International Journal of Computer Sciences and Engineering
Vol.**12**, Issue.**5**, pp.**54-58**, May **2024**
ISSN: 2347-2693 (Online)
Available online at: www.ijcseonline.org

ISSN: 2347-2693 (E)

Review Article

A Scientific Review of Feature Selection Algorithms and Kernel Methods for SVM Classification Models

C. Dharmadevi[1*], S. Thaddeus[2]

[1]Sacred Heart College (Autonomous), Tirupattur, India
[2]Don Bosco College (Co-Ed), Yelagiri, India

**Corresponding Author: udharmadevi@gmail.com*

Received: 27/Mar/2024; **Accepted**: 29/Apr/2024; **Published**: 31/May/2024. **DOI**: https://doi.org/10.26438/ijcse/v12i5.5458

Abstract: Feature selection is crucial for improving the efficiency and effectiveness of machine learning models by identifying and choosing the most pertinent subset of features from the original dataset. This review article comprehensively surveys a diverse range of feature selection techniques in the context of Support Vector Machine (SVM) classification model in machine learning. This research work delves into several prominent techniques, including Mutual Information, Chi-Square, Sequential Feature Selection (SFS), Recursive Feature Elimination (RFE), LASSO, and Random Forest. The study reveals that RFE (Recursive Feature Elimination) emerges as the highly effective feature selection technique, demonstrating superior performance metrics compared to the other methods considered. Additionally, the study proposes the integration of hybrid algorithms to further enhance the performance of SVM classification models.
Furthermore, this review extends its scope to encompass an evaluation of various kernel methods within the SVM classification paradigm, offering a comprehensive perspective on their efficacy and performance.

Keywords: Feature Selection Technique, Chi-Square, RFE (Recursive Feature Elimination), SVM (Support Vector Machine), Kernel methods.

1. Introduction

In this paper, we provide a comprehensive review of feature selection techniques and kernel methods for SVM classification models. Feature selection techniques are crucial for identifying relevant features to enhance model performance and interpretability, particularly in healthcare and disease prediction domains [1],[11]. Various studies have demonstrated the effectiveness of methods like Recursive Feature Elimination (RFE) and hybrid approaches in improving classification accuracy [2],[7],[8],[12]. Additionally, kernel methods, especially the radial basis function (RBF) kernel, have proven to be powerful tools for handling non-linear relationships and improving SVM model performance [15],[20]. The review aims to offer insights into the strengths and applications of these techniques across diverse medical datasets.

This approach helps identify the most suitable feature selection techniques, ensuring higher accuracy, more relevant feature sets, and reduced computational time for the SVM classification model. This assessment applies to various categories of feature selection methods, such as filter, wrapper, and embedded methods. Moreover, the review evaluates different kernel methods within the SVM classification model, providing a detailed analysis of their effectiveness and performance.

2. Related Work

II.I. Feature Selection Techniques

Feature selection techniques are essential in machine learning prediction and classification, facilitating the identification of relevant features to enhance model performance and interpretability. This literature review examines various methods applied across different domains, including healthcare and disease prediction.

Studies such as Mazreati et al. have compared feature selection methods for gastric cancer prediction. Techniques like RFE (Recursive Feature Elimination) and SVM-RFE (Support Vector Machine Recursive Feature Elimination) are commonly employed to achieve this goal [1]. Similarly, Theerthagiri and Siddalingaiah proposed a hybrid Recursive Gaussian SVM-based approach for liver disease classification, improving classification accuracy [3]. Alcaraz et al. introduced a multiobjective feature selection approach using Support Vector Machines (SVM) to optimize classification performance and feature subset size [2]. Gokulnath and Shantharajah developed an optimized feature selection technique for heart disease classification, integrating genetic algorithms and SVM for accurate feature selection [6].

Hybrid methods have gained traction, notably in healthcare. Sheelal and Arun devised a combined PSO–SVM algorithm for Covid-19 detection, emphasizing the importance of feature selection [7]. An enhanced [8], [20] SVM-based approach for diabetic readmission prediction, highlighting the role of feature selection in healthcare management was proposed. Additionally, Enireddy et al. and Shaban et al. presented hybrid approaches for COVID-19 detection, demonstrating the effectiveness of feature selection in medical imaging and classifier optimization [9],[10]. Furthermore, Gholami et al. employed Recursive Feature Elimination (RFE) for brain tumor classification, while Wottschel et al. used SVM recursive feature elimination for predicting disease progression in multiple sclerosis patients [11],[12].

Table.1. Analysis of Features Selection Techniques used with SVM for different domains in healthcare

Author	Data Type / Set	Feature Selection Algorithm	Benefits & Limitations
[1]Hamed Mazreati	Gastric Cancer	Filter, Wrapper and Embedded	Compare to Filter and Wrapper method embedded method shows high performance accuracy
[2] Mei-Ling Huang	Dermatology and Zoo Databases	RFE + Taguchi Parameters Optimization (C and gamma)	Improved Accuracy. Computations time overhead
[3] Theerthagiri	Liver Disease Dataset	Recursive Gaussian Feature Selection	Improved Accuracy with Computational Overhead
[4] Javier Alcaraz	Genetic Data	Metaheuristic algorithm – Multiobjective Feature Selection	Good performance classification result Complexity in multiobjective optimization
[5] Vijayashree, J.	Heart disease	Improved PSO	Accuracy Improved Sensitivity to parameter tuning.
[6] Chandra Babu Gokulnath	Heart disease data	Optimized GA + SVM	Improved Accuracy Complexity in genetic algorithm design
[7] M. Sahaya Sheela, C. A. Arun	Computed Tomography Images-COVID	Hybrid - PCA and PSO (Particle Swam Optimization)	Improved Accuracy using SaaS cloud for huge volume of data. Complexity in hybrid algorithm Design
[8] Shaoze Cui	Diabetic data	Hybrid Feature Selection Algorithm	Accuracy rate increases in Hybrid Algorithms
[9] Vamsidhar Enireddy	COVID-19	Hybrid ResNet	Accuracy rate increases. Implementation Complexity
[10] Warda M. Shaban	COVID-19 Computed Tomography images	Hybrid = Fast Selection Stage (Filter) + Accurate Selection Stage (Genetic Algorithm – Wrapper)	Hybrid Algorithms improves Accuracy. Complexity in hybrid model implementation
[11] Behnood Gholami	Brain Tumor Spectrometry Images	RFE	Requires careful feature engineering
[12] Viktor Wottschelab	Brain MRI	RFE	Dependency on quality of MRI data

The table Table.1 provided portrays a comprehensive overview of feature selection algorithms applied across various medical datasets, highlighting their benefits and limitations. The analysis indicates that the choice of feature selection algorithm significantly impacts the accuracy and computational efficiency of medical data classification tasks. While filter, wrapper, and hybrid methods offer distinct benefits, hybrid approaches notably contribute to improved predictive accuracy and model efficiency [13]. The review also depicts that Recursive Feature Selection (RFE) is widely used with SVM giving better accuracy across various domains of health care.

II.II. Kernel Methods

Kernel methods have emerged as powerful tools in various domains for handling non-linear relationships within data and dataset with huge size to enhance the machine learning model performance. Sanz et al. proposed SVM-RFE, a method integrating Support Vector Machines (SVM) with Recursive Feature Elimination (RFE) to select and visualize the most relevant features using non-linear kernels [14]. This approach enhances model interpretability by iteratively refining feature sets. Alshanbari et al. introduced a weighted radial kernel SVM integrated with RFE for predicting and classifying COVID-19 admissions to ICU (Intensive Care Units) [15]. By leveraging informative features and weighted kernels, the model aids in resource allocation for healthcare management during the pandemic.

Patle and Chouhan discussed various SVM kernel functions and their applications in classification tasks, highlighting their versatility and effectiveness in different scenarios [16]. Additionally, Ben-Hur et al. explored the utility of SVMs and kernels in computational biology, showcasing their significance in analyzing biological data such as gene expression and protein sequences [17]. Liu et al. proposed a novel weighted SVM classification algorithm based on p-norm distance T kernel an improved polarization techniques [18]. This method enhances classification performance, particularly in scenarios with imbalanced class distributions. Similarly, Yanga et al. presented an adaptive parameter selection approach for Gaussian kernel SVMs based on the local density of the training set, improving model generalization across diverse datasets [19].

The table Table.2 shows the analysis of the kernel methods applied in diverse health care data prediction and classification. The table portrays that using kernel methods not only helps in improving the performance but also for feature selection, visualization, prediction, classification of imbalanced data. The study also reveals that kernel methods, especially the RBF kernel, enhance SVM classification models' flexibility and accuracy, making them indispensable tools in tackling distinct and complex data classification challenges.

Table.2. Analysis of SVM model with distinct Kernel methods

Article	Kernel Type	Advantages	Data Used
[14] SVM-RFE	Non-linear Kernel	Feature selection, Visualization	General data
[15] Weighted Radial Kernel SVM	Radial Basis Kernel	COVID-19 ICU prediction, Feature selection	COVID-19 ICU data
[16] SVM Kernel Functions	Distinct Methods	Versatility, Classification performance	General data
[17] SVMs for Computational Biology	Distinct Methods	Biological data analysis	Biological data
[18] Weighted p-norm Distance T Kernel SVM	T Kernel	Improved classification, Imbalanced data handling	General data
[19] Gaussian Kernel SVM	Gaussian Kernel	Adaptive parameter selection	General data

3. Implementation

In this paper, we use a COVID-19 dataset with 5,435 instances and 21 features. The 21 features include 20 symptoms and a target variable, COVID, which indicates whether the virus is present or not. The feature names with indices of the original features are mentioned in Table.3. Since the symptoms and target variables of the COVID-19 dataset are categorical, only feature selection techniques that support categorical data were used in this experiment. The chosen techniques include Mutual Information (Info Gain) and Chi-Square from the filter method, Sequential Feature Selection (SFS) and Recursive Feature Selection (RFE) from the wrapper method, and LASSO (L1 regularization) and Random Forest from the embedded method. Relevant features for COVID-19 prediction were identified using these techniques, and their computational efficiencies were recorded. For all feature selection techniques, the number of selected feature subsets is 12.

Table. 3. Feature Index with Names

Index	Feature Name	Index	Feature Name
0	Breathing Problem	10	Hyper Tension
1	Fever	11	Fatigue
2	Dry Cough	12	Gastrointestinal
3	Sore throat	13	Abroad travel
4	Running Nose	14	Contact with COVID Patient
5	Asthma	15	Attended Large Gathering
6	Chronic Lung Disease	16	Visited Public Exposed Places
7	Headache	17	Family working in Public Exposed Places
8	Heart Disease	18	Wearing Masks
9	Diabetes	19	Sanitization from Market

Table. 4. Comparison of Accuracy and Execution Time of Feature Selection Methods

Feature Selection Method	Accuracy	Execution Time in ms
Chi-Square	95.54%	30.98
Mutual Info Gain	92.65%	672.12
Sequential Feature Selection	95.84%	2495.97
Recursive Feature Selection	96.98%	63.22
Random Forest	96.82%	318.28
Lasso (L1 regularization)	95.98%	75.89

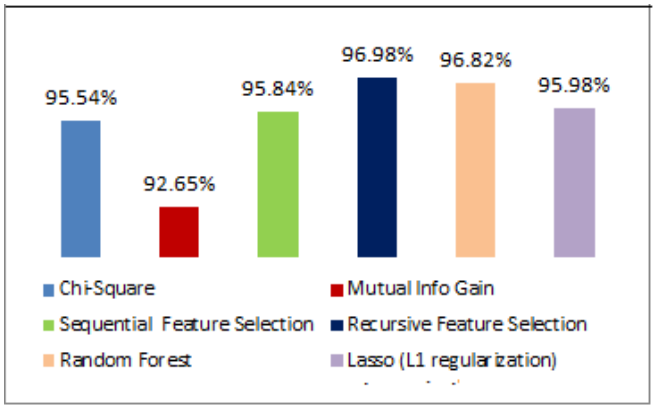

Fig.1. Comparison of Accuracy of Feature Selection Methods

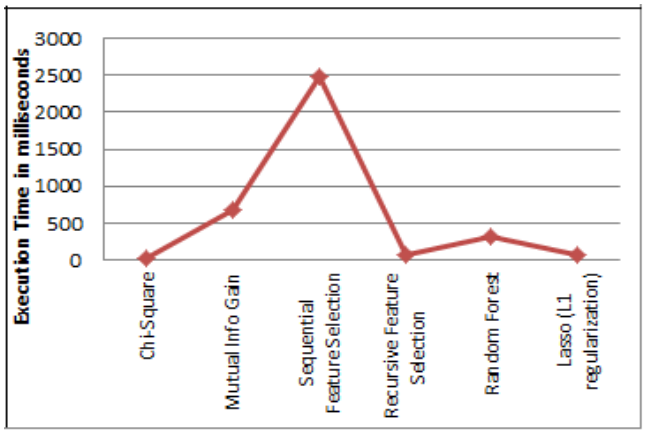

Fig.2. Comparison of Execution Time of Feature Selection Methods

The comparison table of accuracy and execution time for the feature selection methods is shown in Table 4, with graphical representations in Fig. 1 and Fig. 2. Additionally, the number of selected features, as well as the relevance, redundancy, and interpretability of features using different feature selection techniques, are analyzed and presented in Table 5 and Table 6. Table 7 presents the results of our SVM classification models using distinct kernel types.

Table. 5. Indices of Features Selected using different Feature Selection Techniques

Feature Selection Method Type	**Feature Selection Method**	**Indices of Features Selected (for the input num_features_selected = 12)**
Filter Method	Chi-Square	[0 1 2 3 5 6 10 13 14 15 16 17]
	Mutual Info Gain	[0 1 2 3 9 10 11 13 14 15 16 19]
Wrapper Method	Sequential Feature Selection	[1 3 4 5 6 7 8 10 12 13 15 19]
	Recursive Feature Selection	[0 1 2 4 5 6 10 13 14 15 16 17]
Embedded Method	Random Forest	[0 1 2 3 4 6 8 13 14 15 16 17]
	Lasso (L1 regularization)	[1 2 3 4 5 6 7 8 9 10 18 19]

Table. 6. Comparative Analysis Feature Selection Techniques

Algorithm	Relevance	Redundancy	Interpretability
Chi-Square	Moderate	Low	Low
Mutual Info Gain	High	Low	Medium
Sequential Feature Selection (SFS)	Moderate	Low	High
Recursive Feature Selection (RFE)	High	Low	High
Random Forest	High	Low	Moderate
Lasso (L1 Regularization)	High	Moderate	High

Table. 7. Comparison of SVM classification models using different kernel types

Kernel Type	Accuracy	Precision	Recall	Time Taken in ms
Linear	95.40	97.79	96.29	212.45
Rbf	97.57	98.37	98.64	267.02
Sigmoid	74.98	85.21	83.83	666.18
Poly	96.35	97.37	97.37	129.73

4. Discussion

The outcomes in Table 4, Fig. 1, and Fig. 2 show that RFE and Random Forest techniques yield high accuracy rates of 96.98% and 96.82%, respectively. Meanwhile, Lasso, RFE, and Chi-Square techniques are more time-efficient, with Chi-Square being the fastest at 30.98 milliseconds. RFE achieves high accuracy in a relatively short time of 63.22 milliseconds.

The indices list in Table 5 indicates that most methods either missed relevant features like *breathing problems* and *heart disease* or included irrelevant ones such as *gastrointestinal issues* and *sanitization from market*. Among the methods, the filter method (Chi-Square and RFE) identified the most relevant features. Table 6 compares various feature selection techniques based on relevance, redundancy, and interpretability, highlighting the strengths and weaknesses of each. Techniques like Mutual Information Gain, Recursive Feature Selection, and Lasso exhibit high relevance and interpretability, while RFE balances low redundancy with high interpretability, making it particularly suitable for feature selection.

Table. 7. Comparison depicts that the Radial Basis Function (RBF) kernel demonstrated the best overall performance, achieving an accuracy of 97.57%, precision of 98.37%, and recall of 98.64%, indicating its effectiveness in classification. The polynomial kernel followed closely with an accuracy of 96.35%, precision of 97.37%, and recall of 97.37%. The linear kernel, while slightly less accurate, completed classification tasks fastest at 212.45 milliseconds. In contrast, the sigmoid kernel had lower performance metrics and the longest processing time of 666.18 milliseconds. These results emphasize the importance of balancing accuracy and computational efficiency when selecting a kernel type.

The analysis indicates that Chi-Square in filter methods reduces computational time while maintaining accuracy, RFE in wrapper methods offers the most relevant feature sets, and embedded methods provide a balanced trade-off. Among kernel methods, the RBF kernel consistently outperforms others in classification accuracy and robustness, enhancing SVM model performance when combined with appropriate feature selection techniques.

5. Conclusion and Future Scope

In conclusion, this review article provides a comprehensive overview of feature selection algorithms in machine learning, particularly for SVM classification models. By categorizing and discussing filter, wrapper, and embedded methods, we offer insights into their strengths, weaknesses, and applications across various domains. This research indicates that the Chi-Square method from the filter approach and the Recursive Feature Elimination (RFE) technique from the wrapper method produce better results than other techniques. The analysis underscores that while advanced and hybrid feature selection algorithms can significantly improve the accuracy of medical data classification, they often come with increased computational complexity and implementation challenges. Using kernel methods reduces computational complexity. Among the evaluated kernel methods, the RBF kernel consistently outperforms others in classification accuracy and robustness. Therefore, integrating Chi-Square and RFE feature selection techniques with the optimal kernel method, such as the RBF kernel, significantly enhances the performance of the SVM classification model.

Future research could develop efficient hybrid feature selection algorithms that balance accuracy and computational complexity. Integrating feature selection with deep learning and adaptive AI techniques holds promise for large-scale datasets. Exploring these methods across various fields beyond medical data classification will further validate their versatility and robustness.

References

[1] Hamed Mazreati, Reza Radfar, Mohammad-Reza Sohrabi, Babak Sabet Divshali, Mohammad Ali Afshar Kazemi, "*Comparing the Performance of Feature Selection Methods for Predicting Gastric Cancer*", International Journal of Cancer Management, Vol.**16**, Issue **1**, **2023**.

[2] Mei-Ling Huang, Yung-Hsiang Hung, W. M. Lee, R. K. Li and Bo-Ru Jiang, "*SVM-RFE Based Feature Selection and Taguchi Parameters Optimization for Multiclass SVM Classifier*", The Scientific World Journal, Hindawi, **2014**.

[3] Prasannavenkatesan Theerthagiri & Sahana Devarayapattana Siddalingaiah, "*Recursive Gaussian support vector machine based feature selection algorithm for liver disease classification*", Multimedia Tools and Applications, Springer Link, **2023**.

[4] Javier Alcaraz, Martine Labbé and Mercedes Landete, "*Support Vector Machine with feature selection: A multiobjective approach*", Elsevier, Vol.**204**, **2022**.

[5] J. Vijayashree & H. Parveen Sultana, "*A Machine Learning Framework for Feature Selection in Heart Disease Classification Using Improved Particle Swarm Optimization with Support Vector Machine Classifier*", Journal of Russian Law, **2019**.

[6] Chandra Babu Gokulnath, S.P. Shantharajah, "*An optimized feature selection based on genetic approach and support vector machine for heart disease"*, Springer link, Vol.**22**, pp.**14777-14787**, **2018**.

[7] M. Sahaya Sheelal, C. A. Arun, "*Hybrid PSO–SVM algorithm for Covid-19 screening and quantification", International Journal of Information Technology*", Springer Link, Vol.**14**, pp.**2049-2056**, **2022**.

[8] Shaoze Cui, Dujuan Wang, Yanzhang Wang, Pay-Wen Yu, Yaochu Jin, "*An improved support vector machine-based diabetic readmission prediction*", Elsevier, Vol.**166**, pp.**123-135**, **2018**.
[9] Vamsidhar Enireddy, Mathe John Kenny Kumar, Babitha Donepudi and C Karthikeyan, "*Detection of COVID-19 using Hybrid ResNet and SVM*", IOP Science, **2020**.
[10] Warda M. Shaban, Asmaa hamdy Rabie, Ahmed Ibrahim Saleh, Mohy Eldin A. Abo-Elsoud, "*A new COVID-19 Patients Detection Strategy (CPDS) based on hybrid feature selection and enhanced KNN classifier*", Knowledge Based Systems, Research Gate, **2020**.
[11] Behnood Gholami, saiah Norton, Allen R. Tannenbaum, and Nathalie Y. R. Agar, "*Recursive Feature Elimination for Brain Tumor Classification using Desorption Electrospray Ionization Mass Spectrometry Imaging*", PubMed, National Library of Medicine, **2012**.
[12] Viktor Wottschela, Declan T. Chard , Christian Enzingerd , Massimo Filippie, "*SVM recursive feature elimination analyses of structural brain MRI predicts near-term relapses in patients with clinically isolated syndromes suggestive of multiple sclerosis*", Multicenter Study, PubMed, **2019**.
[13] Irfan Ullah Khattak, Aihab Khan, Farhan Hassan Khan, Abdullah Gani, Muhammad Shiraz, "*A Novel Feature Selection Method for Classification of Medical Data using Filters, Wrappers and Embedded Approaches*", Hindawi, Vol.2022, **2022**.
[14] Hector Sanz , Clarissa Valim, Esteban Vegas , Josep M. Oller and Ferran Reverter, "*SVM-RFE: selection and visualization of the most relevant features through non-linear kernels*", BMC Bioinformatics, Nov. **2018**.
[15] Huda M. Alshanbari, Tahir Mehmood, Waqas Sami, Wael Alturaiki, Mauawia A. Hamza and Bandar Alosaimi, "*Prediction and Classification of COVID-19 Admissions to Intensive Care Units (ICU) Using Weighted Radial Kernel SVM Coupled with Recursive Feature Elimination (RFE)*", PubMed, National Library of Medicine, **2022**.
[16] Arti Patle; Deepak Singh Chouhan, "SVM kernel functions for classification", 2013 International Conference on Advances in Technology and Engineering (ICATE), IEEE, **2013**.
[17] Asa Ben-Hur, Cheng Soon Ong, Sören Sonnenburg, Bernhard Schölkopf, Gunnar Rätsch, "*Support Vector Machines and Kernels for Computational Biology*", PubMed, National Library of Medicine, **2008**,
[18] Wenbo Liu, Shengnan Liang, Xiwen Qin, "*Weighted p-norm distance t kernel SVM classification algorithm based on improved polarization*", Scientific Reports, **2022**.
[19] Jiawei Yanga, Zeping Wua,KePenga, Patrick N. Okolob,c, Weihua Zhanga,Hailong Zhaoaand Jingbo Sun, "*Parameter selection of Gaussian kernel SVM based on local density of training set*", Inverse Problem in Science and Engineering, Taylor & Francis", Vol.**29**, No.**4**, pp.**536-548**, **2021**.
[20] C. Dharmadevi and S. Thaddeus, "*Prediction of COVID-19 Infections using Classification Algorithms in Machine Learning*", Indian Journal of Natural Sciences, Vol.**24**, pp.**61005-61011**, **2023**.

