SSRG International Journal of Electronics and Communication Engineering
ISSN: 2348-8549/ https://doi.org/10.14445/23488549/IJECE-V11I3P105

Volume 11 Issue 3, 41-52, March 2024
© 2024 Seventh Sense Research Group®

Original Article

Optimizing Prediction Accuracy in High-Dimensional Data: Comparative Analysis of Feature Selection Methods with Naive Bayes Algorithm

T. Rajendran[1], C. Balakrishnan[2], B. Yamini[3], CH. Srilakshmi[4], B. Maheswari[5], M. Nalini[6], R. Siva Subramanian[7]

[1]Department Computer Science and Engineering, Rajalakshmi Institute of Technology, Tamilnadu, India.
[2,6]Department of Computer Science and Engineering, S.A.Engineering College, Tamilnadu, India.
[3]Department of Networking and Communications, School of Computing, College of Engineering and Technology, SRM Institute of Science and Technology (SRMIST), Tamilnadu, India.
[4]Department of Computer Science and Business Systems, R.M.D. Engineering College, Tamilnadu, India.
[5]Department of Computer Science and Engineering, R.M.K. Engineering College, Tamilnadu, India.
[7]Department of Computer Science and Engineering, R.M.K. College of Engineering and Technology, Tamilnadu, India.

[6]Corresponding Author : nalini.tptwin@gmail.com

Received: 05 January 2024 Revised: 07 February 2024 Accepted: 05 March 2024 Published: 31 March 2024

***Abstract** - This high-dimensional data is becoming increasingly common in various sectors, such as the social sciences, biology, medicine, and finance. It is defined by datasets that include many characteristics or dimensions. This study explores the idea of high-dimensional data, the difficulties it presents, and how it affects prediction results. Developing strategies to extract useful information from high-dimensional data is crucial due to its many issues, including the curse of dimensionality, increased processing complexity, and poor interpretability. This article presents research that employs feature selection techniques to solve problems related to high-dimensional data. The crucial process of feature selection is determining which features are most relevant and keeping them while eliminating those that are unnecessary or redundant. This method seeks to increase prediction accuracy, decrease overfitting, and improve model interpretability by lowering the dimensionality of the data. The present study examines three distinct feature selection methods, including Filter (SU & RELIEF), Wrapper (GENETIC & SFS), and Hybrid (combining filter & wrapper), to choose relevant features from high-dimensional data. We use a real-world high-dimensional customer dataset from UCI to illustrate how well our suggested feature selection methods work with the Naive Bayes machine learning algorithm. We demonstrate how feature selection strategies, both with and without feature selection, lead to improved prediction outcomes in high-dimensional data settings via a number of experiments and evaluations. The results show that using feature selection enhances the accuracy of the predictions. In contrast to filter and wrapper techniques, hybrid FS selects the best feature set from the three FS models. Researchers and practitioners working with high-dimensional data may make better decisions using these insights, eventually improving the prediction models' quality and applicability.*

***Keywords** - Feature Selection, High dimensional data, Machine Learning, Naive Bayes, SFS.*

1. Introduction

High-dimensional data is becoming a common and severe obstacle in data-driven research and applications. Datasets with many attributes or dimensions are called high-dimensional data [1]. Numerous sources, including genetic markers in genomics, sensory data in image processing, and socioeconomic characteristics in social sciences, might give conception to these dimensions.

High-dimensional data presents a number of challenges and complications that need to be carefully considered, in addition to the possibility of revealing rich patterns and insights [2]. High-dimensional data presents a variety of difficulties. A primary obstacle is the "curse of dimensionality." Data points in the feature space are sparser as the number of dimensions rises, making it more challenging to assess significant correlations and patterns precisely. Because of this sparsity, computing becomes more complex, requiring more memory and processing power.

Furthermore, the sheer nature of high-dimensional data often makes overfitting a phenomenon in which models are too customized to the noise in the data. This problem undermines the models' generalisation capacity to novel, unforeseen cases. Furthermore, the large dimensionality of data impedes its interpretability, which makes it difficult for analysts to recognize important elements and understand the underlying structure. High-dimensional data has significant negative implications on prediction results. If appropriate techniques are not used, models trained on high-dimensional data may exhibit worse generalization and prediction accuracy. Making informed decisions and gaining valuable insights from data are the main goals of predictive modelling and data analysis, which are undercut by this.

Several methods and approaches have been proposed to tackle high-dimensional data difficulties. Among the most important approaches to solving this issue is feature selection. A crucial phase in data preprocessing is feature selection, which entails identifying and keeping the most pertinent attributes while removing redundant or superfluous ones. Feature selection aims to increase model interpretability, decrease overfitting, and improve prediction accuracy by lowering the dimensionality of the data.

There are several feature selection approach categories, each with unique advantages and uses. Among these categories are:

1. Filter Methods: These techniques assess the importance of features apart from the selected prediction model. Mutual information-based selection and correlation-based feature selection are two popular filter techniques.
2. Wrapper Techniques: To determine the significance of characteristics, wrapper techniques use a particular prediction model. They include continually training and assessing the model using several feature subsets to find the ideal feature set. Forward selection and Recursive Feature Elimination (RFE) are two examples of wrapper techniques.
3. Hybrid Techniques: These techniques include the best features of both wrapper and filter techniques. These methods include selecting features using a filter and then optimizing the chosen features using a wrapper.

A real-world dataset obtained from the UCI repository is used in this study to demonstrate the effectiveness of feature selection while dealing with high-dimensional data. Three distinct feature selection techniques are taken into consideration here: Hybrid (combining filter and wrapper techniques), wrapper (combining Genetic and SFS), and filter (combining SU and RELIEF) techniques.

To analyse the data, the features selected are connected using the Naïve Bayes machine learning method. This thorough examination aims to show how feature selection may be used to enhance computing efficiency, interpretability of models, and prediction results.

In a broader sense, several difficulties are associated with high-dimensional data, ranging from the dimensionality curse to a decline in prediction accuracy. Because feature selection approaches improve predictive modelling and reduce dimensionality, they effectively address these problems. To provide valuable insights for researchers and practitioners working in high-dimensional data scenarios, this study explores the domain of high-dimensional data, investigates different feature selection techniques, and assesses their effects on predictive modelling within an actual dataset.

1.1. Motivation

This research paper’s inspiration stems from the ubiquity of high-dimensional data in several disciplines. Rich in characteristics or dimensions, high-dimensional data can provide important information and insights that may guide well-informed decision-making.

But it also presents several difficult obstacles that call for creative solutions. Researchers and practitioners are facing more and more information of this kind; thus, it is critical to comprehend the complexities of high-dimensional data and create efficient methods to use its potential fully.

The primary purpose is to take care of the following essential factors:

1. The problem of high-dimensional information to fully comprehend and express the difficulties brought forth by high-dimensional data, such as the dimensionality curse, higher processing requirements, and decreased interpretability. It is essential to acknowledge these obstacles to create reliable data analysis and predictive modelling methods.
2. Effect on the Results of Predictions: To clarify how high-dimensional data influence prediction results. Investigating the impact of several characteristics on predictive models’ accuracy and generalization performance is crucial, which is why these results may not be ideal in some situations.
3. This paper emphasises feature selection as a robust approach to tackle these concerns and improve the overall quality of predictive models.

1.2. Objectives

1. Analysis of the Impact on Predictive Modelling: Assess the impact of high-dimensional data on the outcomes of predictions. This encompasses an empirical examination of the repercussions of handling high-dimensional data, including diminished prediction accuracy and interpretation challenges.
2. Present Proposed Methods for Feature Selection: To propose and evaluate feature selection as a feasible resolution to the difficulties presented by high-dimensional data. The study will investigate various classifications of feature selection techniques, such as wrapper, filter, and hybrid methodologies.
3. Illustrate the Benefits of Feature Selection when Applied to High-Dimensional Data: Perform experiments and evaluations utilizing a real-world dataset to demonstrate the efficacy of feature selection. A diverse range of

machine learning algorithms will illustrate the enhancements in model performance, interpretability, and computational efficiency attained via feature selection.

4. Provide Valuable Insights: To provide researchers and practitioners with guidance and insights regarding efficiently navigating the intricacies associated with high-dimensional data. Our goal is to enable individuals working with these datasets to improve the quality and applicability of their predictive models and make more informed decisions.

By examining these rationales and goals, this scholarly article strives to contribute to the dynamic domain of high-dimensional data analysis and offer pragmatic resolutions to its obstacles. The subsequent sections are structured as follows: 2. Survey of the Literature, 3. Methodology, 4. Selection of Features, 5. ML Model, 6. Experimental Findings, and 7. Conclusion.

2. Literature Survey

[3], Feature selection is an important feature of machine learning, especially in domains such as bioinformatics. Filter techniques are essential for feature selection since they may considerably cut run time and forecast accuracy. This research compared the effectiveness of 22 filter techniques with classification methods on 16 high-dimensional classification data sets.

It determined that no one filter approach consistently outperforms all others, but it gave suggestions for those performing well on various data sets. The R machine learning package mlr was utilised for this study since it provides a consistent programming API for feature selection utilising filter techniques.

[4], Digitization has created large amounts of data in various industries, including healthcare, manufacturing, commerce, IoT devices, and organisations. Machine learning algorithms are employed to find trends in this data and make recommendations for doctors and executives. However, not all dataset properties are useful for training algorithms.

LDA and PCA are two dimensionality reduction approaches investigated in this work on four standard ML algorithms: DT, SVM, NB, and RF. The results reveal that PCA beats LDA in all metrics, whereas classifier performance is unaffected. Experimentation is also being done on datasets from Diabetic Retinopathy (DR) and IDS.

[5], Because of the intricacy of high-dimensional data, the dimensionality issue in healthcare is a significant obstacle. Dimension reduction methods are increasingly in demand to enhance data prediction, analysis, and visualization. Among these methods are feature extraction and feature selection. This evaluation compares several feature extraction and feature selection strategies for dealing with data loss.

Case studies validate improved methodologies, and the review study attempts to aid researchers in selecting the most effective strategy for satisfying high-dimensional data analysis.

[6], FS is an important part of data categorization to reduce the features required to maximize accuracy and save costs. Over-fitting has become a worry with the emergence of high-dimensional datasets and limited sample counts. Various approaches for picking the optimum subset of characteristics have been presented; however, they have encountered challenges such as instability, long convergence times, and selecting a semi-optimal solution.

This work presents a hybrid technique based on the Isar method and the SFLA to identify valuable features in large-scale gene datasets. The algorithm is divided into two phases: filtering and wrapping, with the Relief technique utilized for feature weighting. The experimental findings demonstrate that the suggested strategy produces fewer characteristics while maintaining good accuracy.

[7], This paper explores feature selection in machine learning and data mining, concentrating on its usefulness in eliminating unnecessary and redundant data, lowering computation time, boosting learning accuracy, and improving learning model comprehension.

It goes over different assessment measures, supervised, unsupervised, and semi-supervised feature selection approaches, and how they are used in classification and clustering. The research also addresses future feature selection issues.

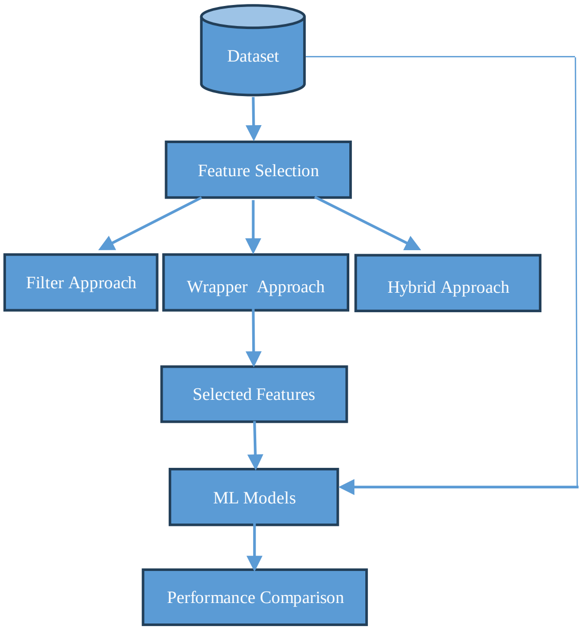

Fig. 1 Overall methodology

3. Methodology

This research paper's methodology is based on a systematic approach, including data collection, feature selection, machine learning model deployment, and evaluation and analysis.

This study aims to examine and illustrate how feature selection methods affect high-dimensional data and how they affect prediction results. The overall methodology is shown in Figure 1.

3.1. Pseudocode for the Proposed Methodology

3.1.1. Data Collection and Preprocessing

```
Dataset = LoadHighDimensionalData()  # Load a high-dimensional dataset
Cleaned Data = DataPreprocessing(dataset) # Handle missing values, outliers, and data consistency
```

3.1.2. Feature Selection

```
selectedFeaturesFilter= FilterFeatureSelection (CleanedData) # Apply filter(SU & ReliefF) FS
selectedFeaturesWrapper=WrapperFeatureSelection (CleanedData) # Apply wrapper(Genetic & SFS) FS
selectedFeaturesHybrid=HybridFeatureSelection (CleanedData) # Apply hybrid feature selection
```

3.1.3. Machine Learning Model Selection and Implementation

```
models = [NaiveBayes]  # machine learning algorithm for the model in models:

# Train and evaluate models with full set of features full
Model = Train Model(model, Cleaned Data)
full Model Performance = Evaluate Model (full Model, Cleaned Data)  Train and evaluate models with selected
features Filter Model = Train Model (model, selected Features Filter) Wrapper model = Train Model (model,
selected Features Wrapper) Hybrid Model = Train Model (model, selected Features Hybrid)
Filter Model Performance = Evaluate model (filter Model, selected Features Filter
Wrapper Model Performance = Evaluate model (wrapper Model, selected Features Wrapper)
 Hybrid Model Performance = Evaluate model (hybrid model, selected Features Hybrid)
```

3.1.4. Evaluation and Analysis

```
Performance Comparisons()  # Visualize performance differences across models and feature selection methods( Accuracy, Recall and specificity)
```

3.1.5. Discussion and Conclusion

```
Discuss Findings()  # Discuss the implications of feature selection on high-dimensional data
Summarize Conclusions()  # Summarize key findings and propose future research directions
```

3.2. Feature Selection

3.2.1. Overview of the Process

Using feature selection strategies is at the core of this study process. Three different strategies are considered: Hybrid, wrapper, and filter techniques. By choosing pertinent characteristics based on statistical metrics like correlation or mutual information, filter techniques are used to pre-process the data. Hybrid techniques use aspects of both filter and wrapper approaches, whereas wrapper methods use iterative feature selection inside the framework of particular machine learning algorithms.

3.2.2. Algorithms for Feature Selection

The high-dimensional dataset is subjected to applying and implementing all three feature selection techniques. The aim is to find the most relevant features that make a substantial contribution to predictive modelling.

3.3. Machine Learning Models

3.3.1. Algorithm Selection

This study uses Naïve Bayes machine learning algorithms to create prediction models. The algorithm highlights how adaptable the study results are to various modelling approaches.

3.3.2. Model Training and Testing

The features chosen using each of the three feature selection techniques and the whole collection of features are used to train the chosen machine learning algorithm. This makes it possible to compare the model's performance with and without feature selection.

3.4. Evaluation and Analysis

Measures of Performance: Performance measurements are used to evaluate the effect of feature selection. The metrics applied in the research are specificity, recall, and accuracy.

3.5. Discussion and Conclusion

The findings are discussed within the framework of the study's goals and intentions. In processing high-dimensional data, the research sheds light on the practical consequences of feature selection. The main conclusions, their ramifications, and possible future study avenues are outlined in the conclusion. In summary, the study article aims to provide a comprehensive and rigorous examination of how feature selection affects high-dimensional data, providing insightful information to practitioners and scholars who deal with these kinds of datasets.

4. Feature Selection

A critical stage in data preparation and analysis is feature selection, especially for high-dimensional datasets. Its main goal is to find and keep the most relevant characteristics (attributes or columns) from the dataset while removing superfluous or unnecessary ones.

The objectives of this method are to decrease the number of dimensions in the data, boost the interpretability of the model, reduce computing complexity, and increase predictive model performance. Three general categories may be used to classify feature selection techniques: filter, wrapper, and hybrid approaches.

4.1. Filter Feature Selection

These techniques evaluate each feature's significance apart from any particular prediction model. Using statistical metrics or tests, these techniques rank or score characteristics according to their relevance. These rankings are used to determine whether features are kept or removed.

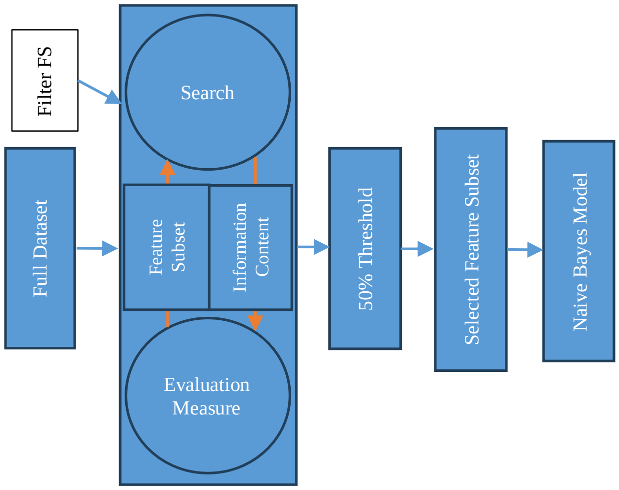

Fig. 2 Overview of filter Feature Selection

Then, the top features are evaluated based on the 50% threshold value. Because of their computing efficiency, filter techniques are often used as an initial stage in the feature selection process. In this research, Symmetrical Uncertainty and ReliefF methods are applied. Filter FS is represented in Figure 2.

4.1.1. Symmetrical Uncertainty

A measure of the amount of information shared by two random variables is called Symmetrical Uncertainty (SU). It is often used in feature selection to find the most illuminating characteristics in a dataset. It assists in identifying features with high mutual information with the target variable when used as a filter for feature selection, taking feature redundancy into account.

1. Calculate Mutual Information: Begin by figuring out how much information each characteristic and the target variable have in common. The amount of information known about one variable (feature) and the other variable (target) based on that knowledge is measured as mutual information.
2. Calculate Entropy: Determine each feature's entropy. A measure of uncertainty or disorder in a collection of values is called entropy. It provides a sense of the data required to characterize the variable.
3. Compute Symmetrical Uncertainty: Determine the Symmetrical Uncertainty for each feature using mutual information and entropy values. The following is the formula for Symmetrical Uncertainty (SU):

$$SU(X,Y)=(2*I(X,Y))/(H(X)+H(Y)) \qquad (1)$$

Where the mutual information between the target variable Y and feature X is represented by: I(X, Y)
The entropies of the target variable Y and the feature X are H(X) and H(Y), respectively.

4. Features Ranking: Sort the features according to their Symmetrical Uncertainty scores. Higher SU values indicate more informative traits.
5. Select Top Features: Assign your features to the top 50% with the most significant Symmetrical Uncertainty values [8].

4.1.2. ReliefF

ReliefF is a feature selection method that assesses the significance of features by considering their overlap and relevance to the target variable [9]. It is adequate for binary and multiclass classification tasks and was first introduced for classification issues.

1. Initialize Weights: Give each feature a starting weight of 0.
2. For Every Instance: Choose an instance randomly from the dataset.
3. Find Nearest Hit and Miss Instances: Determine which instance is closest to the hit class label and which is most relative to the miss class label.
4. Update Weights: Modify the feature weights by comparing the chosen instance's feature values to those of its closest hit-and-miss neighbours.
5. Repeat Steps 2-4: Continue Steps 2-4 until convergence or for a predetermined number of iterations.
6. Calculate Feature Scores: Determining the final score for every feature using the total weights.
7. Rank and Select Features: Choose the top 50% of features to be the subset of the final model by ranking the features according to their scores.

The concept behind ReliefF is that significant characteristics help discriminate between examples of the same class (hits) and instances of other classes (misses). The weights are modified based on the variations in feature values between the chosen instance and its neighbours. Higher weighted features are seen as more significant [10].

4.2. Wrapper Feature Selection

Feature subsets are evaluated using wrapper feature selection techniques, which use a particular machine learning algorithm as a "wrapper."

The optimum feature set is determined via an iterative process, including testing several feature subsets with the selected algorithm and the model's performance on these subsets. Because wrapper approaches need repeatedly retraining the model with distinct feature subsets, they may be computationally demanding. In this research, Symmetrical Uncertainty and Genetic and SFS methods are applied. Wrapper FS is represented in Figure 3.

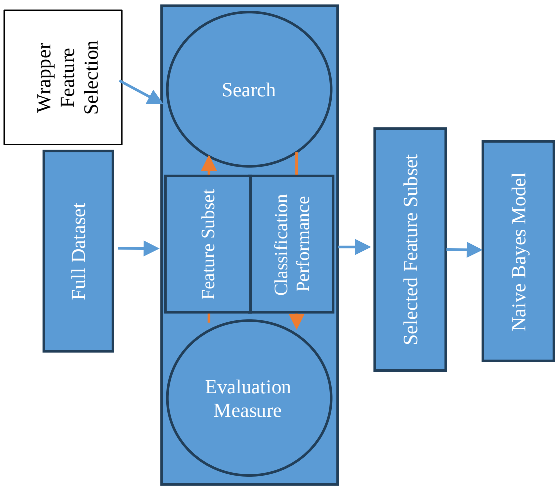

Fig. 3 Overview of wrapper Feature Selection

4.2.1. Genetic Feature Selection

Natural selection serves as the model for optimization techniques known as genetic algorithms. When applied to feature selection, genetic wrapper techniques use a genetic algorithm to find the ideal subset of features that optimises a particular machine learning model's performance [11]. Over many generations, a population of feature subsets must evolve as part of the process.

An overview of the essential operation of a Genetic Wrapper feature selection technique is provided below:

1. Initialization: Make a feature subset beginning population. This may be carried out heuristically or at random.
2. Evaluation: Analyse the population's fitness for each feature subset. Fitness usually determines how a machine learning model performs with the chosen features. Metrics like accuracy, precision, recall, and F1 score are often used.
3. Selection: Based on their fitness, choose feature subsets from the existing population. Subsets with higher performance levels are more likely to be selected.
4. Crossover (Recombination): To produce new offspring, apply crossover, also known as recombination, to pairs of chosen feature subsets. Crossover generates a new feature subset by combining features from two parents.
5. Mutation: Modify the offspring feature subsets by applying mutation. The process of mutation entails randomly altering the chosen traits in tiny ways.
6. Replacement: The combined population of parents and children should replace the previous population.
7. Termination: For a predetermined number of generations or until the convergence requirements are satisfied, repeat steps 2–6.
8. Final Subset Selection: As the final subset to be chosen, choose the feature subset with the best fitness.

To identify a feature subset that maximizes the machine learning model's performance, the genetic algorithm repeatedly develops feature subsets across generations [12].

4.2.2. Sequential Forward Selection (SFS)

A wrapper feature selection technique called Sequential Forward Selection (SFS) adds features one at a time, gradually creating a feature subset. Iteratively adding the most relevant feature at each stage until a predetermined criterion is satisfied, the procedure begins with an empty collection of features [13]. The criteria might be the best performance, the completion of a certain number of features, or any other stopping condition.

An overview of the Sequential Forward Selection algorithm is provided below:

1. Initialization: Begin with a blank slate of chosen characteristics.
2. Iteration: Consider including every feature left to the chosen features for every iteration. Utilise each potential feature and the selected set of features to assess the machine learning model's performance. Add the feature that best enhances the model's performance to the features that have been chosen.
3. Stopping Criterion: Continue the iteration until a certain number of features is achieved, no discernible performance increase is shown, or a stopping criterion is satisfied.
4. Final Subset Selection: The characteristics chosen at the halting criteria make up the final subset [14].

4.3. Hybrid Feature Selection

Hybrid feature selection techniques incorporate aspects of both filter and wrapper techniques to achieve a compromise between feature selection relevance and computational efficiency.

They usually begin with a filter-based feature selection phase to narrow down the original feature set. The final feature subset is then chosen using a wrapper-based technique applied to the filtered features. Hybrid FS is represented in Figure 4.

When dealing with high-dimensional data, these feature selection approaches and their corresponding algorithms provide useful tools for improving interpretability, optimising predictive models, and lowering computing complexity. The particulars of the dataset and the analysis's objectives determine which approach is best.

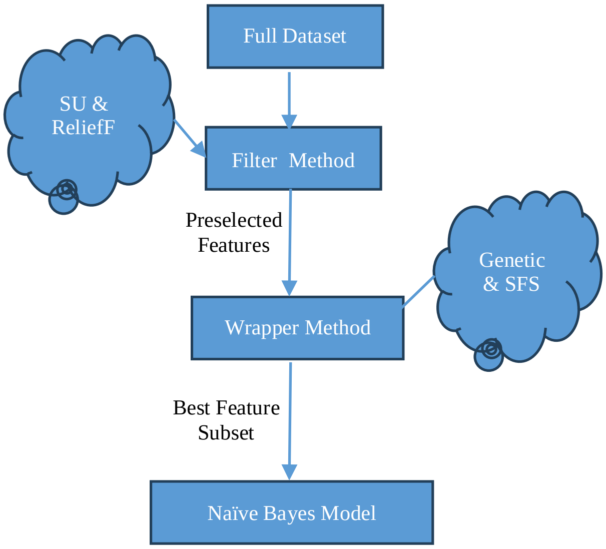

Fig. 4 Overview of hybrid Feature Selection

5. Machine Learning

Machine Learning (ML), a branch of Artificial Intelligence (AI), aims to create models and algorithms that let computers learn from data and make judgements or predictions without the need for explicit programming. Enabling computers to automatically perform better on a given job over time as they are exposed to more data is the aim of Machine Learning. There are several kinds of methods for Machine Learning:

- Supervised Learning: This kind of learning involves training the algorithm using a labelled dataset consisting of matched output labels and input data pairs. By extrapolating patterns from the training data, the algorithm gains the ability to translate the input data into the appropriate output. Two common supervised learning problems are regression and classification.
- Unsupervised Learning: In this case, unlabelled data is supplied to the algorithm, which has to identify patterns or correlations in the data. Unsupervised learning problems include dimensionality reduction and clustering. Similar data points are grouped by the clustering method, whilst the dimensionality reduction technique minimises the number of features while maintaining the relevant information.
- Reinforcement Learning : Behavioural psychology, which teaches agents to make choices by interacting with their surroundings, inspires this learning. Reward or penalty points are given to the agent per the acts it performs. The aim is the agent’s learning of a policy or strategy that maximises the cumulative reward over time.
- Semi-Supervised Learning: This kind of learning combines supervised and unsupervised methods. The dataset used to train the system includes both labelled and unlabelled data. The model applies what it has learned from the labelled input to the unlabelled data.

Several sectors use machine learning, including banking, healthcare, natural language processing, picture and audio recognition, and many more. It is essential for automating difficult processes and making predictions from huge datasets. This research applies the Naive Bayes model with the selected features from the three different feature selections [15].

5.1. Naive Bayes

Naive Bayes is a straightforward probabilistic machine learning method. It is especially well-liked for classification jobs like sentiment analysis and spam email detection. Naive Bayes is computationally efficient and often performs well despite its simplicity.

5.1.1. Fundamentals of Naive Bayes

Bayes’ Theorem

The Bayes theorem, which calculates an event’s probability based on past knowledge of potential contributing factors, is the foundation of the algorithm.

$$P(A|B) = \frac{P(B|A)*P(A)}{P(B)} \quad (2)$$

Regarding classification, this may be stated as:

$$P(Class|Features) = \frac{P(Features|Class)*P(Class)}{P(Features)} \quad (3)$$

Naive Assumption

Naive Bayes assumes that characteristics are conditionally independent, given the class label. Although this is a significant assumption that often proves false in practical situations, the algorithm may function well despite this “naive” assumption.

$$P(Features|class) = P(Feature_1 \, |Class) * (Feature_2 \, |Class) * \dots * (Feature_n \, |Class) \quad (4)$$

5.1.2. Naive Bayes Types

- Gaussian Neural Bayes: Assumes a normal distribution of the characteristics. It works well with continuous data.
- Naive Bayes Multinomial: Used to express counts or frequencies when the characteristics are discrete. Tasks involving text categorization are often used with it.
- Bernoulli Naive Bayes: Suitable for features with binary values. When classifying documents, it is often used to consider each phrase as a binary variable.

5.1.3. Naive Bayes Classification Process

- Data Preprocessing: Transform data into an appropriate format (text data, for example, bag-of-words). Deal with null values.
- Training: Compute the probability of each class - Given the class, determine the conditional probabilities for each feature.

- Prediction: Using the Bayes theorem, determine the probability of each class given the characteristics for a new case. Assign the projected class to the one with the greatest likelihood.

5.1.4. Advantages

- Simplicity: Naive Bayes is a straightforward algorithm to use.
- Efficiency: With big datasets, in particular, it may be computationally efficient.
- Good Performance: It often works well, especially in text categorization and spam filtering, despite its simplicity and the "naive" premise.

5.1.5. Limitations

- Assumption of Independence: Depending on the class, the belief that features are independent may not hold in reality.
- Irrelevant Characteristics Sensitive: It may be susceptible to factors that aren't relevant.
- Estimation Concerns: The probability estimate maybe 0 when a class-feature combination was absent from the training set.

Naive Bayes may be an effective and speedy solution for certain kinds of classification issues despite its drawbacks, particularly provided the independence assumption is not seriously broken [16].

6. Experimental Results

The performance metrics for three distinct scenarios using Naive Bayes (NB) without feature selection, using ReliefF filter feature selection (RELIEFF), and using Symmetrical Uncertainty (SU) filter feature selection are summarized in Figure 5. Symmetrical uncertainty quantifies the information shared by a feature and the target variable in the context of feature selection.

The chosen traits are expected to be instructive for the categorization assignment. The findings imply that while the recall is somewhat low, the accuracy is very good. This suggests that, as the lack of recall means, the model may not be doing an excellent job detecting occurrences of the positive class (low sensitivity).

ReliefF is an algorithm for selecting features that take into account feature repetition as well as relevance. As compared to SU, the findings indicate a little greater accuracy. Still, the recall is much lower, suggesting more false negatives.

Positively, there is an excellent capacity to accurately identify instances of the negative class, as shown by the relatively high specificity. An accuracy comparable to that obtained with feature selection techniques may be obtained using Naive Bayes without feature selection.

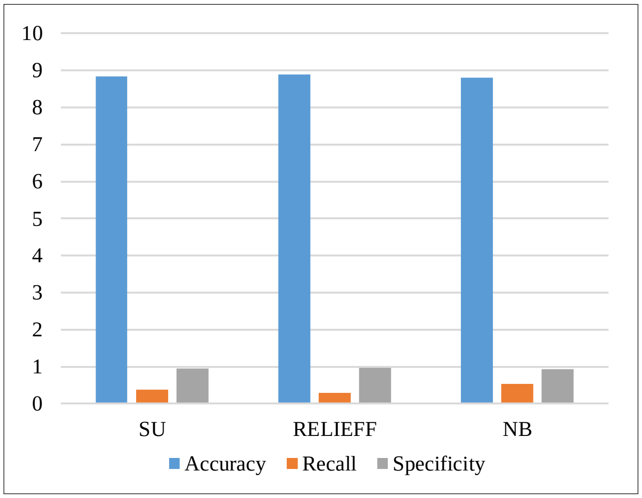

Fig. 5 Results filter NB and NB without FS

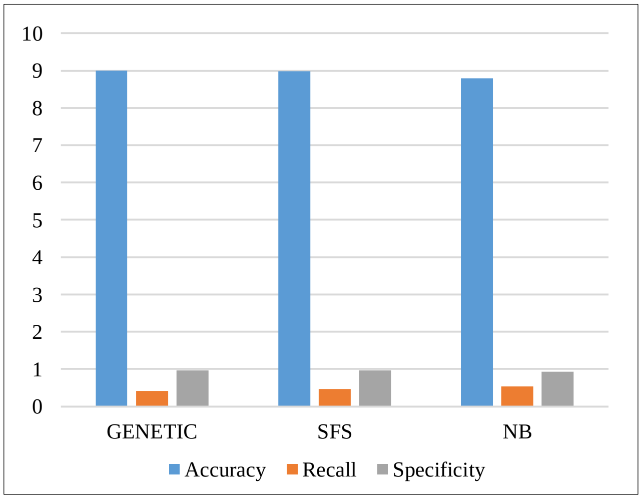

Fig. 6 Results wrapper NB and NB without FS

The recall is greater than that of SU and RELIEFF, suggesting a superior capacity to recognize positive class occurrences. It's crucial to remember that the specificity is somewhat lower than RELIEFF, indicating a more significant likelihood of false positives. Compared to feature selection techniques, Naive Bayes with feature selection (NB) yields comparatively high accuracy and superior recall.

Symmetrical Uncertainty (SU) filter selection has a less fortunate recall but competitive accuracy, suggesting that finding positive cases may be challenging. ReliefF (RELIEFF) has a low recall rate, which means a greater incidence of false negatives while achieving good accuracy and specificity.

Figure 6 displays the performance metrics for three distinct scenarios: utilizing Naive Bayes (NB) without any feature selection, using Genetic Wrapper Feature Selection (GENETIC), and using Sequential Forward Selection (SFS) as a wrapper feature selection approach.

Genetic Wrapper Feature Selection (GENETIC)

The Genetic Wrapper Feature Selection method had the greatest accuracy of the three techniques. This suggests that the genetic algorithm enhanced the model's overall performance by selecting and optimizing a subset of features. The recall is less fortunate, however, indicating that it would be challenging to identify occurrences of the positive class. Positively, there is a high specificity, meaning a solid capacity to recognize instances of the negative class.

Feature Selection with Sequential Forward Selection (SFS) Wrapper

As a wrapper feature selection technique, SFS again showed excellent accuracy. The recall is more significant than the Genetic Wrapper, suggesting a more substantial capacity to recognize positive examples. The specificity is still high, indicating a solid capacity to recognize negative examples.

Naive Bayes without Feature Selection (NB)

Comparable to the SFS technique, a comparable accuracy was obtained using Naive Bayes without feature selection. Recall is the greatest of the three ways, suggesting a superior capacity to recognize good events. That being said, the specificity is somewhat worse than with the wrapper feature selection techniques.

In summary, Sequential Forward Selection and Genetic Wrapper Feature Selection both attained excellent accuracy; however, the genetic approach performed better overall. Naive Bayes showed competitive performance without feature selection, particularly in the recall, indicating that the model's intrinsic simplicity may be sufficient to get good results on the provided dataset without further feature selection.

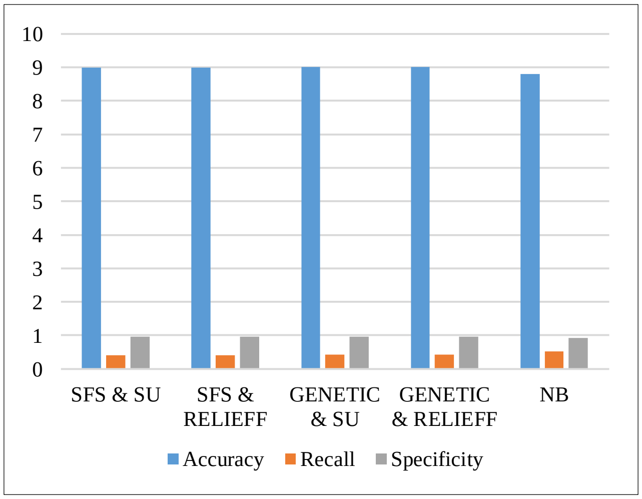

Fig. 7 Results hybrid NB and NB without FS

Results for a hybrid feature selection strategy are shown in Figure 7, which combines two filter feature selection techniques, ReliefF and Symmetrical Uncertainty (SU), with Sequential Forward Selection (SFS) or Genetic Wrapper feature selection. Figure 7 also shows the outcomes of Naive Bayes (NB) without feature selection.

SFS & SU

High accuracy was obtained by integrating SFS and SU in a hybrid technique. While the specificity is high, showing a decent capacity to identify negative cases correctly, the recall is relatively low, suggesting possible difficulty in recognizing positive examples.

SFS & RELIEFF

Similar results were obtained when combining SFS and RELIEFF with SFS & SU. Recall and specificity levels are competitive, and accuracy is good, indicating a well-balanced performance.

GENETIC & SU

Similar to using Genetic Wrapper alone, the hybrid strategy of combining Genetic Wrapper with SU produced results with great accuracy. The recall has somewhat improved compared to SFS & SU, but the specificity has not decreased.

Generic & RELIEFF

The Genetic & RELIEFF hybrid strategy maintained a high accuracy, just as the Genetic & SU combination. The recall is somewhat greater, suggesting that good cases may be more easily recognized.

Naive Bayes without Feature Selection (NB)

Naive Bayes showed competitive performance without any feature selection, particularly in the recall, indicating that the model's intrinsic simplicity may be sufficient to get good results on the provided dataset without any feature selection.

In summary, the hybrid feature selection techniques often obtained high accuracies that combined wrapper (SFS or Genetic) and filter (SU or RELIEFF) methods. The trade-offs between accuracy, recall, specificity, and particular objectives and priorities determine which feature selection method to use Hybrid or individual. In terms of recall, Naive Bayes, without feature selection, continued to be a formidable rival, indicating that it may function effectively without the requirement for further feature selection techniques.

6.1. Results Discussion

Figure 8 thoroughly analyses different feature selection techniques used with Naive Bayes (NB) and NB without feature selection. This analysis includes wrapper techniques like Genetic (GENETIC) and Sequential Forward Selection (SFS), filter methods like Symmetrical Uncertainty (SU) and ReliefF, and hybrid approaches that combine wrapper and filter methods. Accuracy, recall, and specificity are among the measures evaluated, offering a comprehensive picture of these approaches' effectiveness.

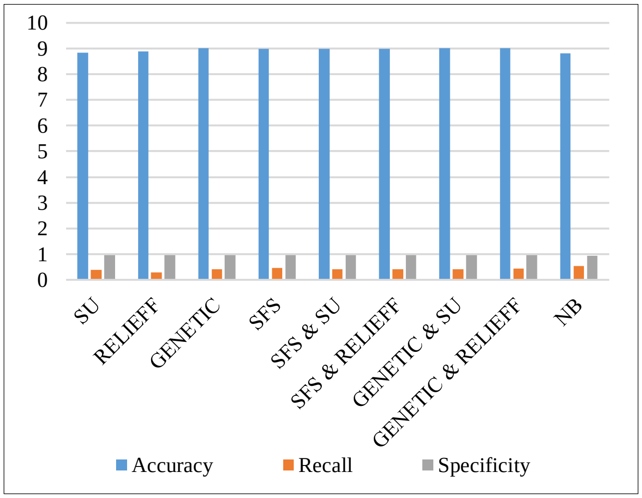

Fig. 8 Results NB with FS and without FS

Based on statistical or information-theoretic metrics, filter techniques are intended to pre-process data and pinpoint the most valuable aspects. Despite its excellent accuracy of 88.41%, SU stands out due to its more excellent recall of 0.385, highlighting its ability to recognize positive situations. RELIEFF, on the other hand, has a lower recall of 0.287 but a somewhat better accuracy of 88.95%.

This disparity raises the possibility of a trade-off between overall accuracy and the capacity to identify successful cases. However, strong specificities show that both SU and RELIEFF can accurately recognise negative instances. The most efficient feature combination for the classifier is found by iterating over subsets of features in wrapper techniques like GENETIC and SFS, which treat feature selection as an optimization issue. GENETIC surpasses SFS by a small margin with an accuracy of 90.01%.

Compared to SFS, GENETIC has a less fortunate recall but a greater specificity. These findings highlight the intrinsic trade-offs in wrapper approaches, where optimising specific metrics could entail sacrificing others. To use the advantages of both, the hybrid feature selection approaches provide a synergistic blend of filter and wrapper techniques.

Accuracy rates for SFS & SU and SFS & RELIEFF are 89.94% and 89.92%, respectively, with a subtle interaction between recall and specificity. This hybrid approach is furthered by GENETIC & SU and GENETIC & RELIEFF, which achieve excellent accuracy of 90.01% and 90.04%, respectively.

These hybrid techniques provide an exciting trade-off between preserving distinctiveness and identifying good examples. With no feature selection, Naive Bayes achieves an accuracy of 88.01% and the most excellent recall of any technique at 0.528. Compared to feature selection with NB, the NB without FS gets less superior results.

The results highlight how crucial it is to modify feature selection techniques per a given classification assignment's particular objectives and priorities. While filter techniques such as SU and RELIEFF show comparable overall performance, wrapper techniques such as GENETIC and SFS provide an alternative viewpoint by focusing on feature subset optimisation.

Hybrid techniques demonstrate the potential for synergy between filter and wrapper methods by attempting to establish a balance. The results prove that the feature selection chooses the best feature subset and improves the NB prediction results.

6.2. Research Findings

A handful of essential insights are revealed by thoroughly examining several feature selection techniques with and without Naive Bayes (NB). The research examined wrapper strategies (Genetic - GENETIC, Sequential Forward Selection - SFS), filter techniques (Symmetrical Uncertainty - SU, ReliefF), and hybrid approaches combining wrapper and filter techniques. The results clarified the trade-offs between performance and each method's advantages and disadvantages.

Filter techniques SU and RELIEFF showed competitive accuracy, with RELIEFF exhibiting stronger specificity and SU emphasizing higher recall. These findings draw attention to the careful decisions that practitioners and researchers need to make depending on the particular objectives of a categorization endeavour. Whereas RELIEFF's focus on specificity may be helpful when reducing false negatives is a top concern, SU's capacity to detect positive occurrences could be critical when sensitivity is critical.

The GENETIC and SFS wrapper techniques treat feature selection as an optimisation issue. There was a little difference in the overall accuracy between GENETIC and SFS. The results imply that whereas wrapper approaches could provide higher accuracy, rigorous trade-off analysis would be necessary, as seen by the complex link between recall and specificity.

Hybrid feature selection techniques presented a synergistic approach by merging filter and wrapper methods. Competitive accuracy was shown by SFS & SU and SFS & RELIEFF, with a well-balanced interaction between recall and specificity. High accuracy was demonstrated by GENETIC & SU and GENETIC & RELIEFF, highlighting the possibility of combining the advantages of filter and wrapper techniques.

Among all the techniques, Naive Bayes without feature selection proved to be a strong rival, attaining competitive accuracy and displaying the greatest recall. This result implies that explicit feature selection may not always be necessary due to the dataset's properties and the Naive Bayes algorithm's inherent simplicity.

The study emphasizes how crucial it is to match feature selection techniques to the particular goals of a classification task. While wrapper methods add optimisation dimensions, hybrid approaches aim to strike a compromise between many techniques, while filter methods could provide simplicity and competitive performance.

To help choose an acceptable feature selection technique, the research recommends having a comprehensive awareness of the classification job's objectives and the dataset's features. Applying these feature selection techniques to various datasets and machine learning algorithms may be the subject of future research.

A deeper understanding of the robustness and generalizability of each feature selection approach may also be gained by examining the effects of changing hyperparameters and tuning procedures within each one. The study's findings add to the larger conversation on feature selection techniques and deepen our understanding of how they affect the performance of machine learning models.

6.3. Drawbacks

It is essential to recognize some of the study's limits and shortcomings even if the research offers insightful information on how different feature selection techniques perform both with and without Naive Bayes (NB).

6.3.1. Restricted Dataset

The study's conclusions are particular to the dataset utilized in it. Results might vary depending on the datasets used, each with unique properties. The inadequacy of variety in datasets restricts the applicability of the results in many application fields.

Algorithms

The selection of machine learning algorithms may impact how well feature selection techniques operate. The study's conclusions may not apply to other classifiers since it concentrated on Naive Bayes. Assessing the resilience of feature selection using various methods may provide a more thorough understanding.

Lack of Ensemble Methods

This study does not explore the possible advantages of using ensemble techniques in addition to feature selection. Compared to single classifiers, ensemble methods-like Random Forests often show resilience and may provide further insights into feature relevance.

A more sophisticated knowledge of feature selection techniques and their application in various contexts may be achieved by addressing these shortcomings and considering them in future research. Extensive analyses spanning several datasets, methods, and evaluation metrics, in conjunction with an exhaustive investigation of optimisation parameters, may augment the resilience and applicability of subsequent studies within this field.

7. Conclusion

This study explored the intricacies and difficulties presented by high-dimensional data or datasets with many characteristics or dimensions. The quality of prediction outputs may be significantly impacted by the main problems associated with high-dimensional data, which include the curse of dimensionality, greater computing loads, and lower interpretability. We suggested an approach based on feature selection methods to overcome these issues.

Three feature selection techniques were explored and implemented: filter, wrapper, and hybrid approaches. These techniques are intended to reduce the dimensionality of the data by identifying and keeping the most essential attributes while eliminating superfluous or unnecessary ones.

Using a Naive Bayes machine learning model, we used this technique for a high-dimensional real-world dataset obtained from the UCI repository. The data was analyzed with and without feature selection.

Our findings illustrated the significant advantages of feature selection in situations involving high-dimensional data. Compared to using the whole feature set, feature selection decreased computing demands while improving model performance and interpretability. In addition, we evaluated the efficacy of the filter, wrapper, and hybrid approaches and discovered that the hybrid technique performed better than the other two.

In conclusion, high-dimensional data poses several obstacles that may reduce the efficacy of predictive modelling. These problems are effectively addressed by feature selection methods, which lower dimensionality while improving the accuracy of prediction models. This study adds to the expanding body of knowledge in high-dimensional data analysis by providing academics and practitioners with valuable insights. It helps them make better judgements when dealing with high-dimensional datasets.

References

[1] Peter Bühlmann, and Sara van de Geer, *Statistics for High-Dimensional Data: Methods, Theory and Applications*, Springer Science & Business Media, 2011. [CrossRef] [Google Scholar] [Publisher Link]

[2] Marius Muja, and David G., "Lowe Scalable Nearest Neighbor Algorithms for High Dimensional Data," *IEEE Transactions on Pattern Analysis and Machine Intelligence*," vol. 36, no. 11, pp. 2227-2240, 2014. [CrossRef] [Google Scholar] [Publisher Link]

[3] Andrea Bommert et al., "Benchmark for Filter Methods for Feature Selection in High-Dimensional Classification Data," *Computational Statistics & Data Analysis*, vol. 143, pp. 1-19, 2020. [CrossRef] [Google Scholar] [Publisher Link]

[4] G. Thippa Reddy et al., "Analysis of Dimensionality Reduction Techniques on Big Data," *IEEE Access*, vol. 8, pp. 54776-5477, 2020. [CrossRef] [Google Scholar] [Publisher Link]

[5] Papia Ray, S. Surender Reddy, and Tuhina Banerjee, "Various Dimension Reduction Techniques for High Dimensional Data Analysis: A Review," *Artificial Intelligence Review*, vol. 54, pp. 3473-3515, 2021. [CrossRef] [Google Scholar] [Publisher Link]

[6] Jamshid Pirgazi et al., "An Efficient Hybrid Filter-Wrapper Metaheuristic-Based Gene Selection Method for High Dimensional Datasets," *Scientific Reports*, vol. 9, pp. 1-15, 2019. [CrossRef] [Google Scholar] [Publisher Link]

[7] Cai Jie et al., "Feature Selection in Machine Learning: A New Perspective," *Neurocomputing*, vol. 300, pp. 70-79, 2018. [CrossRef] [Google Scholar] [Publisher Link]

[8] Hongtao Shi et al., "An Efficient Feature Generation Approach Based on Deep Learning and Feature Selection Techniques for Traffic Classification," *Computer Networks*, vol. 132, pp. 81-98, 2018. [CrossRef] [Google Scholar] [Publisher Link]

[9] Yu Xue, Haokai Zhu, and Ferrante Neri, "A Feature Selection Approach Based on NSGA-II with ReliefF," *Applied Soft Computing*, vol. 134, 2023. [CrossRef] [Google Scholar] [Publisher Link]

[10] Baoshuang Zhang, Yanying Li, and Zheng Chai, "A Novel Random Multi-Subspace Based ReliefF for Feature Selection," *Knowledge-Based Systems*, vol. 252, 2022. [CrossRef] [Google Scholar] [Publisher Link]

[11] Annu Lambora, Kunal Gupta, and Kriti Chopra, "Genetic Algorithm-A Literature Review," *2019 International Conference on Machine Learning, Big Data, Cloud and Parallel Computing (COMITCon)*, Faridabad, India, pp. 380-384, 2019. [CrossRef] [Google Scholar] [Publisher Link]

[12] Negar Maleki, Yasser Zeinali, and Seyed Taghi Akhavan Niaki, "A k-NN Method for Lung Cancer Prognosis with the Use of a Genetic Algorithm for Feature Selection," *Expert Systems with Applications*, vol. 164, 2021. [CrossRef] [Google Scholar] [Publisher Link]

[13] Knitchepon Chotchantarakun, "Optimizing Sequential Forward Selection on Classification Using Genetic Algorithm," *Informatica*, vol. 47, no. 9, pp. 81-90, 2023. [Google Scholar] [Publisher Link]

[14] A. Pasyuk, E. Semenov, and D. Tyuhtyaev, "Feature Selection in the Classification of Network Traffic Flows," *2019 International Multi-Conference on Industrial Engineering and Modern Technologies (FarEastCon)*, Vladivostok, Russia, pp. 1-5, 2019. [CrossRef] [Google Scholar] [Publisher Link]

[15] Feng-Jen Yang, "An Implementation of Naive Bayes Classifier," *2018 International Conference on Computational Science and Computational Intelligence (CSCI)*, Las Vegas, USA, pp. 301-306, 2018. [CrossRef] [Google Scholar] [Publisher Link]

[16] Mucahid Mustafa Saritas, and Ali Yasar, "Performance Analysis of ANN and Naive Bayes Classification Algorithm for Data Classification," *International Journal of Intelligent Systems and Applications in Engineering*, vol. 7, no. 2, pp. 88-91, 2019. [Google Scholar] [Publisher Link]