INTERNATIONAL JOURNAL FOR INNOVATIVE RESEARCH IN MULTIDISCIPLINARY FIELD
ISSN(O): 2455-0620 [Impact Factor: 7.581]
Monthly, Peer-Reviewed, Refereed, Indexed Journal with IC Value : 86.87
Volume - 9, Issue - 10, October - 2023 Publication Date: 05/10/2023

DOIs:10.2015/IJIRMF/202310004 --:-- Research Paper / Article / Review

Feature Selection for Classification in Data Mining: Methods, Benefits and Applications

[1]Nancy Kansal, [2]Dr. Ganesh Gopal Varshney
[1]Research Scholar, Department of Computer Applications, Mewar University, Chittorgarh, Rajasthan
[2]Professor, Mewar University, Chittorgarh, Rajasthan, India
Email Corresponding author: – nancysharmaits@gmail.com

Abstract: *Data mining is a crucial field in extracting valuable insights from vast datasets. One of its key challenges is dealing with high-dimensional data, where irrelevant or redundant features can affect classifiers performance. Feature selection methods play a pivotal role in enhancing the accuracy and efficiency of data mining algorithms by identifying the most informative attributes. This paper explores various feature selection techniques and their impact on classifier performance, highlighting their significance in optimizing the data mining process. In this paper, we have discussed various feature selection approaches such as filter based, Wrapper based and embedded methods while shedding light on their advantages and limitations. By examining the applications of FS methods and support vector machine (SVM) as a wrapper based classifiers in real-world scenarios, this paper serves as a valuable resource for researchers looking for future research directions of feature selection in the field of data mining.*

Key Words: *Feature Selection Methods, Data Mining classifier, Support Vector Machine, Crow Search Algorithm Optimization.*

1. DATA MINING PROCESS AND CLASSIFICATION TASK- AN INTRODUCTION :

Data mining is a process of discovering hidden patterns, relationships and information from large datasets which involves various techniques and steps to extract valuable insights from data and Classification is one of the fundamental tasks in data mining, where the goal is to categorize data points into predefined classes or labels based on their characteristics. Data mining holds immense potential in uncovering hidden patterns within diverse datasets across various domains, enabling the generation of valuable predictions. The process of data pre-processing encompasses critical tasks such as data cleaning, integration, transformation, and reduction. Implementing these pre-processing techniques before the actual mining significantly enhances both the quality of the discovered patterns and the efficiency of the mining process. Data pre-processing plays a pivotal role in the knowledge discovery process and emphasize that the sound decisions depend on the foundation of high-quality data. As data mining plays a crucial role in extracting valuable insights from vast datasets therefore the success of data mining techniques heavily relies on the quality of features used in the analysis [1].

In data mining, classification refers to the process of categorizing data into predefined classes or groups based on its characteristics. It is a fundamental data mining technique that involves training a model to learn patterns and relationships in the data, which can then be used to assign new, unseen data points to one of the established classes. Classification is widely used in various fields including marketing, healthcare, finance, education and more. The goal of classification is to make accurate predictions or decisions based on the available data and due to this, classification is considered as a valuable tool for extracting actionable insights and automating decision-making processes in data-driven industries [2].

Curse of Dimensionality and Dimension Reduction Techniques

Feature selection, also known as attribute selection or dimensionality reduction, is also a crucial technique in data mining and machine learning as it aims to address the problems associated with high-dimensional data sets. Working with high-dimensional data poses challenges related to computational complexity, irrelevant attributes and attribute correlations which can impact the quality and efficiency of data mining tasks, referred to as the "Curse of Dimensionality". It can significantly impact the performance of data mining and machine learning algorithms making them less efficient and accurate as the dimensionality of the data grows. High-dimensional data poses challenges for classification algorithms due to their high computational cost and memory requirements. Increase in the number of features (dimensions) leads to exponential growth in computational costs for data mining tasks and accordingly, reducing the number of variables in a dataset can enhance the accuracy, efficiency, and scalability of data mining algorithms [3].

This research paper begins with an understanding of data mining concepts with the significance to use feature selection techniques in optimizing the data mining process. The study discusses about the significance of feature selection in reducing dimensionality, improving model interpretability and enhancing predictive power. Key feature selection methods such as filter, wrapper, and embedded are also discussed along with their strengths and limitations in the paper. The study will further explore various feature selection techniques & their impact on classifiers' accuracy & efficiency. In this research study, a applications of the feature selection methods and support vector machine as a wrapper based classifier in real world scenarios are also discussed so that it can act as a valued source for researchers looking for future research directions of feature selection in the field of educational data mining.

2. DIMENSIONALITY REDUCTION TECHNIQUES- FEATURE EXTRACTION VS FEATURE SELECTION

In data mining, the common dimensionality reduction techniques used are: feature extraction (dimensionality reduction) and feature selection (FS). While comparing both techniques, feature selection serves multiple purposes including enhancing data comprehension, minimizing computational demands, improving the curse of dimensionality issue and boosting performance of the prediction models. Feature selection retains information about each feature's importance, making it advantageous. However, when a small feature set is needed and original features are diverse, some information loss may occur as certain features are omitted. In contrast to this, feature extraction reduces the feature space size without significant information loss. However, the choice between these two dimensionality reduction methods depends on the data type and application domain, primarily [4].

3. FEATURE SELECTION FOR CLASSIFICATION:

Feature selection, a fundamental step in the data pre-processing pipeline aims to identify and retain the most relevant features while discarding irrelevant or redundant ones. Effective feature selection is crucial in enhancing the performance of classifiers in data mining. The goal of feature selection is twofold: 1. To enhance the predictive performance of classifiers 2. To reduce computational overhead [5].

Feature selection has gained significant attention in research and practical applications as feature selection methods helps in improving classification accuracy, interpretability and efficiency that results in more effective predictive data mining and machine learning models. Feature selection methods are essential in classification tasks for several reasons [6]:

a) Firstly, they help in enhancing model performance by identifying and keeping the most relevant features, reducing noise and overfitting. This improves predictive accuracy and speeds up model training and inference.
b) Secondly, feature selection helps in interpretability that is important in domains like healthcare, education, finance and others by clarifying key factors in classification decisions.
c) Lastly, it helps in reducing computational complexity and resource usage, especially with high-dimensional data.

Advantages of Feature Selection: Feature selection offers numerous advantages in data analysis and machine learning. First and foremost, it efficiently reduces the dimensionality of the feature space, leading to reduced storage requirements and faster algorithm execution. Additionally, it helps in eliminating redundant, irrelevant or noisy data, enhancing data quality that results in more accurate models. By streamlining the feature set, it also saves valuable resources in subsequent data collection and utilization phases. Furthermore, feature selection contributes to performance improvement, leading to higher predictive accuracy. Lastly, it helps in gaining a deeper understanding of the data and the underlying processes that generated it, facilitating knowledge extraction and data visualization [7]. In summary, feature selection is a crucial step in data analysis with far-reaching benefits across various aspects of the process.

4. FEATURE SELECTION PROCESS:

Dash and Liu (1997) introduced a structured approach to feature selection in four key steps: First, the "generation" step involves exploring the feature space to identify the subset that can best predict the class variable. Next, in the "evaluation" phase, an evaluation function, like Information Gain or Correlation Analysis, quantifies the quality of feature subsets, establishing a ranking used in subsequent selection. Thirdly, a "stopping criterion" determines when to halt the search, whether due to no improvement in classification accuracy, reaching a set feature limit, or other predefined conditions. Finally, the "validation" step assesses the chosen feature subset's performance using the selected machine learning induction algorithm, ensuring its effectiveness on unseen data. This systematic process aids in feature selection for enhanced model efficiency and predictive power.

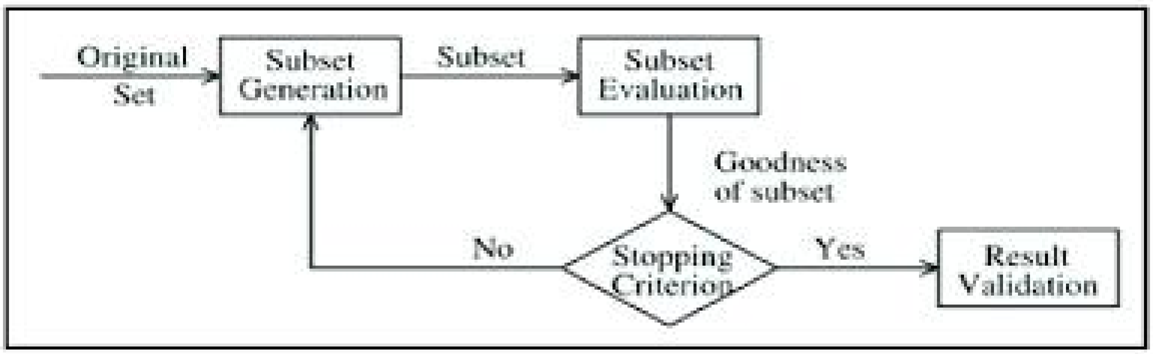

Fig 4.1: Four Key steps in Feature Selection Process [6]

5. CLASSIFICATION OF FEATURE SELECTION METHODS:

Feature selection (FS) methods which are commonly used in data mining and machine learning to choose the most relevant subset of features for a particular task are primarily classified into three main categories: Filter FS methods, Wrapper FS methods and Embedded FS methods [8].

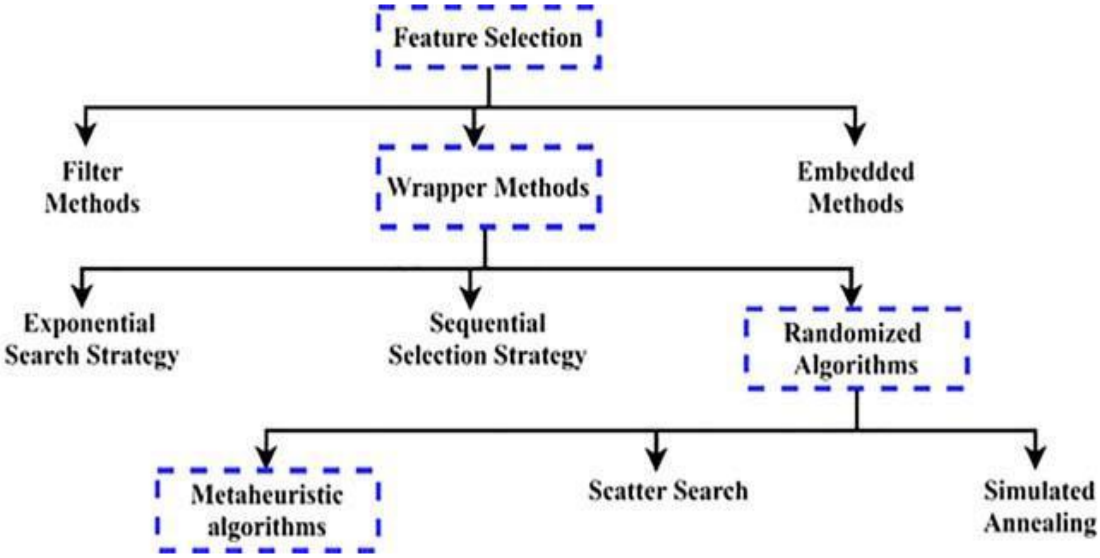

Fig 5.1: Flow chart of categorization of methods for solving Feature Selection(FS) [10]

A. Filter Methods:

Filter methods in feature selection assess the importance of each feature by using statistical measures like correlation, mutual information or statistical tests. These measures rank or score the features and a set number of top-ranked features are chosen for further analysis. Filter feature selection is computationally efficient and can enhance model performance by reducing dimensionality and removing irrelevant or redundant features, making it a valuable tool in feature engineering and data pre-processing [3] [9].

INTERNATIONAL JOURNAL FOR INNOVATIVE RESEARCH IN MULTIDISCIPLINARY FIELD
ISSN(O): 2455-0620 [Impact Factor: 7.581]
Monthly, Peer-Reviewed, Refereed, Indexed Journal with IC Value : 86.87
Volume - 9, Issue - 10, October - 2023 Publication Date: 05/10/2023

One of the main advantages of filter methods in feature selection is that they operate without any reliance on specific learning or classification algorithms [10]. Filter methods assess the relevance of features to the target variable or the dataset's overall structure based on statistical measures such as correlation or mutual information. Since these measures are calculated without considering a specific machine learning algorithm, filter methods can be applied before selecting a particular model, making them versatile and applicable in various contexts without being tied to a specific learning algorithm. This independence from the learning algorithm can simplify the feature selection process and make it more flexible. However, they have limitations in identifying complex feature relationships and may overlook interdependencies between features potentially leading to suboptimal feature subsets [10].
There are several types of filter feature selection methods, each based on different statistical measures or criteria for evaluating feature relevance. Filter feature selection methods cannot be universally applied to all types of data mining tasks. Consequently, these methods are categorized based on the specific data mining tasks they are suited for, which typically include Classification, Regression, or Clustering [11].

Some of the types of Filter FS methods and applicable to the DM task are:
Classification: Information gain, Gain Ratio, Chi-squared test, ANOVA (Analysis of Variance) ETC
Regression: Correlation & Correlation based FS methods, minimum Redundancy maximum Relevance (mRmR) etc
Clustering: Mutual Information, Variance Threshold, ReliefC etc

B. Wrapper Methods:

Wrapper feature selection methods are a class of feature selection techniques that evaluate the performance of a machine learning model using different subsets of features. These methods treat feature selection as a search problem and use the predictive performance of the model as a criterion for selecting the best subset of features [3] [9]. The wrapper methods are advantageous because they work closely with the classification algorithm, enabling a precise selection of features that fit the model's unique requirements. Wrapper methods always incorporate the classification algorithm and interact directly with the classifiers [12].

Wrapper feature selection methods often yield better results compared to filter methods (which rely solely on feature statistics) because they take into account the interaction between features and the model's performance, though they can be computationally expensive and slower as they involve training and evaluating the model multiple times for different feature subsets. Furthermore, wrapper methods might lead to overfitting because they rely on the model's performance on training data to select features, which could result in less effective predictions on new or unseen data. Therefore, the choice of a wrapper method should consider factors such as the dataset size, computational resources, and the nature of the classification problem. Some common wrapper feature selection methods are Forward Selection, Backward Elimination, Recursive Feature Elimination (RFE), Genetic Algorithms, Metaheuristic Algorithms and others [10].

There are several types of wrapper feature selection methods categorized based on the search strategy or modelling algorithm like in classification task, a wrapper method assesses feature subsets by measuring their impact on the performance of a classifier like Support Vector Machine or Naïve Bayesian. On the other hand, in clustering tasks, a wrapper evaluates feature are based on how they influence the performance of a clustering algorithm such as K-means clustering algorithm [11].

C. Embedded Methods:

Embedded feature selection methods, also known as in-model feature selection or intrinsic feature selection that are a category of feature selection techniques integrated directly into the process of training a machine learning model. Unlike wrapper methods that use external evaluation criteria to select features, embedded methods perform feature selection as an inherent part of the model training process [3] [9].

Embedded feature selection methods are techniques that are integrated into the machine learning model training process for automatically selecting features as part of the model building. They are computationally efficient, making them suitable for large datasets and complex models. However, their effectiveness depends on the chosen algorithm and may not always produce the best feature subsets compared to wrapper methods. Embedded methods incorporate feature selection into the model building process and select features as part of the model training. Embedded methods are specific to certain machine learning algorithms like decision trees, Lasso regression, or feature importance in random forests [11].

These algorithms naturally perform feature selection during training by assigning importance scores to features. They are efficient and consider feature relevance within the context of the chosen model, which can lead to good feature subsets for that model. However, they are model-dependent and may not generalize well to other algorithms and may also require fine-tuning of model-specific parameters.

Embedded feature selection methods are advantageous because they are computationally efficient, as feature selection occurs simultaneously with model training [8]. This can be particularly useful when dealing with large datasets and complex models. However, the effectiveness of embedded methods depends on the algorithm and may not always yield the best feature subsets compared to wrapper methods. The choice between embedded and wrapper methods should be made based on the specific problem and the characteristics of the dataset [9].

6. UNLOCKING CLASSIFICATION PERFORMANCE: THE POWER OF SUPPORT VECTOR MACHINE (SVM) AS WRAPPER-BASED CLASSIFIER:

Classification is a widely recognized and frequently used data mining technique. Various types of classification algorithms include C4.5, the Decision Tree algorithm, K-Nearest Neighbor (K-NN), and Support Vector Machine (SVM) algorithms. Classification algorithms serve as the foundation of supervised machine learning, facilitating the organization of data into predefined classes or labels. Each classification algorithm is paired with a specific classifier, employing unique mathematical methodologies and strategies to formulate predictions. For example, Logistic Regression utilizes a linear model for binary classification tasks, Decision Trees adopt a tree-like structure to differentiate between classes and Random Forests leverage multiple decision trees to enhance predictive accuracy. Whether one opts for Support Vector Machines, K-Nearest Neighbor, Naive Bayes, Neural Networks, Gradient Boosting, or other options, it would depend on the specific problem and data characteristics. This diversity makes classification algorithms a versatile toolbox applicable across a broad spectrum of tasks from text categorization to image recognition and many others [13].

In a research study, the researcher had conducted a study on the impact of feature selection on the accuracy of three distinct classifiers: Naïve Bayes, Artificial Neural Network utilizing the Multilayer Perceptron architecture and the J48 decision tree classifier. These classifiers were systematically compared using fifteen real datasets that had been pre-processed with various feature selection methods. The analysis revealed a substantial enhancement in classification accuracy with improvements of up to 15.55% observed across the datasets. Notably, the Multilayer Perceptron emerged as the most responsive classifier to feature selection underscoring its significance in this research context [14].

In 1963, Vladimir N. Vapnik and Alexey Ya. Chervonenkis introduced the SVM algorithm, which falls within the domain of machine learning. Its primary objective is to discover the most suitable hyperplane with the maximum margin [15]. This algorithm proficiently separates data samples that are linearly separable into two separate classes. When confronted with non-linearly separable data, SVM addresses this challenge by mapping the data into a high-dimensional feature space and subsequently conducting the classification [15]. Support Vector Machines (SVMs) when used as a wrapper-based classifier can significantly enhance classification performance. SVMs are employed to select the most relevant features and optimize hyper parameters in a systematic manner. This process improves the accuracy and efficiency of the classifier, particularly in scenarios with high-dimensional data. Additionally, SVMs can be integrated into ensemble methods and utilized in cross-validation, ensuring robust and well-generalized models. The flexibility of SVMs in handling complex decision boundaries makes them a valuable tool for enhancing classification model quality, making them a popular choice in machine learning applications [16].

Support Vector Machines (SVMs) can be combined with feature selection techniques and other classification algorithms to identify the most relevant attributes for a particular classification task and the process involves training the SVM using different subsets of features & evaluating its performance. Features that have the most significant impact on improving the SVM's performance are selected for inclusion in the final model, while less influential attributes can be excluded. This approach can lead to a classifier that is not only more efficient but also more accurate especially when dealing with datasets containing a large number of dimensions.

In an important research study, the authors, five different wrapper-based algorithms namely Crow Search Algorithm (CSA), Particle Swarm Optimization (PSO), Tree Seed Algorithm (TSA), Slime Mould Algorithm (SMA), and Artificial Bee Colony (ABC) were employed to reduce and select input attributes for the diagnosis of diabetes disease. The results of these algorithms were then compared using traditional machine learning algorithms including Support Vector Machine (SVM), Decision Tree (DT), Random Forest (RF), K Nearest Neighbor (kNN) and Feed Forward Neural Networks (FFNN). The initial dataset that contained 16 features for diabetes diagnosis after applying the wrapper-based

feature selection and reduction methods were determined 13 for CSA, 10 for PSO and subsequently, these features were subjected to a combination of CSA+FFNN & PSO + SVM classifiers. The CSA+ FFNN yielded the best classification performances with a remarkable success rate of 99.04% with PSO+SVM having second highest success rate of 97.5% in comparison to other combinations like ABC+SVM (96.73%), SMA+FFNN (94.23%) and TSA + SVM (96.15%) [17].
In an another study, the researchers focus on enhancing support vector machines through the utilization of the Crow Search Algorithm and that refined support vector machine classifier is known as CSA-SVM and is applied to achieve precise diagnosis in the context of Indian liver disease data [18]. To demonstrate its effectiveness, the authors considered several state-of-the-art algorithms for comparison and the performance evaluation of CSA-SVM reveals its exceptional performance in comparison to all other methods in terms of various assessment metrics with a remarkable classification accuracy of 99.49%. In their literature study itself, the researcher presented that a number of researchers have carried out experiments for the performance evaluation of SVM classifier including the implementation of a diverse range of classification algorithms for liver disease diagnosis like Naïve Bayes, J48, Random Tree through the WEKA tool; logistic Regression, Support Vector Machine (SVM), Random Tree techniques were compared for their classification accuracy; a Multi-layer Feed-Forward Deep Neural Network (MLFFDNN) was trained with Back-Propagation Network (BPN) and combining Particle Swarm Optimization (PSO) with SVM for feature selection & liver data classification. In this research work, it is highlighted that SVM has shown promising results in liver disease diagnosis [18].

Therefore, it is evident that Support Vector Machines (SVMs) have the potential to substantially boost classification performance, when used as a wrapper-based classifier.

7. SUMMARY AND FUTURE RESEARCH DIRECTIONS:

At the conclusion, this paper has provided a comprehensive overview of the significance of feature selection in the field of data mining and its profound impact on classifier performance. We have discussed various feature selection techniques categorized as filter, wrapper, and embedded methods, highlighting their strengths and limitations. Filter methods efficiently reduce dimensionality and enhance model efficiency by eliminating irrelevant and redundant features based on statistical measures. Wrapper methods, on the other hand, treat feature selection as a search problem, utilizing the performance of machine learning models to select the best feature subsets. Embedded methods seamlessly integrate feature selection into the model training process, offering computational efficiency and relevance to specific algorithms. Furthermore, we emphasized the critical role of Support Vector Machines (SVMs) as wrapper-based classifiers in optimizing classification tasks. SVMs can significantly enhance classification performance by systematically selecting relevant features and optimizing hyper parameters. The flexibility of SVMs in handling complex decision boundaries makes them invaluable in data mining applications.

However, through this research study, we tried to highlight valuable insights related to feature selection and its impact on classification, though there are several promising research directions that need exploration in future. For the future research directions in the field of feature selection, it is proposed that future research should prioritize the development of advanced techniques tailored to specific data types and mining tasks. These innovations should not only cater to the unique characteristics of diverse datasets but also explore the integration of deep learning and reinforcement learning into the feature selection process, potentially producing more effective methods. Moreover, in domains like healthcare, education, finance and others where transparency and interpretability are important for ensuring the decision-making processes of data mining models, more related experiments are required to be carried out for the development of novel predictive models in the respective domain. An another important aspect is the scalability as datasets continue to grow in size and complexity. Future researchers, can also focus on crafting feature selection techniques that can efficiently handle massive datasets, necessitating exploration into distributed and parallel computing approaches to address the challenges posed by high-dimensional data. Lastly, by considering multiple criteria at the same time, such as model accuracy, interpretability and computational efficiency, researchers can develop more comprehensive and adaptable feature selection algorithms and processes that can further help them in improving the performance of the predictive models. These research directions shall further advance the field of feature selection and contribute to the development of more efficient and accurate data mining and machine learning models across various domains and applications.

REFERENCES:

1. Han, J., Pei, J., & Tong, H. (2022). Data mining: concepts and techniques. Morgan kaufmann.

2. Kesavaraj, G., & Sukumaran, S. (2013, July). A study on classification techniques in data mining. In *2013 fourth international conference on computing, communications and networking technologies (ICCCNT)* (pp. 1-7). IEEE.
3. Guyon, I., & Elisseeff, A. (2003). An introduction to variable and feature selection. *Journal of machine learning research*, *3*(Mar), 1157-1182.
4. Janecek, A., Gansterer, W., Demel, M., & Ecker, G. (2008, September). On the relationship between feature selection and classification accuracy. In *New challenges for feature selection in data mining and knowledge discovery* (pp. 90-105). PMLR.
5. Dash, M., & Liu, H. (2003). Consistency-based search in feature selection. *Artificial intelligence*, *151*(1-2), 155-176.
6. Dash, M., & Liu, H. (1997). Feature selection for classification. *Intelligent data analysis*, *1*(1-4), 131-156.
7. Ladha, L., & Deepa, T. (2011). Feature selection methods and algorithms. *International journal on computer science and engineering*, *3*(5), 1787-1797.
8. Kumar, V., & Minz, S. (2014). Feature selection: a literature review. *SmartCR*, *4*(3), 211-229.
9. Chandrashekar, G., & Sahin, F. (2014). A survey on feature selection methods. *Computers & Electrical Engineering*, *40*(1), 16-28.
10. Agrawal, P., Abutarboush, H. F., Ganesh, T., & Mohamed, A. W. (2021). Metaheuristic algorithms on feature selection: A survey of one decade of research (2009-2019). *IEEE Access*, *9*, 26766-26791.
11. Jović, A., Brkić, K., & Bogunović, N. (2015, May). A review of feature selection methods with applications. In *2015 38th international convention on information and communication technology, electronics and microelectronics (MIPRO)* (pp. 1200-1205). IEEE.
12. Akinola, O. O., Ezugwu, A. E., Agushaka, J. O., Zitar, R. A., & Abualigah, L. (2022). Multiclass feature selection with metaheuristic optimization algorithms: a review. *Neural Computing and Applications*, *34*(22), 19751-19790.
13. Wu, X., Kumar, V., Ross Quinlan, J., Ghosh, J., Yang, Q., Motoda, H., ... & Steinberg, D. (2008). Top 10 algorithms in data mining. *Knowledge and information systems*, *14*, 1-37.
14. Karabulut, E. M., Özel, S. A., & Ibrikci, T. (2012). A comparative study on the effect of feature selection on classification accuracy. *Procedia Technology*, *1*, 323-327.
15. Cortes, C., & Vapnik, V. (1995). Support-vector networks. *Machine learning*, *20*, 273-297.
16. Gad, A. G., Sallam, K. M., Chakrabortty, R. K., Ryan, M. J., & Abohany, A. A. (2022). An improved binary sparrow search algorithm for feature selection in data classification. *Neural Computing and Applications*, *34*(18), 15705-15752.
17. Yasar, A. (2021). Data classification of early-stage diabetes risk prediction datasets and analysis of algorithm performance using feature extraction methods and machine learning techniques. *International Journal of Intelligent Systems and Applications in Engineering*, *9*(4), 273-281.
18. Devikanniga, D., Ramu, A., & Haldorai, A. (2020). Efficient diagnosis of liver disease using support vector machine optimized with crows search algorithm. *EAI Endorsed Transactions on Energy Web*, *7*(29), e10-e10.