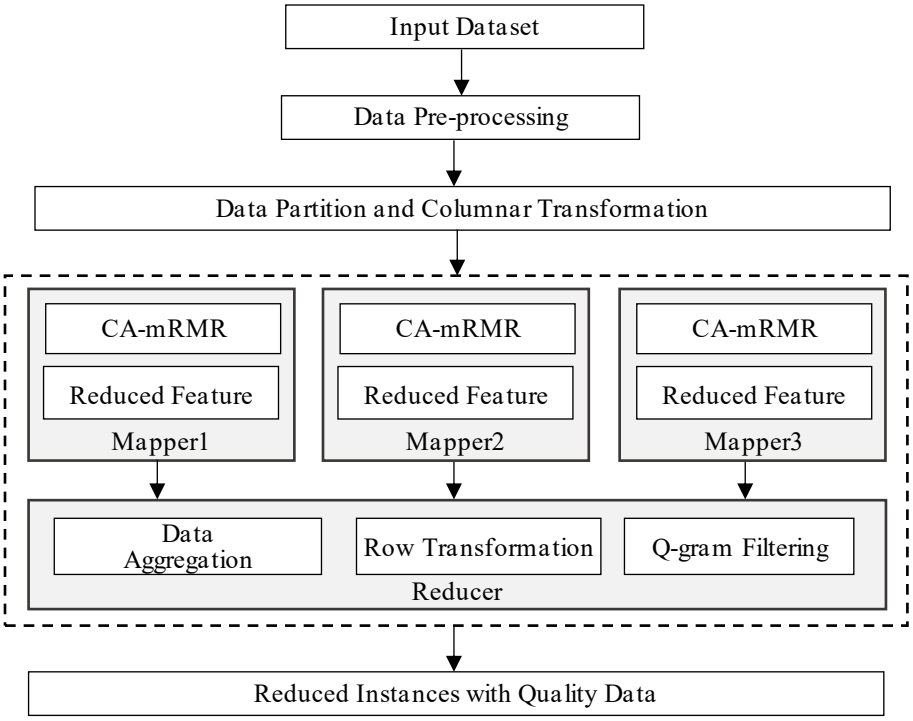

Fig. 3 Workflow of the proposed quality improvement framework for big data

7.1. Experiment Setup

For performing the trials and to analyze the effectiveness of the proposed approach, a set of four nodes was used in the clustering step, with a single node as the master, whereas the remaining three nodes were identified as slaves. The system configuration and the configuration of the computing nodes are as follows: Intel Core i3 CPU processor, two cores per processor with 3.00 GHz, 64 GB RAM, and 1000 GB Hard Disk. The computer has the following software configuration: Open Source Apache Hadoop with Apache Spark and machine learning library (MLlib) version of 1.2.2, and HDFS with a block size of 128 MB. The master node manages the HDFS, controls, and coordinates with slave nodes. It uses a MapReduce framework with a set of nodes at a map phase and a single reduce phase, in which the input data is split into several partitions and then fed as input for the mapper nodes. Upon applying the CA-mRMR algorithm to the data in the mapper nodes, a reduced set of instances is obtained as the output, which is then fed to the reduce phase for performing aggregation and Q-gram-based filtering. The final result of the reduce phase is the quality of the dataset after removing irrelevant and redundant attributes and instances.

7.2. Datasets Used

This study employed three frequently used datasets to analyse the efficacy of the feature selection algorithm for big data. The details of the dataset used for the proposed study, including the Number of instances and attributes, along with the source details, are presented in Table 1. The Epsilon dataset is a part of LibSVM and has many attributes, while ECBDL14 is a dataset that has many instances. The Susy dataset has a minimum number of 18 attributes and a maximum number of records and is available at the UCI repository.

Table 1. Dataset used

Dataset	#Attr.	#Inst.	Source
Epsilon	2000	400000	LibSVM
Susy	18	5000000	UCI repository
ECBDL14	631	7994298	GECCO-2014 International conference

The original datasets were preprocessed to address missing values, outliers, and inconsistencies during collection to ensure data quality. According to the 3σ rule, outliers were defined as any result that did not fall within the interval $(\mu - 3\sigma, \mu + 3\sigma)$. Furthermore, both missing values and identified outliers were replaced with representative values from comparable cases by using the KNN imputation technique [40]. Duplicate records and unnecessary features were eliminated during the initial cleaning. To improve the classifier's performance and interpretability, min–max normalization was used to bring all features onto a unified scale.

By changing the dataset size (50 K, 100 K, and 200 K instances), a sensitivity analysis was performed to assess the model's resilience and scalability. Using several classifiers and datasets, the effects of these differences on classification accuracy and execution time were examined. This allowed for better evaluation of the stability of the proposed CA-mRMR model across varying data volumes. The data quality was

maintained during the preprocessing and selection procedures, which included utility-based feature scoring and instance filtering. These methods reduce redundancy and improve overall information density. Ethical data management is a crucial element of research that utilizes high-dimensional information. Only publicly accessible and anonymised datasets were used in this study, ensuring that no personal information could be directly linked to any individual. All the data complied with the relevant licenses and terms of service. During processing and analyzis, privacy and confidentiality were carefully maintained. Feature selection and classification were consistently implemented to prevent bias and promote equity. To minimize overfitting, five-fold cross-validation was used to evaluate the final results, and all the reported metrics adhered to the established procedures.

8. Results and Discussion

8.1. Result for the Proposed Model

This section provides various analyses based on experimental data. Initially, a comprehensive study was conducted for the suggested model, utilizing the various datasets provided in Table 1. The classification accuracy of the quality-improved big data was evaluated using various standard classifiers such as SVM, NB, KNN, LR, RF, and DT with default parameters. The experiment was performed in batches by varying the size of the datasets, which refers to the Number of instances as 50 K, 100 K and 200 K. Table 2 presents the findings detailing the classification accuracy as well as the duration required for training and testing the proposed model using the Epsilon dataset.

Table 2. Results of the proposed model with the Epsilon dataset

Classifiers	Size	Acc. (%)	Train. (sec)	Test. (sec)	Total (sec)
SVM	50K	80.91	1.37	1.12	2.49
	100K	79.62	1.68	1.43	3.11
	200K	78.77	1.9	1.51	3.41
NB	50K	79.93	1.4	1.21	2.61
	100K	78.96	1.58	1.18	2.76
	200K	76.25	1.81	1.49	3.3
KNN	50K	80.75	1.27	1.05	2.32
	100K	79.11	1.41	1.19	2.6
	200K	77.95	1.66	1.4	3.06
LR	50K	78.64	1.16	0.93	2.09
	100K	77.23	1.4	1.16	2.56
	200K	75.65	1.59	1.23	2.82
RF	50K	81.99	1.68	1.4	3.08
	100K	80.72	2.05	1.76	3.81
	200K	79.94	2.36	1.89	4.25
DT	50K	81.66	1.46	1.23	2.69
	100K	80	1.58	1.34	2.92
	200K	78.88	1.84	1.56	3.4

The average classification accuracies with varying instance counts of 50 K, 100 K, and 200 K using SVM, NB, KNN, LR, RF, and DT were 80%, 78%, 79%, 77%, 81%, and 80%, respectively. The average time taken for both training and testing the model using SVM, NB, KNN, LR, RF, and DT is 3.00s, 2.89s, 2.66s, 2.49s, 3.71s, and 3.00s, respectively. From this analysis, it is evident that the accuracy and execution duration of the model are closely related to the dataset size. Notably, LR achieved the lowest accuracy with the shortest runtime, whereas RF achieved the highest accuracy with the longest computational time, highlighting the tradeoff between performance and time complexity. The results obtained by varying the size of the datasets with sample counts of 50 K, 100 K and 200 K for the suggested work on the Susy dataset using different classifiers, such as SVM, NB, and KNN, are shown in Table 3. The mean classification accuracy with 50 K, 100 K and 200 K instances using SVM, NB, KNN, LR, RF, and DT are 74.44%, 73.22%, 74.12%, 71.86%, 75.56%, and 75.12%, respectively. Training and testing the model with SVM, NB, KNN, LR, RF, and DT required an average of 2.41, 2.29, 2.07, 1.93, 3.04 and 2.40 seconds, respectively.

Table 3. Results of the proposed model with the Susy dataset

Classifiers	Size	Acc. (%)	Train. (sec)	Test. (sec)	Total (sec)
SVM	50K	75.29	1.08	0.77	1.85
	100K	74.49	1.41	1.11	2.52
	200K	73.56	1.69	1.18	2.87
NB	50K	74.57	1.17	0.82	1.99
	100K	73.96	1.36	0.85	2.21
	200K	71.14	1.55	1.13	2.68
KNN	50K	75.77	1.03	0.74	1.77
	100K	73.91	1.21	0.8	2.01
	200K	72.69	1.39	1.03	2.42
LR	50K	73.02	0.89	0.56	1.45
	100K	72.17	1.17	0.86	2.03
	200K	70.39	1.4	0.9	2.3
RF	50K	76.37	1.37	1.05	2.42
	100K	75.59	1.68	1.39	3.07
	200K	74.73	2.08	1.56	3.64
DT	50K	76.67	1.2	0.91	2.11
	100K	74.78	1.36	0.97	2.33
	200K	73.6	1.58	1.19	2.77

The results obtained by changing the size of the datasets with the Number of records as 50 K, 100 K and 200 K for the proposed quality improvement framework on the ECBDL14 dataset using different classification algorithms, such as SVM, NB, and KNN, are shown in Table 4.

Table 4. Results of the proposed model with the ECBDL14 dataset

Classifiers	Size	Acc. (%)	Train. (sec)	Test. (sec)	Total (sec)
SVM	50K	85.01	1.18	1.02	2.2
	100K	83.73	1.49	1.33	2.82
	200K	82.98	1.71	1.41	3.12

NB	50K	84.15	1.21	1.11	2.32
	100K	82.97	1.39	1.08	2.47
	200K	80.49	1.62	1.39	3.01
KNN	50K	84.94	1.08	0.95	2.03
	100K	83.27	1.22	1.09	2.31
	200K	82.00	1.47	1.3	2.77
LR	50K	82.49	1	0.85	1.85
	100K	81.21	1.26	1.09	2.35
	200K	80.46	1.43	1.14	2.57
RF	50K	86.09	1.47	1.3	2.77
	100K	84.83	1.84	1.67	3.51
	200K	84.15	2.03	1.74	3.77
DT	50K	85.84	1.25	1.12	2.37
	100K	84.14	1.37	1.26	2.63
	200K	82.91	1.66	1.41	3.07

The mean accuracies with varying instance counts (50 K, 100 K, and 200 K) using SVM, NB, KNN, LR, RF, and DT classifiers were 83.9%, 82.54%, 83.4%, 81.4%, 85%, and 84.3%, respectively. As shown in these classifiers are 2.71s, 2.6s, 2.37s, 2.26s, 3.35s, and 2.69s, respectively. The average values for the three datasets obtained using various standard classifiers are shown as a bar graph in Figure 4. The average classification accuracies across the six classifiers for each dataset were as follows: epsilon (M = 79.28%, Var = 1.77), Susy (M = 74.04%, Var = 1.77), and ECBDL14 (M = 83.43%, Var = 1.70). According to these descriptive statistics, ECBDL14, Epsilon, and Susy exhibited varying levels of average accuracy. The relatively small variances indicate consistent classifier performance within each dataset. A one-way ANOVA was performed to determine whether the mean classification accuracy of the six classifiers varied significantly across the datasets. The analysis revealed a statistically significant difference ($F = 75.93$, $p < 0.001$) with the computed F-value exceeding the critical threshold ($F_{crit} = 3.682$). This result confirms that the type of dataset substantially affects the classification performance, highlighting the importance of evaluating the proposed CA-mRMR model across diverse datasets to ensure robustness and generalizability. The average execution times for the proposed CA-mRMR-based feature selection method, evaluated across the epsilon, suspicious, and ECBDL14 datasets using the SVM, NB, KNN, LR, RF, and DT classifiers, are presented in Figure 5.

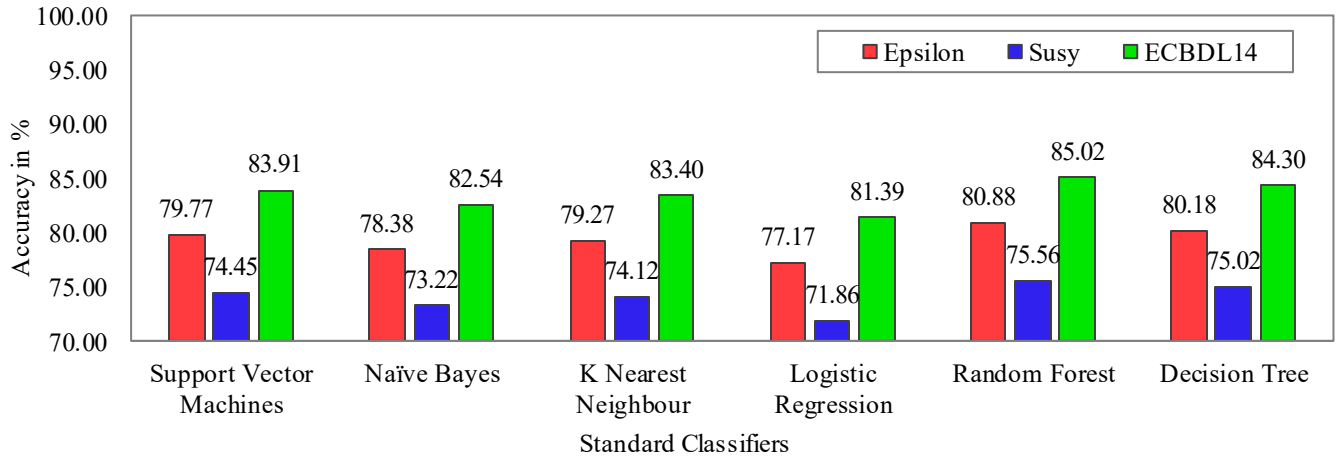

Fig. 4 Execution time analysis for the proposed model

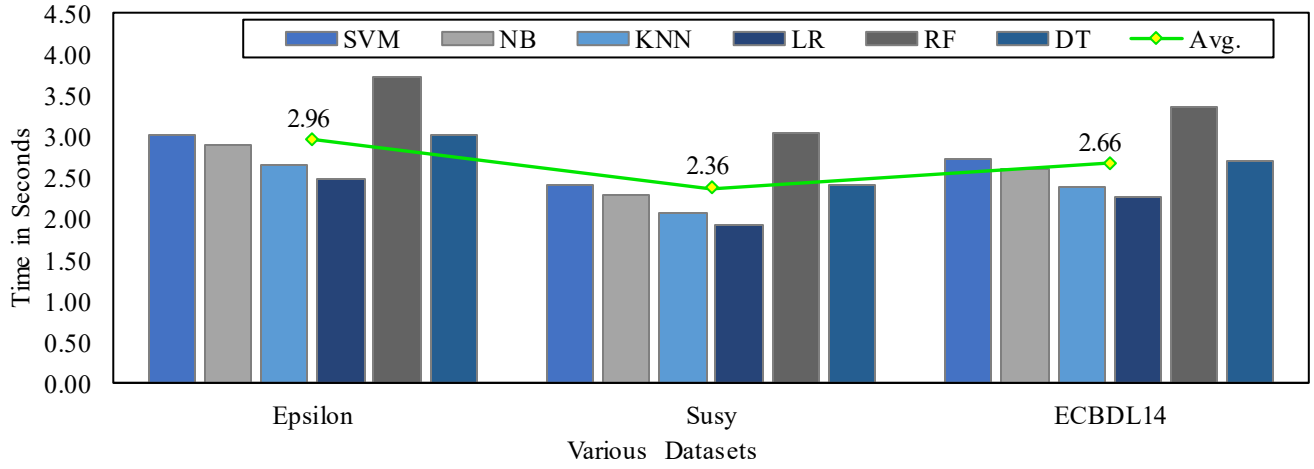

Fig. 5 Execution time analysis for the proposed model

A one-way ANOVA was performed to ascertain whether the mean execution times of the six classifiers differed significantly. The results showed a statistically significant difference in execution time ($F = 5.18$, $p = 0.009$). The null hypothesis was rejected because the computed F-value exceeded the critical value ($F\_crit = 3.11$), indicating that the classifier type significantly affected the execution time. Descriptive statistics revealed that LR had the shortest average execution time ($M = 2.22$s), whereas RF had the longest ($M = 3.37$s). These findings emphasize the importance of considering computational efficiency and accuracy when selecting a classifier.

8.2. Comparison with State-of-the-Art Models

A comparison has been made for the proposed CA-mRMR algorithm, and the results are compared with other algorithms, such as the traditional mRMR [31], EmRMR [32], MR-mRMR [33], ImRMR [40], and FmRMR [39] algorithms.

The analysis was performed by varying the selection criteria for 100, 150, and 200 feature sizes. The execution times of both the suggested and existing algorithms are assessed, and the time required to process the complete datasets, such as Epsilon, Susy, and ECBDL14, are presented in Table 5.

Table 5. Execution time comparison by varying the number of attributes

Dataset	#Attr. Sel.	CA-mRMR	EmRMR	MR-mRMR	mRMR	ImRMR	FmRMR
Epsilon	100	1.05	1.84	1.03	7.49	2.13	1.93
	150	1.272	2.50	1.41	9.67	2.95	1.45
	200	1.69	2.51	1.72	12.45	3.09	2.17
Susy	5	0.41	1.20	0.57	8.9	2.29	1.27
	10	0.96	1.73	1.04	13.42	2.62	1.37
	15	1.05	2.44	1.12	17.85	2.14	1.69
ECBDL14	100	0.62	1.20	0.69	7.78	2.02	0.97
	150	0.71	1.21	0.84	12.11	1.80	0.89
	200	1.01	2.17	1.06	15.41	2.20	1.02

The average time taken by the Epsilon dataset for the proposed model is 1.34s, whereas that of other models, such as MR-mRMR, mRMR, EmRMR, ImRMR, and FmRMR models, are 1.4s, 9.87s, 2.29s, 2.72s, and 1.85s, respectively. Similarly, the average time taken by the Sysu dataset for the proposed model is 0.74s, whereas for the other models, such as MR-mRMR, mRMR, EmRMR, ImRMR, and FmRMR models, it is 0.81s, 13.39s, 1.79s, 2.35s, and 1.44s, respectively. For the ECBDL14 dataset, the average time taken for CA-mRMR, MR-mRMR, mRMR, EmRMR, ImRMR and FmRMR models is 0.78s, 0.86s, 11.77s, 1.53s, 2.01s and 0.96s, respectively. More specifically, the average time taken by various methods, including CA-mRMR, MR-mRMR, ImRMR, FmRMR, EmRMR, and mRMR models for the three different datasets was 0.97s, 1.05s, 2.36s, 1.42s, 1.87s and 11.68s, respectively. Thus, the time required for the proposed CA-mRMR algorithm is minimal compared with the traditional mRMR algorithm. Compared to other recent variants, such as MR-mRMR, EmRMR, ImRMR, and FmRMR, the CA-mRMR algorithm also shows lower or comparable execution times in most cases.

Although some models, such as MR-mRMR or FmRMR, perform slightly better in isolated cases, the difference is minimal. It can be considered insignificant when weighed against the overall data quality and performance. The values provided in Table 5 are presented as a graph in Figure 6 to visualize the time differences clearly.

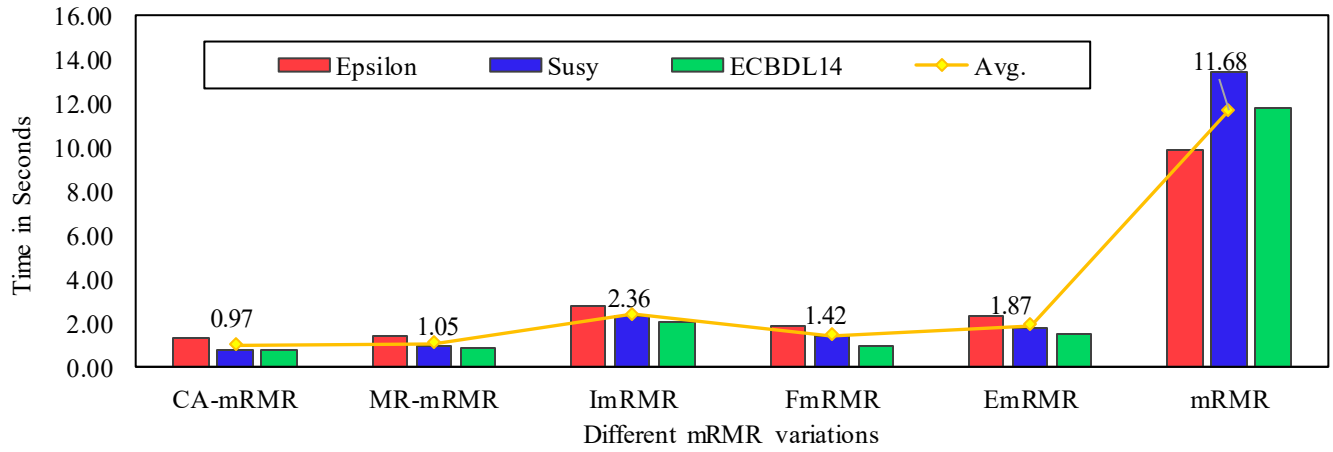

Fig. 6 Average execution time comparison for mRMR variations

The execution time and accuracy of the proposed model were compared with those of existing state-of-the-art models, such as mRMR, MR-mRMR, EmRMR, and ImRMR FmRMR. Various analyses were carried out by selecting 100 attributes from the Epsilon and ECBDL14 datasets and ten attributes from the Sysu dataset with different classifiers, such as SVM, NB, KNN, LR, RF, and DT classifiers. The times required to execute the model are listed in Table 6.

Table 6. Execution time comparison for mRMR variations with different classifiers

Classifiers	Datasets	Various Feature Selection Methods					
		mRMR	MR-mRMR	EmRMR	ImRMR	FmRMR	CA-mRMR
SVM	Epsilon	44.00	11.00	10.90	13.53	15.29	9.23
	Susy	36.00	13.23	13.08	15.39	16.89	11.12
	ECBDL14	29.00	10.51	9.25	11.29	13.30	8.12
NB	Epsilon	224.00	54.00	28.74	31.02	31.80	27.00
	Susy	251.00	42.00	22.17	24.43	25.88	21.00
	ECBDL14	182.00	29.00	18.40	20.91	21.92	17.00
KNN	Epsilon	40.45	9.31	8.56	10.55	12.19	6.85
	Susy	32.10	11.92	10.16	12.79	13.55	8.71
	ECBDL14	25.91	8.76	6.77	8.69	9.67	5.84
LR	Epsilon	39.88	7.39	8.34	11.59	13.66	7.01
	Susy	31.88	10.12	10.69	13.03	15.46	9.03
	ECBDL14	24.88	7.02	6.84	9.21	12.13	5.13
RF	Epsilon	50.72	16.71	17.29	20.86	22.99	15.17
	Susy	42.68	19.19	19.35	21.93	24.26	16.30
	ECBDL14	35.72	15.90	15.80	18.13	21.00	13.53
DT	Epsilon	41.76	10.73	10.04	12.39	14.32	8.28
	Susy	33.32	12.84	11.81	15.27	15.83	9.83
	ECBDL14	27.23	9.78	8.28	10.49	11.87	6.84

The speed-up rates of the proposed CA-mRMR model over the traditional mRMR algorithm (mRMR/CA-mRMR) with respect to the Epsilon, Susy, and ECBDL14 datasets using the SVM classifier were 4.77, 3.24, and 3.57, respectively. Compared with MR-mRMR (MR-mRMR/CA-mRMR), the speed-up rates were 1.19, 1.19, and 1.29, respectively. When considering all the methods, the maximum speed-up rate using SVM was achieved by CA-mRMR over mRMR (4.77×). Similarly, for the NB classifier, the speed-up rates of CA-mRMR over mRMR are 8.30 (Epsilon), 11.95 (Susy), and 10.71 (ECBDL14), and 2.00, 2.00, and 1.71, respectively. When extended to other classifiers, such as KNN, LR, RF, and DT, CA-mRMR consistently demonstrates a lower execution time than other methods, including EmRMR, ImRMR, and FmRMR.

While some models, such as EmRMR and ImRMR, occasionally show comparable performance, CA-mRMR provides the best overall tradeoff between execution time and selection quality. These results confirm that the CA-mRMR algorithm offers a consistently superior speed with minimal compromise, making it an efficient choice across different classifiers and datasets.

The analysis of the classification accuracy with that of the SVM, NB, KNN, LR, RF, and DT classifiers for different datasets, such as Epsilon, Sysu, and ECBDL14, was performed by selecting 100 attributes from the Epsilon and ECBDL14 datasets and 10 from the Sysu datasets. The values obtained for the accuracy of the classifier using different mRMR variations are listed in Table 7.

Table 7. Execution accuracy comparison for mRMR variations with different classifiers

Classifiers	Datasets	Various Feature Selection Methods					
		mRMR	MR-mRMR	EmRMR	ImRMR	FmRMR	CA-mRMR
SVM	Epsilon	61.80	80.32	81.47	80.71	80.96	83.26
	Susy	69.14	78.69	79.78	79.01	79.21	86.89
	ECBDL14	71.77	83.40	82.56	83.50	83.72	85.87
NB	Epsilon	59.31	78.80	79.43	78.81	78.80	82.05
	Susy	67.82	78.00	77.91	78.13	78.51	85.08
	ECBDL14	69.53	81.41	79.62	81.28	81.78	85.07
KNN	Epsilon	60.42	81.22	81.07	80.79	81.25	84.59
	Susy	69.51	80.35	79.44	80.35	80.16	87.51
	ECBDL14	70.79	82.85	80.51	82.46	83.62	86.31

LR	Epsilon	58.77	79.01	79.99	79.09	79.17	83.71
	Susy	67.76	77.94	78.11	78.59	77.88	85.62
	ECBDL14	69.23	81.06	78.60	80.71	81.65	84.43
RF	Epsilon	63.13	82.43	82.50	83.06	83.36	85.40
	Susy	70.34	80.40	81.53	80.75	81.26	89.27
	ECBDL14	73.73	85.42	83.90	85.32	85.25	88.45
DT	Epsilon	60.98	78.87	81.36	79.43	80.32	83.11
	Susy	68.88	77.63	79.58	78.16	78.25	87.26
	ECBDL14	71.03	82.11	81.87	82.94	82.85	85.96

The increase in the accuracy rate of the proposed CA-mRMR approach with that of the traditional mRMR algorithm for the Epsilon, Susy, and ECBDL14 datasets using the SVM classifier was 34.72%, 25.67%, and 20.76%, respectively, and the rate of increase with that of the MR-mRMR algorithm for the same datasets was 3.66%, 10.43%, and 2.96%, respectively. For the NB classifier, the increases with respect to mRMR are 22.74%, 17.26%, and 22.53%, and those with respect to MR-mRMR are 4.13%, 9.07%, and 4.49%, respectively. Similarly, the increases for KNN with mRMR are 24.17%, 25.85%, and 21.63%, and those with MR-mRMR are 4.14%, 8.91%, and 4.17%, respectively.

For LR, the rate of increase in accuracy with respect to mRMR was 24.94%, 26.31%, and 22.01%, and MR-mRMR was 5.95%, 9.83%, and 4.16%, respectively. In the case of RF, the improvements with mRMR were 22.27%, 26.91%, and 19.72%, and those with MR-mRMR were 2.97%, 8.87%, and 3.54%, respectively. Finally, the increases in accuracy for the DT classifier with mRMR were 22.13%, 26.68%, and 20.98%, and those with MR-mRMR were 4.24%, 12.39%, and 3.85%, respectively. Thus, the proposed CA-mRMR model consistently outperformed both mRMR and MR-mRMR across all classifiers and datasets.

The tradeoff between accuracy and execution time across various feature selection algorithms is evident from the experimental data. Although techniques such as EmRMR and ImRMR may offer competitive accuracy, CA-mRMR consistently requires a shorter execution time. The proposed CA-mRMR method reliably delivers the lowest execution time and highest classification accuracy across all classifiers and datasets.

Compared to mRMR (SVM, Epsilon), CA-mRMR achieves a notable accuracy improvement of up to 34.72%, along with a maximum speed-up of 4.77×. This balance demonstrates that the model can offer superior predictive performance without compromising computational efficiency, making it well-suited for real-time or large-scale applications where both accuracy and speed are critical.

8.3. Comparison of Methods with Biomarker Datasets

Another comparison was made between the proposed CA-mRMR algorithm and the EmRMR [32] and traditional mRMR [31] algorithms and conventional feature selection methods with other biomarker datasets such as lung, NCI, Colon, Leukemia, Lymphoma, DLBCL and Gastric. A complete list of these datasets is available at GitHub (https://github.com/xwdshiwo/BioFSDatasets_and_code).

Table 8 displays the execution time calculated using the SVM classifier for both the proposed and existing methods using five-fold cross-validation. The table displays the datasets, including the Number of attributes and instances, along with the execution times for the different models.

Table 8. Execution time comparison for biomarker datasets

Datasets	#Attr.	#Inst.	mRMR	EmRMR	CA-mRMR
Lung	73	326	23.27	0.06	0.26
NCI	60	9173	39.71	2.02	1.93
Colon	62	2000	40.24	0.37	1.45
Leukaemia	72	7129	43.54	1.51	1.62
Lymphoma	45	4026	41.81	0.95	0.82
DLBCL	77	7129	48.56	1.67	1.73
Gastric	65	22,645	70.31	3.47	4.01

The speed-up rates of the proposed CA-mRMR model over the traditional mRMR algorithm (mRMR/CA-mRMR) using the SVM classifier for the Lung, NCI, Colon, Leukemia, Lymphoma, DLBCL, and Gastric datasets were 89.50, 20.58, 27.75, 26.88, 50.99, 28.06, and 17.54, respectively. In comparison, the speed-up rates of the EmRMR algorithm (EmRMR/CA-mRMR) for the same datasets were 0.23, 1.05, 0.43, 0.93, 1.16, 0.97, and 0.87, respectively.

Among all datasets, the maximum speed-up with respect to mRMR is observed for the Lung dataset (89.50×), while the highest improvement over EmRMR is seen for the NCI dataset (1.05×). These results indicate that CA-mRMR achieves substantial reductions in execution time compared to conventional methods, particularly when compared to mRMR, while maintaining or exceeding the performance efficiency of faster variants, such as EmRMR. The values presented in Table 8 are shown as a graph in Figure 7 to understand the time variations easily. From the experimental analysis, it is clear that the proposed model has the minimum execution time in many cases and yet has the best performance in terms of accuracy compared to many other existing models used for the study.

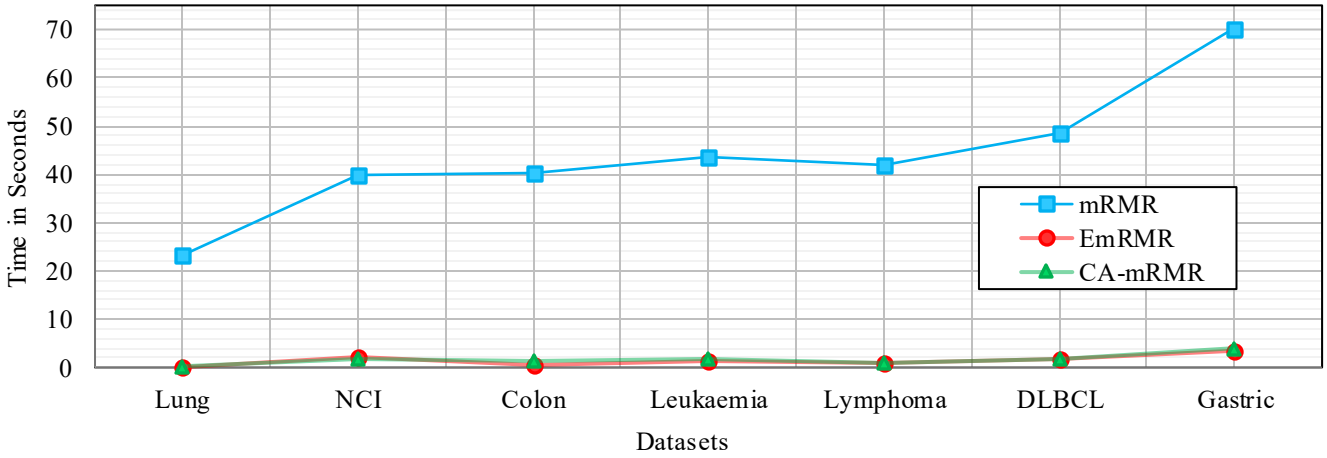

Fig. 7 Execution time comparison for different datasets

Moreover, using five biomarker datasets (colon, leukemia, lymphoma, DLBCL, and gastric), the efficacy of the suggested CA-mRMR feature selection method was compared with ten traditional and state-of-the-art methods: Lasso, Random Forest (RF), Logistic Regression (LR), Ridge, Correlation Coefficient (Corr), Decision Tree (DT), Mutual Information Coefficient (MIC), t-test, Stability Selection (Stab), and ImRMR. Accuracy, precision, recall, and F1-score were the four main performance metrics used for the evaluation, with SVM serving as the base classifier. SVMs were consistently used across all approaches to ensure impartial and unbiased evaluation, and the final performance indicator was the average classification accuracy derived from a five-fold cross-validation. This technique allows for uniform evaluation across datasets by ensuring robustness and reducing the possibility of overfitting. The results were compared with several classical feature-selection algorithms reported by Yu et al. (2024) [40]. To facilitate meaningful comparisons, the Number of selected features for each approach was kept constant.

In addition, the proposed approach showed little difference in performance across datasets, demonstrating its statistical dependability and generalizability across varying dataset sizes and feature distributions. Table 9 summarizes the accuracy, precision, recall, and F1-score results.

Table 9. Comparison with traditional feature selection methods

Datasets	Lasso	RF	LR	Ridge	Corr	DT	MIC	t-test	Stab	ImRMR	Proposed
Accuracy Values											
Colon	0.92	0.95	0.91	0.92	0.93	0.92	0.91	0.84	0.93	0.93	0.92
Leukemia	0.89	0.92	0.92	0.91	0.93	0.92	0.88	0.82	0.91	0.96	0.96
Lymphoma	0.99	0.97	0.97	0.95	0.95	0.96	0.9	0.87	0.99	1	0.99
DLBCL	0.94	0.94	0.94	0.96	0.94	0.96	0.91	0.89	0.94	0.97	0.98
Gastric	0.91	0.93	0.92	0.82	0.9	0.92	0.86	0.86	0.93	0.94	0.95
Precision values											
Colon	0.86	0.87	0.96	0.92	0.92	0.92	0.96	0.92	0.78	0.93	0.94
Leukemia	0.93	0.93	1.00	0.93	0.93	0.93	0.93	0.67	0.66	0.94	0.95
Lymphoma	0.98	0.97	0.84	0.99	0.98	0.96	0.96	0.96	0.71	1.00	0.99
DLBCL	0.88	0.88	0.96	0.96	0.86	0.82	0.96	0.88	0.69	0.98	0.98
Gastric	0.77	0.77	0.93	0.93	0.97	0.97	0.97	0.93	0.67	0.96	0.96
Recall Values											
Colon	0.93	0.90	1.00	0.88	0.93	0.93	0.93	0.93	0.90	0.95	0.94
Leukemia	0.92	0.92	0.72	0.96	0.96	0.92	1.00	0.60	0.74	0.94	0.96
Lymphoma	0.96	0.96	0.87	0.91	0.91	0.91	0.96	0.91	0.86	1.00	0.99
DLBCL	0.80	0.80	0.80	1.00	0.95	0.95	1.00	0.85	0.86	0.96	0.97
Gastric	0.98	0.97	0.89	0.93	0.93	0.89	0.86	0.89	0.83	0.96	0.97
F-measure											
Colon	0.89	0.89	0.98	0.90	0.92	0.92	0.94	0.92	0.84	0.94	0.94
Leukemia	0.93	0.93	0.84	0.95	0.95	0.92	0.97	0.63	0.70	0.94	0.95
Lymphoma	0.97	0.97	0.86	0.95	0.94	0.93	0.96	0.93	0.77	1.00	0.99

DLBCL	0.84	0.84	0.87	0.98	0.90	0.88	0.98	0.87	0.77	0.97	0.97
Gastric	0.86	0.86	0.91	0.93	0.95	0.93	0.91	0.91	0.74	0.96	0.96

Across all datasets, the proposed CA-mRMR approach consistently produced robust and stable accuracy. It showed competitive performance for lymphoma (0.99), leukemia (0.96), and colon cancer (0.92), either matching or slightly trailing the best methods. While Ridge attained good accuracy for DLBCL (0.96) and RF for Colon (0.95), their performance lacked consistency across datasets. Methods such as DT and MIC demonstrated moderate to good accuracy in isolated cases but showed variability. In contrast, CA-mRMR proved resilient to changes in the data structure and illustrated adaptability to high-dimensional gene expression datasets.

In addition, CA-mRMR was among the top performers in terms of precision. It reached near-maximum values for gastric cancer (0.96), leukemia (0.95), colon cancer (0.94), and lymphoma (0.99). Although LR achieved perfect precision (1.00) for leukemia, its low recall (0.72) indicated a tendency to misclassify true positives. Similarly, MIC displayed high precision on Colon (0.96), but this was not reflected in the recall or F1-score. CA-mRMR's ability to sustain high precision across datasets points to lower false positive rates and greater reliability in real-world classifications.

The recall values further validated the generalization capacity of the CA-mRMR. It consistently reported high values, including 0.94 for Colon, 0.96 for Leukaemia, 0.99 for Lymphoma, 0.97 for DLBCL, and 0.97 for Gastric, which either matched or exceeded those of other techniques. Although MIC achieved perfect recall (1.00) for DLBCL, it was offset by lower precision, weakening its overall classification quality. CA-mRMR demonstrated the ability to avoid the precision-recall tradeoff, which is a common issue in high-dimensional classification tasks.

The F1-score, which balances precision and recall, confirmed CA-mRMR's overall classification strength. It yielded high and consistent F1-scores across all datasets: 0.94 for Colon, 0.95 for Leukaemia, 0.99 for Lymphoma, 0.97 for DLBCL, and 0.96 for Gastric. While traditional methods such as Ridge and RF produced competitive F1-scores on individual datasets, they failed to maintain performance consistency across all evaluation metrics. This highlights the limitations of classical techniques when applied to various biomarker data. The average performance of these ten methods was evaluated across various datasets, and the results are presented in Figure 8. The proposed CA-mRMR method outperformed all the classical approaches, achieving the highest accuracy (96.0%), precision (97.0%), recall (96.5%), and F-measure (95.5%). Although ImRMR and RF performed comparably, their metrics were slightly lower and less consistent, demonstrating the superior overall performance and balance of CA-mRMR across all the evaluation parameters.

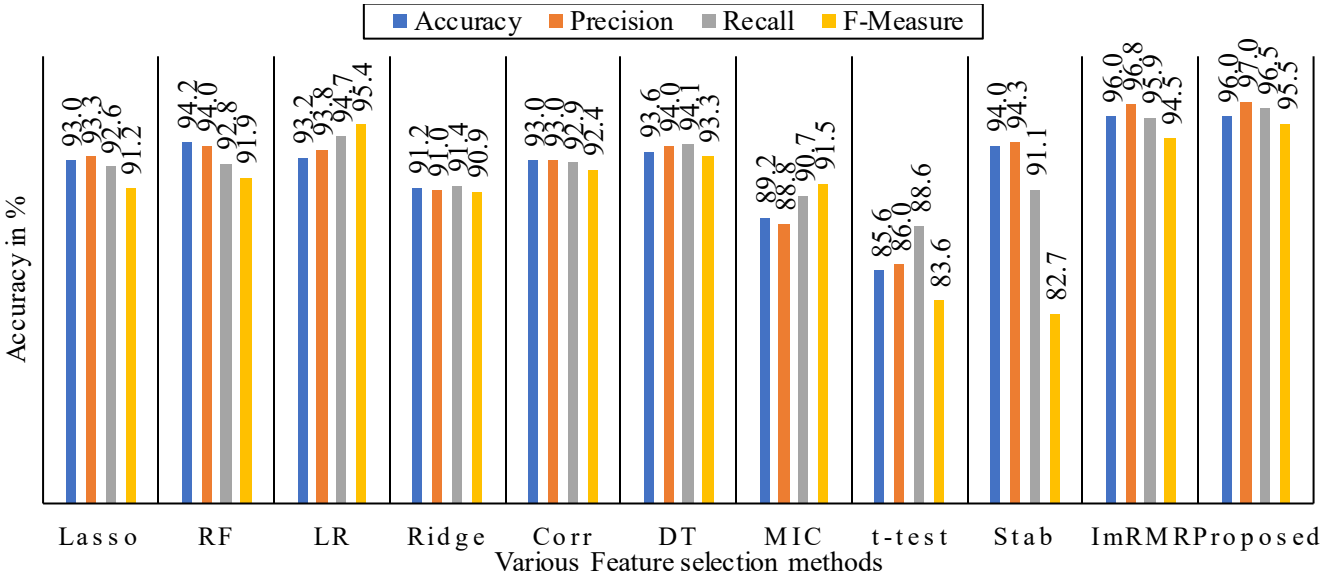

Fig. 8 Performance comparison of different feature selection models

In addition to comparing the results with the conventional feature selection models, a comparative analysis was performed with other hybrid models that employed biomarker datasets.

Table 10 presents a comparison analysis that compares the Number of selected features and classification accuracy for a range of biomarker datasets using the traditional, hybrid, and suggested CA-mRMR models.

Table 10. Comparison with hybrid models

Datasets	Methods	Acc.	Features
Lung	Peng et al. (2005) [31]	78.32	15
	Saravanan et al. (2022) [32]	89.36	32
	Proposed	95.85	24
NCI	Peng et al. (2005) [31]	72.13	13
	Saravanan et al. (2022) [32]	88.63	23
	Proposed	92.77	18
Colon	Gao et al. (2017) [51]	90.32	3
	Sun et al. (2018) [52]	84.30	5
	Lu et al. (2017) [53]	89.09	19
	Wang et al. (2017) [54]	85.70	11
	Lin et al. (2019) [55]	84.00	3
	Yu et al. (2024) [40]	93.33	4
	Peng et al. (2005) [31]	75.69	14
	Saravanan et al. (2022) [32]	90.36	19
	Proposed	94.56	19
Leukemia	Aziz et al. (2017) [56]	98.68	12
	Tumuluru et al., (2017) [57]	94.59	-
	Sun et al. (2018) [52]	92.73	3
	Lu et al. (2017) [53]	97.62	7
	Wang et al. (2017) [54]	96.10	8.3
	Lin et al. (2019) [55]	95.20	9
	Yu et al. (2024) [40]	97.23	6
	Peng et al. (2005) [31]	88.76	14
	Saravanan et al. (2022) [32]	93.46	29
	Proposed	97.11	21
Lymphoma	Vanitha et al. (2015) [58]	90.90	4
	Yu et al. (2024) [40]	97.77	5
	Peng et al. (2005) [31]	86.59	13
	Saravanan et al. (2022) [32]	94.78	18
	Proposed	98.11	16
DLBCL	Peng et al. (2005) [31]	83.57	16
	Saravanan et al. (2022) [32]	95.16	29
	Proposed	98.19	18
Gastric	Peng et al. (2005) [31]	80.71	13
	Saravanan et al. (2022) [32]	93.69	20
	Proposed	97.15	17

The results clearly show that the proposed method performed better in terms of feature reduction and classification accuracy. In particular, CA-mRMR outperformed Peng et al. (78.32% with 15 features) and Saravanan et al. (89.36% with 32 features) in the lung dataset, achieving an accuracy of 95.85% with 24 features. Similarly, the proposed approach outperformed Saravanan et al. (88.63%) and Peng et al. (72.13%) on the NCI dataset, achieving 92.77% accuracy with 18 features.

The suggested strategy accurately used a few features while demonstrating a solid balance on high-dimensional datasets, such as leukemia and colon cancer. It produced results comparable to those of Yu et al. (93.33%, four features) and Aziz et al. (98.68%, 12 features), achieving 94.56% accuracy with 19 features in the colon dataset and 97.11% accuracy with 21 features in leukemia. With 16 features, CA-mRMR achieved 98.11% in the lymphoma dataset, which is among the highest recorded accuracies. The DLBCL and Gastric datasets exhibited comparable patterns. Therefore, the CA-mRMR approach proved its efficacy in managing high-dimensional biomarker data by exhibiting exceptional classification accuracy while preserving a small feature set. Its reliable performance on many datasets attests to its efficiency, generalizability, and usefulness in real-world big-data applications, especially in the biomedical field.

The proposed CA-mRMR consistently outperformed or matched the classical methods across all key evaluation metrics. Its robustness, efficiency, and generalizability in feature selection underscore its suitability for biomarker data classification tasks, offering both theoretical value and practical impact. These findings significantly strengthen the contribution of this study and support its relevance to real-world bioinformatic applications.

9. Research Implications

9.1. Theoretical and Practical Implications

According to the experimental and result analyses, the proposed CA-mRMR feature selection technique showed consistent performance with a variety of classifiers, including SVM, NB, KNN, LR, RF, and DT, on both large-scale and biomarker datasets. The method's domain adaptability and classifier independence underscore its potential for wider applications in fields such as natural language processing, cybersecurity threat identification, financial forecasting, and environmental modelling. The resilience and scalability of the method are confirmed by the consistent feature count decrease across classifiers without compromising the classification quality.

There are numerous real-world biomedical and data-intensive applications in which the proposed CA-mRMR feature selection method is highly beneficial. CA-mRMR can improve the accuracy of disease classification models by identifying the most pertinent biomarkers using high-dimensional gene expression data. This facilitates early diagnosis and individualized treatment planning by enabling more accurate detection of diseases, including cancer subtypes (such as leukemia and lymphoma). The use of CA-mRMR extends beyond the medical field to drug discovery, where the prediction of drug-target interactions depends on the selection of pertinent molecular descriptors from enormous chemical datasets. This technique improves model interpretability in bioinformatics by removing redundant or noisy features, which helps in pathway analysis and gene function prediction.

In real-time data analysis in sensor networks and the Internet of Things, where dimensionality reduction is essential for rapid decision making, CA-mRMR is especially appropriate owing to its stability across various classifiers and datasets. Moreover, it is useful for other real-time applications

where accuracy and speed are essential, such as fraud detection, financial risk analysis, and cybersecurity monitoring, owing to its reliable performance and minimal processing overhead. These uses demonstrate the method's versatility and applicability to fields that demand effective handling of high-dimensional data, thereby enhancing its usefulness in both scholarly and real-time practical applications.

9.2. Challenges and Future Work

Although the proposed CA-mRMR approach has shown progress, a number of more general issues in data science still need to be addressed. Managing extremely large and high-dimensional datasets, particularly those produced in real-time or streaming environments, is a major challenge. Effective, flexible, and scalable feature selection methods are required to ensure prompt and precise decision-making in these situations.

Another major issue is ensuring the model results are transparent and interpretable, particularly in sensitive fields like healthcare and finance. Although CA-mRMR offers a condensed and relevant subset of features, building confidence and actionable knowledge requires additional Integration with Explainable AI (XAI) techniques. Another difficulty is the increasing complexity of data, such as multimodal data, which consists of text, pictures, and sensor inputs. Future studies must concentrate on creating unified frameworks that can handle various types of data while preserving computing efficiency and performance.

Furthermore, ethical issues, including privacy protection, equity, and data bias, are becoming increasingly significant. It will be essential for ethical data science approaches to include privacy-preserving techniques, such as differential privacy or federated learning, as well as fairness-aware feature selection. Finally, there are major issues regarding generalizability and repeatability. Standardized benchmarking across diverse datasets, workloads, and contexts is necessary to validate the suggested models and techniques consistently. In addition to improving the applicability of feature selection algorithms, such as CA-mRMR, addressing these issues will help create data science solutions that are reliable, moral, and prepared for the future.

Future studies should focus on applying CA-mRMR to multimodal datasets, such as multi-omics or sensor fusion data, expanding its use for unsupervised and semi-supervised learning tasks, and incorporating it into deep learning pipelines for improved feature interpretability. Comparative research incorporating distributed or federated learning contexts may potentially assess the effectiveness of CA-mRMRs in settings with limited resources and privacy. Such investigations would further demonstrate the generalizability and practical applicability of the proposed approach in real-world, high-dimensional data contexts.

10. Conclusion

A framework for quality enhancement was presented in this research to choose a meaningful feature subset that represents the complete dataset to enhance the quality of large datasets. By eliminating redundant and irrelevant attributes, the suggested conjoint analysis with the minimum Redundancy Maximum Relevance (mRMR) approach was used in the map phase to identify the most important attributes.

In the reduce phase, the q-gram-based filtering approach is used to identify pertinent instances by eliminating redundant and irrelevant instances from the big data. The dataset is partitioned using the Apache Spark environment to increase the effectiveness of the proposed model. The performance of the proposed framework was experimentally examined using three datasets and evaluated across different models currently in use. According to research, the suggested model improves the quality of large data by reducing the Number of features and instances while maintaining classification accuracy and reducing execution time. The model outperformed other existing models, with speed-up rates ranging from 4.77 to 1.19. In addition, compared to the other models used for comparison, the accuracy percentage increased from 34.72% to 3.66%. The proposed CA-mRMR approach has certain drawbacks despite its excellent performance across a variety of datasets and classifiers. High-dimensional datasets are the major focus of experimental assessment, which may restrict direct application to other domains, such as text or images, without further adaptation. Furthermore, depending on task-specific constraints, dynamic feature selection may be necessary in real-world circumstances, even though the Number of chosen features is kept constant for a fair comparison. Moreover, only the accuracy and execution time were considered when assessing the performance of the proposed model. Future studies should use other deep evaluations to validate its efficiency further.

The approach was further verified using a variety of classifiers, including SVM, NB, KNN, LR, RF, and DT, on the large-scale and heterogeneous datasets Epsilon, Susy, and ECBDL14 to improve external validity. Compared to mRMR, MR-mRMR, EmRMR, ImRMR, and FmRMR, CA-mRMR consistently produced smaller feature subset sizes across all classifiers, indicating effectiveness and generalizability in crucial fields such as medicine. This cross-domain validation demonstrates the broader applicability of the method and strengthens its resilience across a range of data distributions and classification scenarios. Integration with deep learning pipelines and adaptive feature selection techniques that dynamically respond to dataset properties will be the subject of future research. Assessing external validation on real-time datasets is also necessary to evaluate generalization and scalability. Future research will examine the proposed framework using a wider variety of datasets and propose a classification method that is more appropriate for large datasets.

References

[1] Hamed Ghorban Tanhaei et al., "Predictive Analytics in Customer Behavior: Anticipating Trends and Preferences," *Results in Control and Optimization*, vol. 17, pp. 1-17, 2024. [CrossRef] [Google Scholar] [Publisher Link]

[2] Arunraj Gopalsamy, and B. Radha, "Feature Selection Using Multiple Ranks with Majority Vote-Based Relative Aggregate Scoring Model for Parkinson Dataset," *Proceedings of International Conference on Data Science and Applications: ICDSA 2021*, vol. 2, pp. 1-19, 2021. [CrossRef] [Google Scholar] [Publisher Link]

[3] Khaledun Nahar et al., "Mining Educational Data to Predict Students Performance: A Comparative Study of Data Mining Techniques," *Education and Information Technologies*, vol. 26, no. 5, pp. 6051-6067, 2021. [CrossRef] [Google Scholar] [Publisher Link]

[4] C.V. Swetha, Sibi Shaji, and B. Meenakshi Sundaram, "Feature Selection Using Chi-Squared Feature-Class Association Model for Fake Profile Detection in Online Social Networks," *International Conference on Advanced Computing and Intelligent Technologies*, Imphal, India, pp. 259-276, 2023. [CrossRef] [Google Scholar] [Publisher Link]

[5] Saravanan Arumugam, "An Effective Hybrid Encryption Model Using Biometric Key for Ensuring Data Security," *International Arab Journal Information Technology*, vol. 20, no. 5, pp. 796-807, 2023. [CrossRef] [Google Scholar] [Publisher Link]

[6] Zhiying Fan, "E-Commerce Data Mining Analysis Based on User Preferences and Association Rules," *Scalable Computing: Practice and Experience*, vol. 25, no. 3, pp. 1765-1772, 2024. [CrossRef] [Google Scholar] [Publisher Link]

[7] Sonika Gupta, and Sushil Kumar Mehta, "Data Mining-Based Financial Statement Fraud Detection: Systematic Literature Review and Meta-Analysis to Estimate Data Sample Mapping of Fraudulent Companies Against Non-Fraudulent Companies," *Global Business Review*, vol. 25, no. 5, pp. 1290-1313, 2024. [CrossRef] [Google Scholar] [Publisher Link]

[8] K. Vani, and S.P. Swornambiga, "Adaptive Intrusion Detection Framework for Enhanced Cloud Security in Fog and Edge Computing Environments," *International Journal of Advanced Technology and Engineering Exploration*, vol. 11, no. 121, pp. 1613-1640, 2024. [CrossRef] [Google Scholar] [Publisher Link]

[9] Liuchao Jin et al., "Big Data, Machine Learning, and Digital Twin Assisted Additive Manufacturing: A Review," *Materials & Design*, vol. 244, pp. 1-53, 2024. [CrossRef] [Google Scholar] [Publisher Link]

[10] Stephen Kaisler et al., "Big Data: Issues and Challenges Moving Forward," *2013 46th Hawaii International Conference on System Sciences*, Wailea, HI, USA, pp. 995-1004, 2013. [CrossRef] [Google Scholar] [Publisher Link]

[11] Mallikarjuna Paramesha, Nitin Liladhar Rane, and Jayesh Rane, "Big Data Analytics, Artificial Intelligence, Machine Learning, Internet of Things, and Blockchain for Enhanced Business Intelligence," *Partners Universal Multidisciplinary Research Journal (PUMRJ)*, vol. 2, no. 3, pp. 110-133, 2024. [CrossRef] [Google Scholar] [Publisher Link]

[12] Mirza Golam Kibria et al., "Big Data Analytics, Machine Learning, and Artificial Intelligence in Next-Generation Wireless Networks," *IEEE Access*, vol. 6, pp. 32328-32338, 2018. [CrossRef] [Google Scholar] [Publisher Link]

[13] Nikolaos Stylos, and Jeremy Zwiegelaar, *Big Data as a Game Changer: How does it Shape Business Intelligence within a Tourism and Hospitality Industry Context?*, Big Data and Innovation in Tourism, Travel, and Hospitality, Springer, Singapore, pp. 163-181, 2019. [CrossRef] [Google Scholar] [Publisher Link]

[14] Dirk Hölscher et al., "A Big Data Quality Preprocessing and Domain Analysis Provisioner Framework Using Cloud Infrastructures," *ALLDATA 2018: The 4th International Conference on Big Data, Small Data, Linked Data and Open Data*, Athens, Greece, pp. 53-58, 2018. [Google Scholar] [Publisher Link]

[15] Ikbal Taleb, and Mohamed Adel Serhani, "Big Data Pre-Processing: Closing the Data Quality Enforcement Loop," *2017 IEEE International Congress on Big Data (BigData Congress)*, Honolulu, HI, USA, pp. 498-501, 2017. [CrossRef] [Google Scholar] [Publisher Link]

[16] Katherine Rucinski et al., "Challenges and Opportunities in Big Data Science to Address Health Inequities and Focus the HIV Response," *Current HIV/AIDS Reports*, vol. 21, no. 4, pp. 208-219, 2024. [CrossRef] [Google Scholar] [Publisher Link]

[17] Noura AlNuaimi et al., "Streaming Feature Selection Algorithms for Big Data: A Survey," *Applied Computing and Informatics*, vol. 18, no. 1/2, pp. 113-135, 2020. [CrossRef] [Google Scholar] [Publisher Link]

[18] Haowen Guan et al., "SLOF: Identify Density-Based Local Outliers in Big Data," *2015 12th Web Information System and Application Conference (WISA)*, Jinan, China, pp. 61-66, 2015. [CrossRef] [Google Scholar] [Publisher Link]

[19] Jundong Li, and Huan Liu, "Challenges of Feature Selection for Big Data Analytics," *IEEE Intelligent Systems*, vol. 32, no. 2, pp. 9-15, 2017. [CrossRef] [Google Scholar] [Publisher Link]

[20] N.N. Misra et al., "IoT, Big Data and Artificial Intelligence in Agriculture and Food Industry," *IEEE Internet of Things Journal*, vol. 9, no. 9, pp. 6305-6324, 2020. [CrossRef] [Google Scholar] [Publisher Link]

[21] Fakhitah Ridzuan, and Wan Mohd Nazmee Wan Zainon, "A Review on Data Quality Dimensions for Big Data," *Procedia Computer Science*, vol. 234, pp. 341-348, 2024. [CrossRef] [Google Scholar] [Publisher Link]

[22] Jingran Wang et al., "Overview of Data Quality: Examining the Dimensions, Antecedents, and Impacts of Data Quality," *Journal of the Knowledge Economy*, vol. 15, no. 1, pp. 1159-1178, 2024. [CrossRef] [Google Scholar] [Publisher Link]

[23] Adebunmi Okechukwu Adewusi et al., "Business Intelligence in the Era of Big Data: A Review of Analytical Tools and Competitive Advantage," *Computer Science & IT Research Journal*, vol. 5, no. 2, pp. 415-431, 2024. [CrossRef] [Google Scholar] [Publisher Link]

[24] Yazeed Alkatheeri et al., "The Mediation Effect of Management Information Systems on the Relationship between Big Data Quality and Decision making Quality," *Test Engineering and Management*, pp. 12065-12074, 2020. [Google Scholar]

[25] Anandhi Ramasamy, and Soumitra Chowdhury, "Big Data Quality Dimensions: A Systematic Literature Review," *JISTEM-Journal of Information Systems and Technology Management*, vol. 17, pp. 1-13, 2020. [CrossRef] [Google Scholar] [Publisher Link]

[26] Agung Wahyudi, George Kuk, and Marijn Janssen, "A Process Pattern Model for Tackling and Improving Big Data Quality," *Information Systems Frontiers*, vol. 20, no. 3, pp. 457-69, 2018. [CrossRef] [Google Scholar] [Publisher Link]

[27] R. Joseph Manoj, M.D. Anto Praveena, and K. Vijayakumar, "An ACO-ANN Based Feature Selection Algorithm for Big Data," *Cluster Computing*, vol. 22, no. 2, pp. 3953-3960, 2019. [CrossRef] [Google Scholar] [Publisher Link]

[28] Noha Shehab, Mahmoud Badawy, and H. Arafat Ali, "Toward Feature Selection in Big Data Preprocessing Based on Hybrid Cloud-Based Model," *The Journal of Supercomputing*, vol. 78, no. 3, pp. 3226-3265, 2021. [CrossRef] [Google Scholar] [Publisher Link]

[29] Ibrahim M. El-Hasnony et al., "Improved Feature Selection Model for Big Data Analytics," *IEEE Access*, vol. 8, pp. 66989-67004, 2020. [CrossRef] [Google Scholar] [Publisher Link]

[30] Ioannis Tsamardinos et al., "A Greedy Feature Selection Algorithm for Big Data of High Dimensionality," *Machine Learning*, vol. 108, no. 2, pp. 149-202, 2019. [CrossRef] [Google Scholar] [Publisher Link]

[31] Hanchuan Peng, Fuhui Long, and C. Ding, "Feature Selection based on Mutual Information Criteria of Max-Dependency, Max-Relevance, and Min-Redundancy," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 27, no. 8, pp. 1226-1238, 2005. [CrossRef] [Google Scholar] [Publisher Link]

[32] A. Saravanan, C. Stanly Felix, and M. Umarani, "Maximum Relevancy and Minimum Redundancy Based Ensemble Feature Selection Model for Effective Classification," *Advanced Computing and Intelligent Technologies: Proceedings of ICACIT 2022*, Singapore, pp. 131-146. 2022. [CrossRef] [Google Scholar] [Publisher Link]

[33] Blessy Trencia Lincy S.S., and Suresh Kumar Nagarajan, "MR-mRMR Feature Selection Approach with an Incremental Classifier Model in Big data," *International Journal of Pharmaceutical Research*, vol. 10, no. 4, pp. 365- 379, 2018. [Google Scholar] [Publisher Link]

[34] Blessy Trencia Lincy S.S., and Suresh Kumar Nagarajan, "An Enhanced Pre-Processing Model for Big Data Processing: A Quality Framework," *2017 International Conference on Innovations in Green Energy and Healthcare Technologies*, Coimbatore, India, pp. 1-7, 2017. [CrossRef] [Google Scholar] [Publisher Link]

[35] Thee Zin Win, and Nang Saing Moon Kham, "Mutual Information-Based Feature Selection Approach to Reduce High Dimension of Big Data," *MLMI '18: Proceedings of the International Conference on Machine Learning and Machine Intelligence*, Ha Noi Viet Nam, pp. 3-7, 2018. [CrossRef] [Google Scholar] [Publisher Link]

[36] Vinaya Keskar, Jyoti Yadav, and Ajay Kumar, "Perspective of Anomaly Detection in Big Data for Data Quality Improvement," *Materials Today: Proceedings*, vol. 51, pp. 532-537, 2021. [CrossRef] [Google Scholar] [Publisher Link]

[37] Ka Yee Wong, and Raymond K. Wong, "Big Data Quality Prediction Informed by Banking Regulation," *International Journal of Data Science and Analytics*, vol. 12, no. 2, pp. 147-164, 2021. [CrossRef] [Google Scholar] [Publisher Link]

[38] Sarwar Kamal et al., "A MapReduce Approach to Diminish Imbalance Parameters for Big Deoxyribonucleic Acid Dataset," *Computer Methods and Programs in Biomedicine*, vol. 131, pp. 191-206, 2016. [CrossRef] [Google Scholar] [Publisher Link]

[39] Bibhuprasad Sahu et al., "Novel Hybrid Feature Selection using Binary Portia Spider Optimization Algorithm and Fast mRMR," *Bioengineering*, vol. 12, no. 3, pp. 1-26, 2025. [CrossRef] [Google Scholar] [Publisher Link]

[40] Kun Yu et al., "A Hybrid Feature-Selection Method Based on mRMR and Binary Differential Evolution for Gene Selection," *Processes*, vol. 12, no. 2, pp. 1-21, 2024. [CrossRef] [Google Scholar] [Publisher Link]

[41] Ziqiang Ye et al., "Identification of OSAHS Patients based on ReliefF-mRMR Feature Selection," *Physical and Engineering Sciences in Medicine*, vol. 47, no. 1, pp. 99-108, 2024. [CrossRef] [Google Scholar] [Publisher Link]

[42] Kuraganty Phani Rama Krishna, and Ramakrishna Thirumuru, "A Balanced Intrusion Detection System for Wireless Sensor Networks in a Big Data Environment using CNN-SVM Model," *Informatics and Automation*, vol. 22, no. 6, pp. 1296-1322, 2023. [CrossRef] [Google Scholar] [Publisher Link]

[43] Osama Mohareb Khaled et al., "Evaluating Machine Learning Models for Predictive Analytics of Liver Disease Detection using Healthcare Big Data," *International Journal of Electrical and Computer Engineering (IJECE)*, vol. 15, no. 1, pp. 1162-1174, 2025. [CrossRef] [Google Scholar] [Publisher Link]

[44] Jintong Yang, Yiling Guo, and Xinling Cai, "Wildlife Development Prediction Based on Big Data and Bayesian Logistic Regression," *2024 2nd International Conference on Mechatronics, IoT and Industrial Informatics (ICMIII)*, Melbourne, Australia, pp. 419-423, 2024. [CrossRef] [Google Scholar] [Publisher Link]

[45] Hitham Al-Manaseer et al., *A Novel Big Data Classification Technique for Healthcare Application using Support Vector Machine, Random Forest and J48*, Classification Applications with Deep Learning and Machine Learning Technologies, pp. 205-215, 2022. [CrossRef] [Google Scholar] [Publisher Link]

[46] Salvador García et al., "Big Data Preprocessing: Methods and Prospects," *Big Data Analytics*, vol. 1, no. 1, pp. 1-22, 2016. [CrossRef] [Google Scholar] [Publisher Link]

[47] Keith D. Foote, Big Data Integration 101: The What, Why and How, Dataversity, 2019. [Online]. Available: https://www.dataversity.net/big-data-integration-101-the-what-why-and-how/

[48] Fakhitah Ridzuan, and Wan Mohd Nazmee Wan Zainon, "A Review on Data Cleansing Methods for Big Data," *Procedia Computer Science*, vol. 161, pp. 731-738, 2019. [CrossRef] [Google Scholar] [Publisher Link]

[49] Sebastián Maldonado, Ricardo Montoya, and Julio López, "Embedded Heterogeneous Feature Selection for Conjoint Analysis: A SVM Approach using L1 Penalty," *Applied Intelligence*, vol. 46, no. 4, pp. 775-787, 2017. [CrossRef] [Google Scholar] [Publisher Link]

[50] Stefan Burkhardt et al., "Q-Gram Based Database Searching using a Suffix Array (QUASAR)," *RECOMB '99: Proceedings of the Third Annual International Conference on Computational Molecular Biology*, Lyon, France, pp. 77-83, 1999. [CrossRef] [Google Scholar] [Publisher Link]

[51] Lingyun Gao et al., "Hybrid Method Based on Information Gain and Support Vector Machine for Gene Selection in Cancer Classification," *Genomics, Proteomics & Bioinformatics*, vol. 15, no. 6, pp. 389-395, 2017. [CrossRef] [Google Scholar] [Publisher Link]

[52] Lin Sun et al., "Joint Neighborhood Entropy-Based Gene Selection Method with Fisher Score for Tumor Classification," *Applied Intelligence*, vol. 49, no. 4, pp. 1245-1259, 2018. [CrossRef] [Google Scholar] [Publisher Link]

[53] Lu Huijuan et al., "A Hybrid Feature Selection Algorithm for Gene Expression Data Classification," *Neurocomputing*, vol. 256, pp. 56-62, 2017. [CrossRef] [Google Scholar] [Publisher Link]

[54] Aiguo Wang et al., "Wrapper-Based Gene Selection with Markov Blanket," *Computers in Biology and Medicine*, vol. 81, pp. 11-23, 2017. [CrossRef] [Google Scholar] [Publisher Link]

[55] Lin Sun et al., "Feature Selection Using Neighborhood Entropy-Based Uncertainty Measures for Gene Expression Data Classification," *Information Sciences*, vol. 502, pp. 18-41, 2019. [CrossRef] [Google Scholar] [Publisher Link]

[56] Rabia Aziz, C.K. Verma, and Namita Srivastava, "A Novel Approach for Dimension Reduction of Microarray," *Computational Biology and Chemistry*, vol. 71, pp. 161-169, 2017. [CrossRef] [Google Scholar] [Publisher Link]

[57] Praveen Tumuluru, and Bhramaramba Ravi, "GOA-Based DBN: Grasshopper Optimization Algorithm-Based Deep Belief Neural Networks for Cancer Classification," *International Journal of Applied Engineering Research*, vol. 12, no. 24, pp. 14218-14231, 2017. [Google Scholar] [Publisher Link]

[58] C. Devi Arockia Vanitha, D. Devaraj, and M. Venkatesulu, "Gene Expression Data Classification Using Support Vector Machine and Mutual Information-Based Gene Selection," *Procedia Computer Science*, vol. 47, pp. 13-21, 2015. [CrossRef] [Google Scholar] [Publisher Link]