Journal of Soft Computing Paradigm (ISSN: 2582-2640)
www.irojournals.com/jscp/

Robust Breast Cancer Prognosis Prediction: Adaptive Outlier Removal using SVM and K-Means Clustering

S. Vivekanandan[1], S. Mounika[2], P. Monisha[3], M. Balaganesh[4]

[1]Assistant Professor, Department of Computer Science and Engineering, Velalar College of Engineering and Technology, Anna University, Erode, India

[2,3,4]Student, Department of Computer Science and Engineering, Velalar College of Engineering and Technology, Anna University, Erode, India

E-mail: [1]erodeviveks@gmail.com, [2]mounikasenthil1010@gmail.com, [3]monishaperiyasamy79@gmail.com, [4]Balag8563@gmail.com

Abstract

Analyzing several datasets is essential to breast cancer research in order to find trends and prognostic markers. For this reason, the Wisconsin Prognostic Breast Cancer (WPBC) dataset offers a valuable source of data. Outliers, however, have the potential to seriously affect how accurate predictive models are. This work suggests using the Support Vector Machine (SVM) algorithm in an adaptive outlier removal method to improve the resilience of prediction models that were trained on the WPBC dataset. To ensure optimum SVM performance, the technique includes pre-processing processes, including addressing missing data and standardizing features. Tailored elimination of outliers is made possible by their dynamic identification, depending on how they deviate from the support of the SVM model. To increase generalization, the SVM is then retrained using the outlier-adjusted dataset. Test set evaluation shows the effectiveness of the method with improved F1-score, recall, and accuracy. With datasets similar to WPBC, this adaptive outlier elimination technique offers a useful tool for improving breast cancer prediction models, leading to increased model performance and dependability in prognostic tasks.

Extended version of article received from 4th INTERNATIONAL CONFERENCE on Artificial Intelligence,5G Communication and Network Technologies
DOI: https://doi.org/10.36548/jscp.2024.1.007
Received: 04.03.2024, received in revised form: 01.04.2024, accepted: 13.04.2024, published: 30.04.2024

Keywords: Adaptive Outlier Removal, Machine Learning, Breast Cancer, Prognostic Factors.

1. Introduction

Analyzing a variety of datasets is essential for identifying patterns and prognostic variables in the field of breast cancer research, which is critical for enhancing therapeutic results [3]. One notable resource is the WPBC dataset, which provides an extensive information repository. However, the existence of outliers may have a substantial effect on the prediction models' performance that are used in these kinds of investigations. This work uses the Support Vector Machine (SVM) algorithm to suggest a novel method for outlier elimination in answer to this difficulty. This adaptive technique tries to improve the robustness of prediction models trained on the WPBC dataset by dynamically detecting outliers based on their divergence from the support of the SVM model [4-7]. To guarantee the best possible SVM performance, the technique includes all necessary pre-processing procedures, such as addressing missing data and feature standardization. The work attempts to enhance model generalization by retraining the SVM on the outlier-adjusted dataset afterwards. This method's efficacy is thoroughly assessed on a test set, demonstrating improved F1-score, recall, and accuracy. This adaptive outlier elimination method shows promise as a useful tool for improving breast cancer prediction models' accuracy and dependability in prognostic tasks [8-10].

2. Related Work

The related work section presents a short review on the methods used in detecting and eliminating the outliers in different application applying the machine learning.

In this study, [1] the authors proposed using Independent Component Analysis (ICA) as a blind source separation method. They developed ICA Mixture Models (ICAMMs), which are two-layer neural networks, to handle dependencies between sources. They introduced a novel metric called Probabilistic Distance (PDI) to measure differences between ICAMMs, which outperformed traditional metrics like KLD in detecting changes over time. They tested their approach on material defect detection and EEG signal analysis, demonstrating its effectiveness. ICAMM offers flexibility in modeling various probability density functions and

can aid in data mining and pattern identification tasks. The PDI facilitates model differentiation and has wide applicability in various domains.

In this research, [2] the authors addressed the issue of k-means clustering in the presence of outliers. They proposed a straightforward technique based on local searches, allowing the exclusion of a limited number of outliers. Their method aims to minimize the variance of points within each cluster. They demonstrated scalability to large datasets and achieved constant-factor approximation solutions when combined with sketching techniques. Empirical evaluations on real-world and synthetic data showed superior performance compared to existing heuristic approaches. Clustering involves grouping similar points into k clusters, aiming to minimize the sum of squared distances from each point to its nearest cluster center. Despite being NP-hard, k-means clustering remains a widely studied challenge in data mining.

3. System Specification (Tools)

3.1 Hardware Requirements

- Processor Type : AMD RYZEN 7
- Speed : 4.40GHZ
- RAM : 16 GB RAM
- Hard disk : 1 TB
- Keyboard : 101/102 Standard Keys
- Mouse : Optical Mouse

3.2 Software Specification

- Operating System: Windows 10
- Front End: MATLAB
- Back End: JAVA

4. Proposed Work

4.1 System Workflow

Dynamic and complex, adaptive outlier elimination has become a key strategy for improving the precision and dependability of prediction models on a variety of datasets. Outliers, or data items that dramatically differ from the norm, may have a considerable impact on model performance in the context of data analysis. The subtleties of complex datasets may be too complicated for the traditional one-size-fits-all techniques for outlier removal to handle. By using sophisticated algorithms, such the Support Vector Machine (SVM), to dynamically detect and remove outliers based on their divergence from the underlying data distribution, adaptive outlier removal marks a paradigm leap in data processing. With a customized and nuanced approach, our proactive and data-driven technique makes sure that the outlier elimination procedure closely matches the dataset's features. Therefore, adaptive outlier elimination is a crucial technique in fields like breast cancer research where accuracy and dependability are crucial. It not only improves the resilience of prediction models but also helps with better generalization. Machine Learning revolutionizes problem-solving by enabling computers to learn from data, find hidden connections, and improve decision-making without explicit programming. Breast cancer, a complex disease with various subtypes, demands early identification and precise prognostication for effective treatment planning. Continuous research using datasets like WPBC is crucial for enhancing prediction models and improving detection, diagnosis, and treatment methods. Prognostic variables, including pathological, clinical, and molecular indicators, guide treatment planning and improve patient outcomes by forecasting disease progression and treatment efficacy. The Figure : 1 depicts the workflow overview.

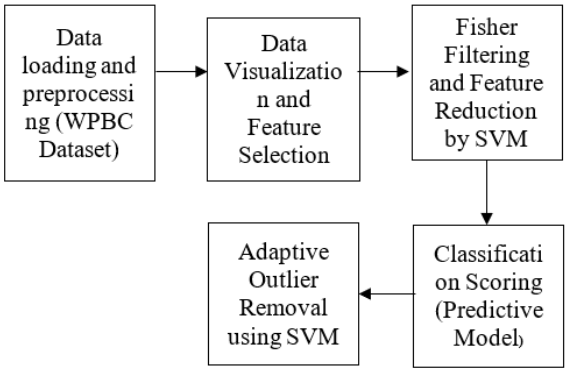

Figure 1. Workflow Overview

4.2 Overview of the Dataset

With an emphasis on the WPBC dataset, the suggested method uses an adaptive outlier removal strategy using Support Vector Machines (SVM) to improve the robustness and reliability of the analysis. The WPBC dataset is first loaded into the system, and any required preparation operations such as managing missing values and encoding categorical variables are then carried out. After dividing the data into training and testing sets, standardization is performed to guarantee that feature values scale consistently. The standardized training data is used to train the SVM model, which enables the system to recognize outliers using the patterns it has learnt. The training and testing sets are then cleared of these outliers. The suggested approach makes sure that the model is trained on a more reliable and representative dataset in addition to offering a technique for adaptively detecting and removing outliers. In situations where the existence of outliers might have a major influence on the efficacy of predictive models for breast cancer prediction, this method helps to increase model performance and generalization. The Figure: 2 depicts the dataset selection and the overview of the dataset is illustrated in Figure.3

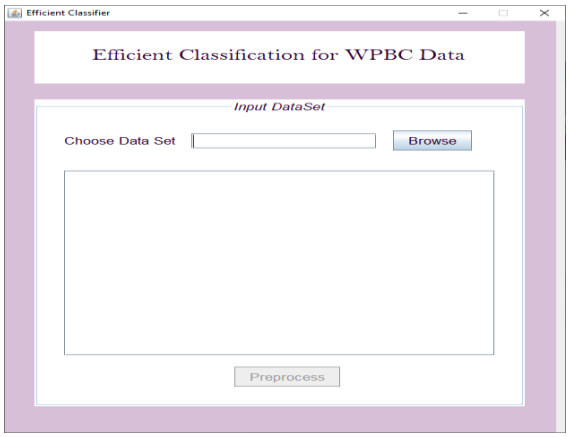

Figure 2. Dataset Selection

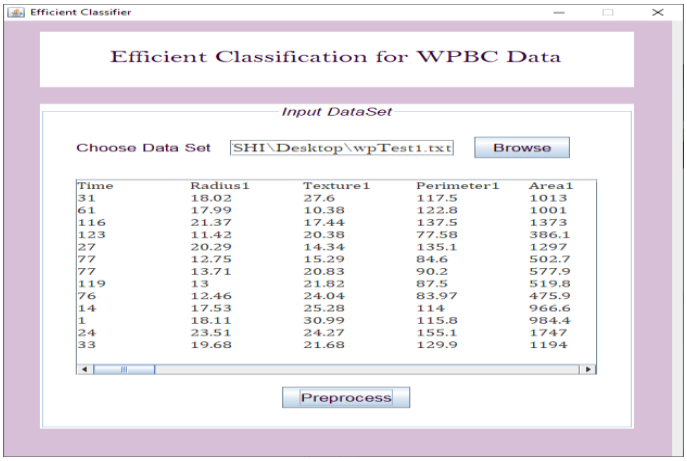

Figure 3. Overview of the Dataset

4.2.1 Sample Dataset (Wisconsin Prognostic Breast Cancer Dataset (WPBC))

The Figure .4 and 5 shows the sample dataset and the classification results

Time	Radius1	Texture1	Perimeter1	Area1	Smoothness1	Compactness1	Concavity1	ConcavePoints1	Symmetry1	FractalDim1	Radius2	Texture2	Perimeter2
	Area2	Smoothness2	Compactness2	Concavity2	ConcavePoints2	Symmetry2	FractalDim2	Radius3	Texture3	Perimeter3	Area3	Smoothness3	Compactness3
	Concavity3	ConcavePoints3	Symmetry3	FractalDim3	Tumor	Lymphnode							
31	18.02	27.6	117.5	1013	0.09489	0.1036	0.1086	0.07055	0.1865	0.06333	0.6249	1.89	3.972
	71.55	0.004433	0.01421	0.03233	0.009854	0.01694	0.003495						
	21.63	37.08	139.7	1436	0.1195	0.1926	0.314	0.117	0.2677	0.08113	5	5	
61	17.99	10.38	122.8	1001	0.1184	0.2776	0.3001	0.1471	0.2419	0.07871	1.095	0.9053	8.589
	153.4	0.006399	0.04904	0.05373	0.01587	0.03003	0.006193	25.38					
	17.33	184.6	2019	0.1622	0.6656	0.7119	0.2654	0.4601	0.1189	3	2		
116	21.37	17.44	137.5	1373	0.08836	0.1189	0.1255	0.0818	0.2333	0.0601	0.5854	0.6105	3.928
	82.15	0.006167	0.03449	0.033	0.01805	0.03094	0.005039	24.9					
	20.98	159.1	1949	0.1188	0.3449	0.3414	0.2032	0.4334	0.09067	2.5	0		
123	11.42	20.38	77.58	386.1	0.1425	0.2839	0.2414	0.1052	0.2597	0.09744	0.4956	1.156	3.445
	27.23	0.00911	0.07458	0.05661	0.01867	0.05963	0.009208	14.91	26.5				
	98.87	567.7	0.2098	0.8663	0.6869	0.2575	0.6638	0.173	2	0			
27	20.29	14.34	135.1	1297	0.1003	0.1328	0.198	0.1043	0.1809	0.05883	0.7572	0.7813	5.438
	94.44	0.01149	0.02461	0.05688	0.01885	0.01756	0.005115	22.54	16.67				
	152.2	1575	0.1374	0.205	0.4	0.1625	0.2364	0.07678	3.5	0			
77	12.75	15.29	84.6	502.7	0.1189	0.1569	0.1664	0.07666	0.1995	0.07164	0.3877	0.7402	2.999
	30.85	0.007775	0.02987	0.04561	0.01357	0.01774	0.005114	15.51					
	20.37	107.3	733.2	0.1706	0.4196	0.5999	0.1709	0.3485	0.1179	2.5	0		
77	13.71	20.83	90.2	577.9	0.1189	0.1645	0.09366	0.05985	0.2196	0.07451	0.5835	1.377	3.856
	50.96	0.008805	0.03029	0.02488	0.01448	0.01486	0.005412	17.06					
	28.14	110.6	897	0.1654	0.3682	0.2678	0.1556	0.3196	0.1151	4	10		
119	13	21.82	87.5	519.8	0.1273	0.1932	0.1859	0.09353	0.235	0.07389	0.3063	1.002	2.406
	24.32	0.005731	0.03502	0.03553	0.01226	0.02143	0.003749	15.49					
	30.73	106.2	739.3	0.1703	0.5401	0.539	0.206	0.4378	0.1072	2	1		
76	12.46	24.04	83.97	475.9	0.1186	0.2396	0.2273	0.08543	0.203	0.08243	0.2976	1.599	2.039
	23.94	0.007149	0.07217	0.07743	0.01432	0.01789	0.01008	15.09	40.68				
	97.65	711.4	0.1853	1.058	1.105	0.221	0.4366	0.2075	6	20			
14	17.53	25.28	114	966.6	0.09278	0.09175	0.1105	0.06741	0.1424	0.05563	0.4587	1.318	3.032
	52.27	0.00454	0.01312	0.0172	0.007299	0.009783	0.001526						
	21.75	40.14	142.6	1483	0.1287	0.2472	0.2753	0.1372	0.2404	0.07156	10	9	
1	18.11	30.99	115.8	984.4	0.08625	0.0924	0.06214	0.05598	0.1603	0.05468	0.4236	1.176	2.189
	36.01	0.004733	0.01621	0.02044	0.01142	0.01411	0.002098	21.05					
	38.93	128	1214	0.1194	0.2088	0.2385	0.1333	0.2652	0.07006	2.7	4		
24	23.51	24.27	155.1	1747	0.1069	0.1283	0.2308	0.141	0.1797	0.05506	1.009	0.9245	6.462
	164.1	0.006292	0.01971	0.03582	0.01301	0.01479	0.003118	30.67					
	30.73	202.4	2906	0.1515	0.2678	0.4819	0.2089	0.2593	0.07738	3.5	4		
33	19.68	21.68	129.9	1194	0.09797	0.1339	0.1863	0.1103	0.2082	0.05715	0.6226	2.284	5.173
	67.66	0.004756	0.03368	0.04345	0.01806	0.03756	0.003288	22.75					
	34.66	157.6	1540	0.1218	0.3458	0.4734	0.2255	0.4045	0.07918	1.5	13		

Figure 4. Sample Dataset

4.2.2 Result

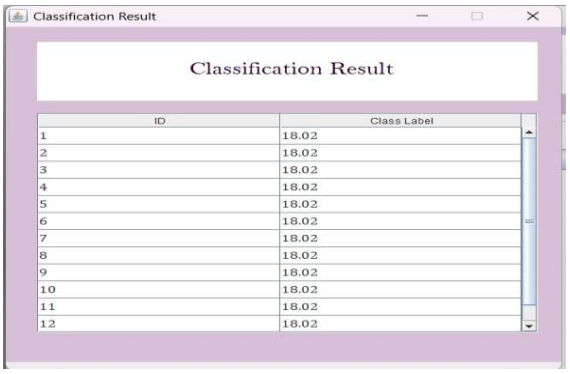

Figure 5. Classification Result

4.3 Methodology

A. Data Visualization

In the proposed system, the initial step involves loading the Wisconsin Prognostic Breast Cancer (WPBC) dataset and performing essential pre-processing steps. The WPBC dataset is acquired from the UCI Machine Learning Repository website and saved in text format. Subsequently, the dataset is imported into an Excel spreadsheet, with the values organized under corresponding attribute column headers. This import process encompasses handling missing data and encoding categorical variables as part of the preparation. Before delving into adaptive outlier removal, the system emphasizes the importance of visualizing the data to gain insights and conducting pre-processing to ensure the dataset is suitable for subsequent analysis

B. Data Pre- Processing

The WPBC dataset must be loaded into the suggested system, and necessary pre-processing must be carried out. You may save the WPBC dataset as a text file after downloading it from the UCI Machine Learning Repository website. The values are then stored with the matching characteristics as column headers once this file is imported into an Excel spreadsheet. This covers encoding categorical variables and dealing with missing data. The approach highlights the significance of pre-processing to make sure the dataset is ready for further analysis and displaying the data to get insights before diving into adaptive outlier elimination. The Figure.6 shows the results of data pre-processing.

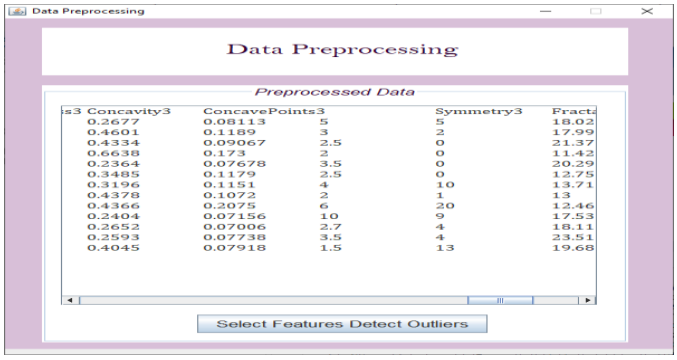

Figure 6. Data Pre-Processing

C. Feature Selection Algorithms

Fisher filtering is used in the feature selection process to highlight each feature's capacity for discrimination. Feature selection algorithms are included into the data processing pipeline to determine and preserve the most relevant characteristics for the prognosis of breast cancer. These techniques help concentrate attention on the characteristics that make a substantial contribution to the prediction model while lowering dimensionality and focusing on the features that contribute significantly to the predictive model. The Figure.7 illustrates the feature selection algorithm.

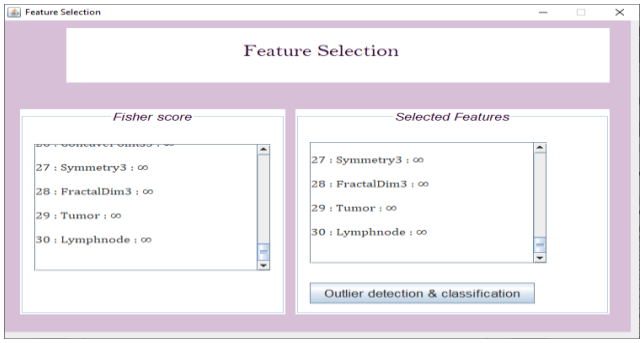

Figure 7. Feature Selection Algorithm

5. System Analysis

5.1 Existing System

The existing system utilizes the K-Mean algorithm based on Turkey's rule in conjunction with new distance metric. The disadvantages of the existing method is listed below.

1. The proposed modification based on Tukey' s rule for outlier removal may be sensitive to the choice of parameters, impacting its adaptability to different datasets.

2. The adaptive outlier elimination process adds computational complexity to the k-means algorithm, potentially slowing down the clustering process.

3. The effectiveness of the new distance metric in improving clustering accuracy may be dataset -dependent, limiting its general applicability.

4. The overall performance gains of up to 80.57% in clustering accuracy may not be consistently achievable across diverse datasets, raising concerns about the algorithm's robustness.

5.2 Proposed System

In the proposed system, an adaptive outlier removal technique using Support Vector Machines (SVM) is employed for enhancing the robustness and reliability of the analysis, with a focus on the WPBC dataset. The system begins by loading the WPBC dataset and conducting necessary preprocessing steps, including handling missing values and encoding categorical variables. The data is then split into training and testing sets, followed by standardization to ensure consistent scaling of feature values. The SVM model is trained on the standardized training data, allowing the system to identify outliers based on the learned patterns. These outliers are subsequently removed from both the training and testing sets. The proposed system not only provides a mechanism for adaptively identifying and eliminating outliers but also ensures that the model is trained on a more robust and representative dataset. This approach contributes to the improvement of model generalization and performance, particularly in scenarios where the presence of outliers can significantly impact the effectiveness of predictive models for breast cancer prognosis. The advantages of the proposed system are listed below.

- Adaptive outlier removal using SVM enhances model robustness by identifying and eliminating outliers, ensuring a more reliable and representative training dataset.
- The system contributes to improved model generalization by mitigating the influence of outliers, leading to better performance on unseen data.
- Removal of outliers from the Wisconsin Prognostic Breast Cancer dataset improves the accuracy of diagnostic predictions, supporting more accurate breast cancer prognosis.

6. The Method and Metrics Used in Evaluating the Performance of The SVM and K - Means

The method combines SVM and K-means for breast cancer prognosis prediction while addressing outliers using adaptive outlier removal:

1. SVM (Support Vector Machines):

Method: SVM is a supervised learning algorithm used for classification tasks. It works by finding the hyperplane that best separates the data points into different classes while maximizing the margin between the classes.

2. K-Means:

Method: K-means clustering is an unsupervised learning algorithm used for clustering tasks. It aims to partition the data into K clusters, where each data point belongs to the cluster with the nearest mean.

3. Adaptive Outlier Removal:

Method: The paper proposes an adaptive outlier removal technique to enhance the performance of K-means clustering. This technique aims to iteratively remove outliers from the data based on their distances from cluster centroids, thereby improving the robustness of the clustering results.

4. Performance Evaluation:

Cross-Validation: Both SVM and K-means clustering models may undergo cross-validation to assess their performance robustness. This involves splitting the dataset into training and testing subsets multiple times to ensure reliable estimates of performance metrics.

Comparison: The performance of SVM and K-means models, with and without adaptive outlier removal, can be compared using appropriate evaluation metrics. This comparison helps determine the effectiveness of the proposed method in enhancing breast cancer prognosis prediction.

7. SVM and K-Means Clustering

- This approach combines the strengths of SVM and K-means while addressing their limitations.
- SVM is used to classify data points and identify potential outliers, which are then removed from the dataset.
- Evaluation metrics for both classification (SVM) and clustering (K-means) are used to assess the performance of the combined approach.
- Adaptive outlier removal helps enhance the robustness of K-means clustering by iteratively refining the dataset based on SVM predictions.
- The combined approach aims to improve breast cancer prognosis prediction by leveraging both supervised and unsupervised learning techniques while mitigating the impact of outliers and noise. The Figure.8 illustrates the fisher filtering applied for feature selection.

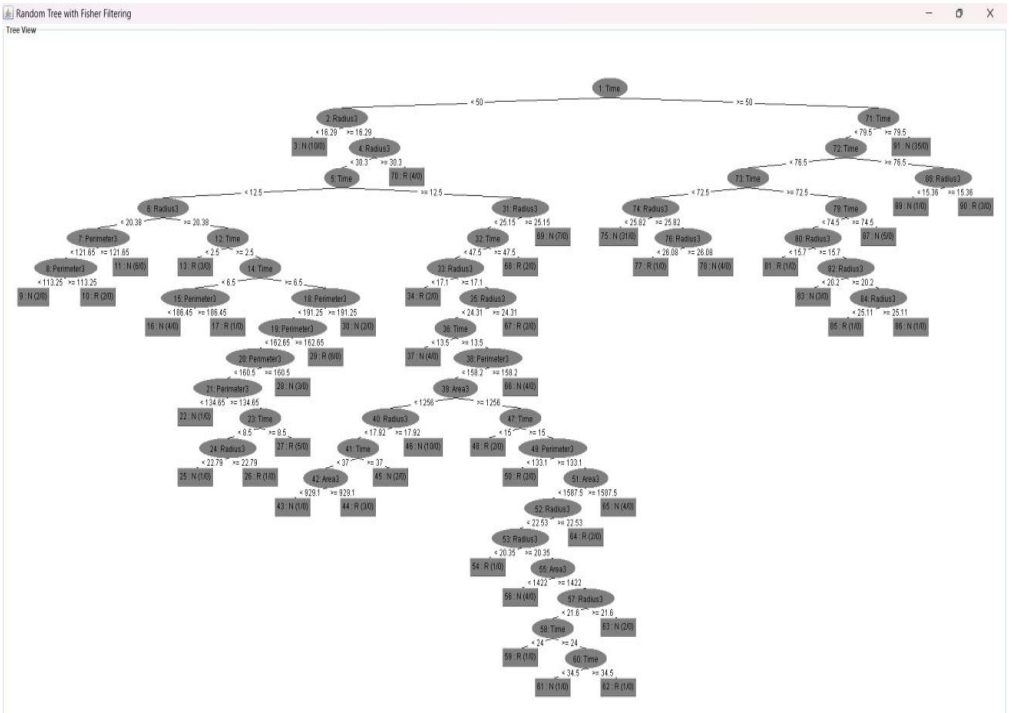

Figure 8. Fisher Filtering

8. Results and Discussions

With an accuracy of 88%, the Support Vector Machine (SVM) method outperforms the K-means algorithm, which only manages 75% accuracy. While SVM is a supervised learning method used for classification problems and is renowned for its ability to handle complicated decision boundaries, K-means is a clustering technique that divides data into K groups based on similarity. Because SVM can locate the best separating hyper planes in high-dimensional spaces, it is a recommended method for classification problems with well-defined borders. Its greater accuracy means that it performs better when categorizing data points. Table 1 presents a comparison of the accuracy achieved by both K-Means and SVM, while Figure 9 provides a graphical representation of the same comparison.

Table 1. Comparison Table

Algorithm	Accuracy
K means	75
SVM	88

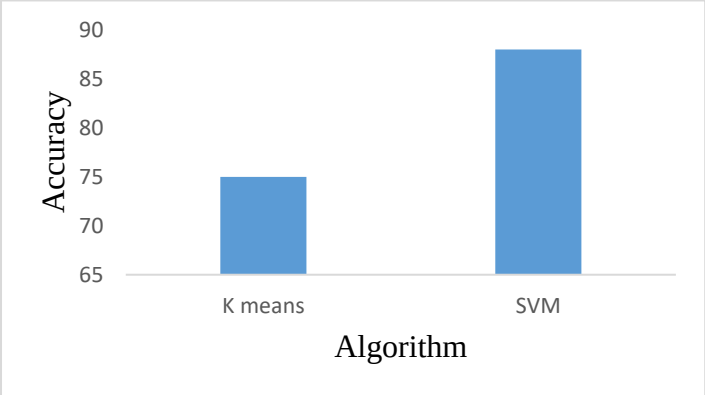

Figure 9. Comparison Graph

9. Conclusion

To sum up, the suggested approach is a big step forward in using cutting-edge technology to meet particular needs or obstacles. With careful planning, thorough testing, and cautious execution, the system is ready to provide noticeable advantages inside the desired

operating environment. The result of these efforts guarantees that the system meets user expectations, accomplishes its intended function, and performs dependably in practical situations. The end of one stage of development heralds the start of an iterative process that offers chances for ongoing adaptation and improvement as technology advances. In the end, the system's effectiveness depends on its capacity to improve productivity, simplify procedures, and provide insightful information or services.

References

[1] Hoxha, Genc, Farid Melgani, and Jacopo Slaghenauffi. "A new CNN-RNN framework for remote sensing image captioning." In 2020 Mediterranean and Middle-East Geoscience and Remote Sensing Symposium (M2GARSS), pp. 1-4. IEEE, 2020.

[2] Yu, Niange, Xiaolin Hu, Binheng Song, Jian Yang, and Jianwei Zhang. "Topic-oriented image captioning based on order-embedding." IEEE Transactions on Image Processing 28, no. 6 (2018): 2743-2754.

[3] Lo, Owen, William J. Buchanan, Paul Griffiths, and Richard Macfarlane. "Distance measurement methods for improved insider threat detection." Security and Communication Networks 2018 (2018): 1-18.

[4] Cegielski, Andrzej. "Bibliography on the Kaczmarz method." Journal of Mathematical Analysis and Applications 343 (2008): 427-435.

[5] " Safont, Gonzalo, Addisson Salazar, Luis Vergara, Enriqueta Gomez, and Vicente Villanueva. "Probabilistic distance for mixtures of independent component analyzers." IEEE Transactions on Neural Networks and Learning Systems 29, no. 4 (2017): 1161-1173.

[6] Walker, Shalika, Waqas Khan, Katarina Katic, Wim Maassen, and Wim Zeiler. "Accuracy of different machine learning algorithms and added-value of predicting aggregated-level energy performance of commercial buildings." Energy and Buildings 209 (2020): 109705.

[7] Yin, Shizhuang, and Tao Wang. "An unknown Protocol improved k-means clustering algorithm based on Pearson distance." Journal of Intelligent & Fuzzy Systems 38, no. 4 (2020): 4901-4913.

[8] Shrifan, Nawaf HMM, Ghassan Nihad Jawad, Nor Ashidi Mat Isa, and Muhammad Firdaus Akbar. "Microwave nondestructive testing for defect detection in composites based on K-means clustering algorithm." IEEE Access 9 (2020): 4820-4828.

[9] "Evolutionary static and dynamic clustering methods based on multi-verse optimizer," J. Chen and H. Zhuge, "2019 15th International Conference on Semantics, Knowledge and Grids (SKG), Guangzhou, China, 2020, pp. 123-126, doi: 10.1109/SKG49510.2019.00029.

[10] " Zhang, Mingxing, Yang Yang, Hanwang Zhang, Yanli Ji, Heng Tao Shen, and Tat-Seng Chua. "More is better: Precise and detailed image captioning using online positive recall and missing concepts mining." IEEE Transactions on Image Processing 28, no. 1 (2018): 32-44.

Author's Biography

Mr.S.Vivekanandan.,ME., Assistant Professor, Department Of Computer Science And Engineering, Velalar College Of Engineering And Technology, Anna University, Erode, India.

Ms. S.Mounika, Student, Department Of Computer Science And Engineering, Velalar College Of Engineering And Technology, Anna University, Erode, India.

Ms. P. Monisha, Student, Department of Computer Science And Engineering, Velalar College Of Engineering And Technology, Anna University, Erode, India.

Mr. M. Balaganesh, Student, Department of Computer Science And Engineering, Velalar College Of Engineering And Technology, Anna University, Erode, India.