Journal of Intelligent Systems and Internet of Things
Vol. 16, No. 01, PP. 41-48, 2025
DOI: https://doi.org/10.54216/JISIoT.160104

Feature Selection and Stability Analysis using Ensemble Techniques

Dipti Theng[1,*], K. K. Bhoyar[2], Prashant Pawade[3]
[1]Department of Computer Science and Engineering, Symbiosis Institute of Technology Pune, Symbiosis International (Deemed University), Pune, India
[2]Department of Computer Science and Engineering YCCE Nagpur, India
[3]Department of Civil Engineering GHRCE Nagpur, India
Emails: deepti.theng@gmail.com; kkbhoyar@gmail.com; prashant.pawade@raisoni.net

Abstract

Selecting the most relevant feature subset for a task is demanded and recommended for high accuracy and reduced model training time. Ensemble learning has shown superior results in classification; hence, we propose an ensemble method for feature selection and shown stability analysis for the selected feature set. The research question being investigated is whether ensemble methods are effective at selecting informative features in a dataset and if the selected features are stable compared to other feature selection methods. This paper presented a tree-based ensemble learning approach for feature selection. Our approach for ensemble feature selection includes function perturbation with the voting ensemble, an ensemble with a fixed number of features, and an ensemble with a contiguous number of features. Ensemble learning is found to be superior to other traditional feature selection algorithms. Ensemble learning algorithms are implemented on two high-dimensional microarray biomedical datasets. From our experimental study, it is observed that the voting ensemble outperforms other ensemble techniques, thereby reducing feature subset size and achieving higher accuracy. Stability analysis of all the algorithms has been studied and it is found that all ensemble techniques have higher stability than the traditional feature selection methods. Thus, ensemble learning proves to be a superior technique for feature selection. Our results demonstrate that the proposed method is effective in identifying relevant features and stable features and can improve the performance of machine learning models.

Received: October 24, 2024 Revised: January 02, 2025 Accepted: January 31, 2025

Keywords: Feature selection; Ensemble technique; Stability; Microarray dataset; Biomarker selection

1. Introduction

Feature selection stands as a pivotal data preprocessing step in machine learning and data science, aimed at identifying a compact and highly discriminative subset from the original feature set. The problem being addressed is the high dimensionality of modern datasets, which can lead to overfitting and poor generalization of models. Reduced dimensionality using feature selection improves the model accuracy and reduces training time compared to the complete feature set. Ensemble techniques can be used for feature selection because they provide feature importance, which can be used to rank and select features. Ensemble learning is implemented in two ways as function perturbation and data perturbation. In function perturbation, multiple functions (algorithms) are implemented, and each algorithm's advantage is taken into consideration by aggregating (ensemble) the results of these algorithms [1]. However, in data perturbation, one algorithm is implemented on different samples of the same dataset. Here multiple sets of samples (by cross-validation or random spacing) yield diverse feature set selection, which is then aggregated to generate a single feature set [2-3]. Ensemble learning approach for feature selection improves classification accuracy compared to the traditional feature selection techniques [4]. Other than classification accuracy, the stability of the selected feature set is an important parameter to analyze the feature set's robustness [5-6]. Stability is the sensitivity of the selected features to the data variation. The higher the stability index more stable the feature set. Various stability indices are defined according to the feature category divided into similarity-based and frequency-based approaches [7].

The research question being investigated is whether ensemble methods are effective at selecting informative features in a dataset and if the selected features are stable compared to other feature selection methods. It will also investigate how the performance of models trained using the selected features compares to those trained using other feature selection methods or all the features. This paper implemented a tree-based ensemble learning approach for feature selection. Tree-based algorithms are highly effective for classification in machine learning. Therefore, they may also be used to identify the most discriminating feature responsible

for most accurate categorization. Our approach for ensemble feature selection includes function perturbation with the voting ensemble, an ensemble with a fixed number of features, and an ensemble with a contiguous number of features. Ensemble learning is found to be superior to other traditional feature selection algorithms. Ensemble learning algorithms are implemented on two high-dimensional microarray bioinformatics datasets. From our experimental study, it is observed that the voting ensemble outperforms other ensemble techniques, thereby reducing feature subset size and achieving higher accuracy. Stability analysis of all the algorithms has been studied and it is found that all ensemble techniques have higher stability than the traditional feature selection techniques. Thus, ensemble learning proves to be a prominent technique for feature selection.

2. Literature Review

Review on the recent development of ensemble learning for feature selection is studied and presented in this section. Review has mainly focused on studying ensemble approach, which has claimed good accuracy using ensemble feature selection. Target was to understand the way ensemble is applied and algorithms contributed in the ensemble. The literature review discusses the challenges and limitations of using ensemble methods for feature selection and stability analysis, and suggest future directions for research in this area.

Table 1: Review Summary

S.N.	Major Area/ Keywords	Important findings	Future scope identified
1	Classification, Ensemble methods, Feature selection, High-dimensionality	An ensemble feature selection method has been proposed, which is grounded in the assessment of the reliability of feature selectors.	In future research, it would be intriguing to fully investigate the potential of hybrid ensemble methods [8].
2	Breast Cancer; Feature Selection; Ensemble Classifier; Gene Expression Data	A proposed ensemble classifier employs Correlation-Based Feature Selection combined with forward search to analyze microarray breast cancer gene expression datasets.	Future research could explore using various feature selection techniques alongside the ensemble classifier to identify significant gene subsets in other cancer datasets [9].
3	ensemble feature selection, gene expression, microarray	Differential gene expressions derived from microarray technology were analyzed for patients with various diseases. The models demonstrated a high accuracy rate in predicting diagnoses for most patients.	Future efforts could focus on leveraging the advantages of integrating hierarchical functional organization [10].
4	machine learning, classification, random forest	The construction process of Random Forests is outlined, along with an examination of their capacity improvements and performance metrics. The application of Random Forests across various domains, including medicine, agriculture, and astronomy, is also discussed.	Random Forest is highly effective for feature selection due to its advantageous characteristics, such as Variable Importance Measure, Out-of-Bag error, and Proximities [11].
5	Extra-trees classifier, deep-learning, multi-class SVM	The Extra Trees classifier plays a crucial role in selecting the most significant features, which helps in cutting down the execution time. This approach not only speeds up computations but also lowers the classification error rate.	Extra tree classifier can be used independently for feature selection application [12].

From review presented in table I, it is observed that tree-based learning algorithms have good capability to identify most discriminating features. These algorithms rather than for classification can also be used effectively for feature selection tasks. Thus, our proposed ensemble technique considers implementing tree-based algorithms as one of the functions in function perturbation ensemble.

3. Methodology and Results

This section briefs the dataset used for experimentation and ensemble techniques applied for the feature selection. The dataset description section provided a detailed description of the dataset used in the experimentation. Ensemble technique section describes the ensemble function perturbation used to conduct the research or experiment.

A. Dataset Description

Dataset used for research experimentation are as detailed in table-II. These are biomedical high dimensional datasets. Datasets are of binary classification tasks and class represents a type of cancer. Low instances to features ration in table shows the high dimensionality and low sample size characteristic of the dataset. Dataset imbalance may exist due to low sample size and it is handled using SOMTE (Synthetic Minority Oversampling Technique) algorithm. Instead of oversampling with replacement, SOMTE creates "synthetic" representatives of the minority class. Data preprocessing and imbalance handling is the first step of proposed approach as indicated in figure 1. Preprocessed data is then given to feature selection algorithm in stage-1.

Table 2: Dataset Description

Sr. No.	Source	Dataset	#Features	#Samples	#Classes	Ratio (instances-to-features)	Domain
1	UCI ML repository	Colon	2000	63	2	0.032	Biomedical Dataset
2	UCI ML repository	Leukemia	7129	72	2	0.010	Biomedical Dataset

Such high dimensional datasets lead to poor classification accuracy when worked with full feature set [13]. This demands feature set reduction by selecting most relevant features to the task/class.

B. Ensemble Technique

The methodology for a study on feature selection and stability analysis using ensemble techniques involve the following steps:

- Selection of a dataset: The study would start by selecting a dataset that is relevant to the research question and has a sufficient number of features and samples.
- Preprocessing of the dataset: This would involve cleaning and normalizing the data, as well as handling missing values or outliers.
- Ensemble feature selection: Using ensemble techniques like as bagging or random forests, ensemble feature selection identifies the most important characteristics in a dataset. Additionally, comparing and contrasting alternative ensemble methods or optimizing certain parameters within these approaches may be part of the research.
- Stability analysis: In this step, the stability of the selected features against alterations in the dataset or in the feature selection procedure is examined in order to evaluate how robust the features are. This can be accomplished by adding noise to the data or by repeatedly running the feature selection process using various data subsets.
- Performance evaluation: This entails creating a model, such as a classifier or regression model, using the chosen features, and then assessing the model's performance using metrics like accuracy, precision, or recall.

For ensemble learning, two tree-based algorithms are used in addition to more conventional algorithms like sequential forward selection (SFS) and exhaustive feature selection (EFS): recursive feature elimination with a random forest estimator (RFE-RF) and recursive feature elimination with a gradient boost estimator (RFE-GB). Tree-based algorithms are highly effective for classification in machine learning [14-15], therefore they may also be used to identify the most discriminating feature that are responsible for most accurate categorization. Thus, including tree-based algorithms in function perturbation-based ensemble have increased the possibility of gaining higher accuracy and more stable feature selection.

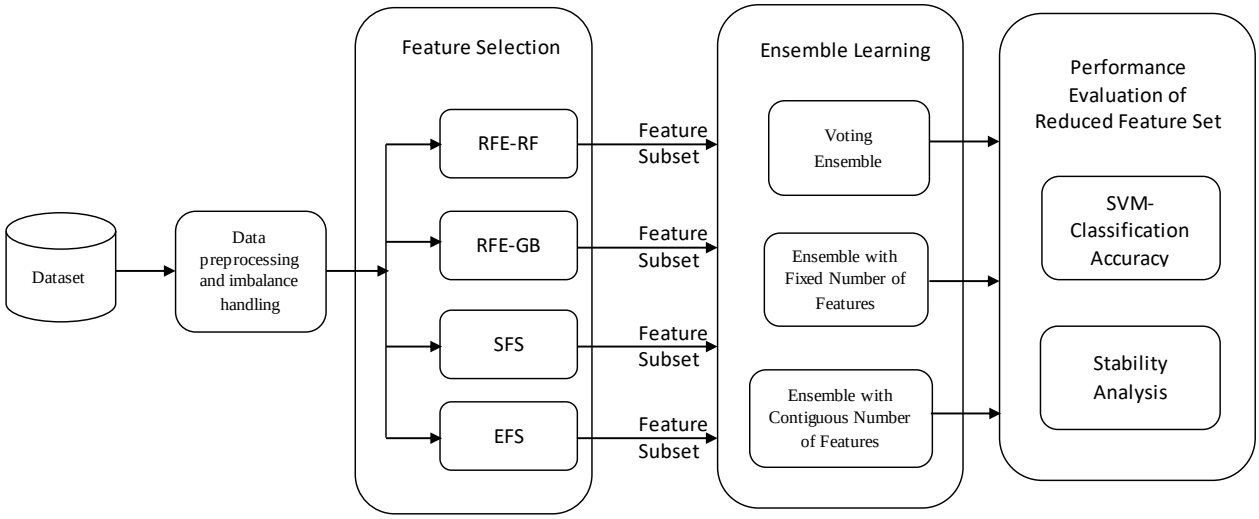

Figure 1. Proposed ensemble learning approach for feature selection

Figure 1 shows the proposed ensemble learning approach for feature selection. Each of the specified algorithm in feature selection stage is executed independently on both the datasets to check individual algorithm performances. Feature selection stage generates a feature subset selected by each algorithm independently. All these feature sets are given as input to ensemble learning making function perturbation where different algorithm's varying feature set is combined by various ensemble-learning approaches. In ensemble, learning stage three strategies are implemented namely, voting ensemble, ensemble with fixed number of features, and ensemble with contiguous number of features. Last stage evaluates the performance of reduced feature set generated by three ensemble-learning approaches by accuracy and stability analysis. Here two novelty in proposed approach we claimed as, i) Included and implemented two tree-based recursive feature selection algorithms for generating intermediate feature subset, and ii) Implemented and compared three promising ensemble techniques as voting, considering top features for each algorithm (fixed number of features), and considering all features selected by each algorithm (varying number of features-continuous features). This gives a clear vision to researchers for ensemble learning analysis based on accuracy and stability of feature set generated by each. Number of features selected by each individual algorithm and an ensemble learning with their accuracy is given in table III and table IV for Colon and Leukemia dataset respectively.

Table 3: Results of Implementation on Colon Dataset

Sr. No.	Algorithms Implemented	Total Dataset Feature	#Features Selected by the Algorithm	Accuracy
1.	RFE-RF	2000	361	0.8462
2.	RFE-GB	2000	250	0.8462
3.	SFS	2000	300	0.8381
4.	EFS	2000	286	0.8541
5.	Voting Ensemble	2000	356	0.8536
6.	Ensemble With Fixed Number of Features	2000	350	0.8479
7.	Ensemble With Contiguous Number of Features	2000	293	0.8520

Voting ensemble outperforms with highest accuracy of 0.8536 and 0.8671 for colon and leukemia dataset respectively over the other algorithms. It is observed from table II and III that all ensemble techniques have higher accuracy compared to individual algorithm outcome.

Table 4: Results of Implementation on Leukemia Dataset

Sr. No.	Algorithms Implemented	Total Dataset Feature	#Features Selected by the Algorithm	Accuracy
1.	RFE-RF	7129	465	0.8324
2.	RFE-GB	7129	320	0.8351
3.	SFS	7129	400	0.8298

4.	EFS	7129	356	0.8647
5.	Voting Ensemble	7129	420	0.8671
6.	Ensemble With Fixed Number of Features	7129	451	0.8465
7.	Ensemble With Contiguous Number of Features	7129	391	0.8562

Figure 2 and figure 4 shows the noticeable feature set reduction by feature selection algorithm still achieving good accuracy. This indicates the presence of irrelevant features in the datasets, which degrades the accuracy and increase the training time. Thus, feature selection is an important and must needed machine learning and data science tasks to improve the model performance.

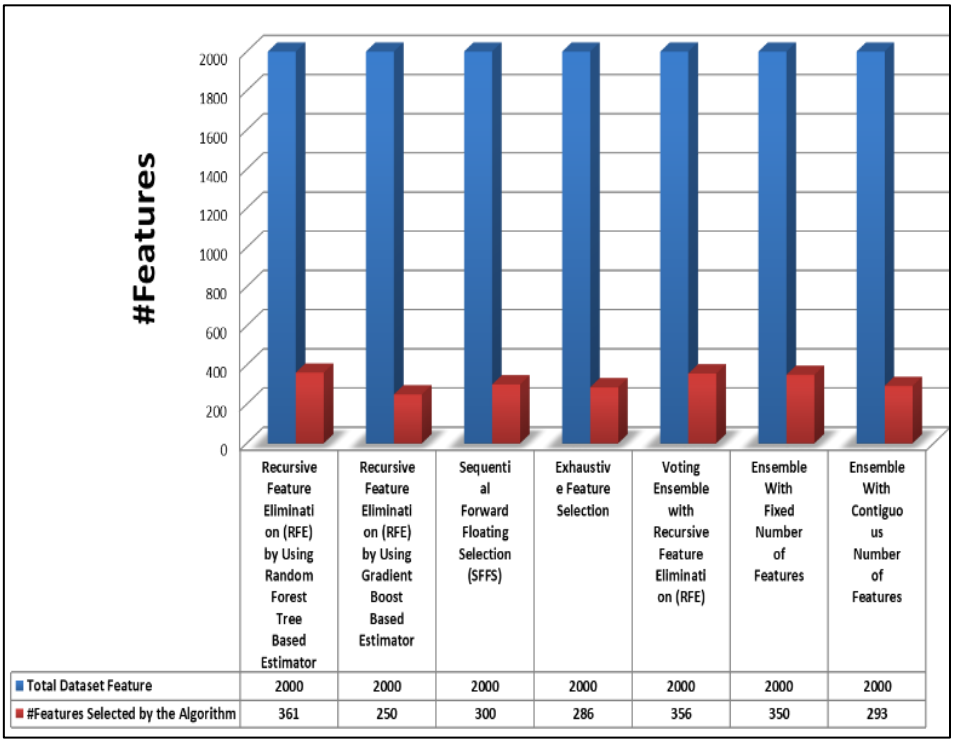

Figure 2. Total features selected over the original feauture set: Colon dataset

Figure 2 show the number of features selected over the original feature set for the colon dataset. Having a smaller number of relevant and informative features can lead to a more accurate model. This is because having too many features can add noise to the model and make it more difficult to extract the important information from the data, which can lead to overfitting and lower accuracy. Smaller feature set can help to reduce the dimensionality of the data and improve the performance of the model by reducing the impact of irrelevant or redundant features.

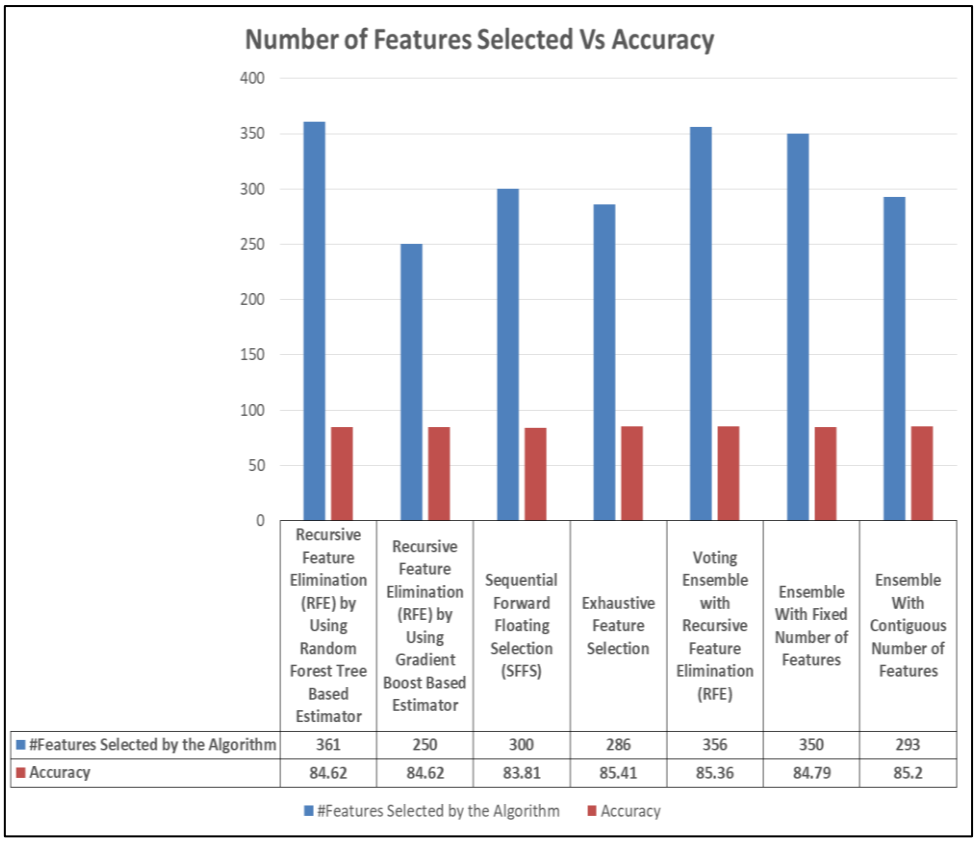

Figure 3. Prediction accuracy for the selected number of features: Colon dataset

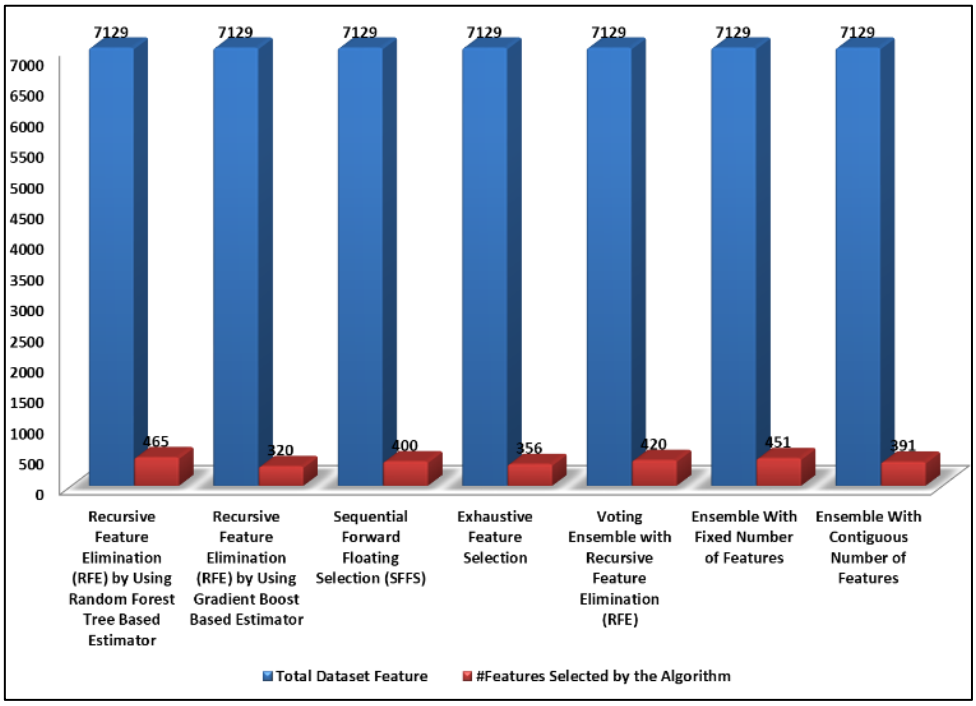

Figure 4. Total features selected over the original feature set: Leukemia dataset

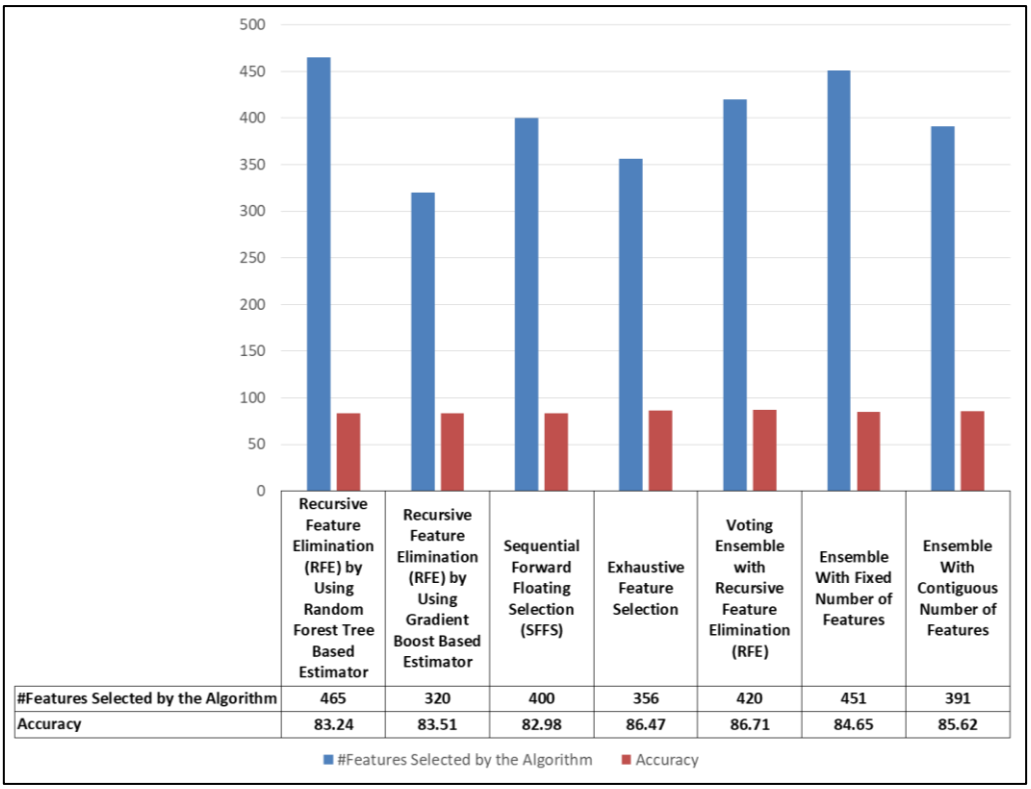

Figure 5. Prediction accuracy for the selcted number of features: Leukemia dataset

Figure 3 and figure 5 compares classification accuracy of each algorithm and number of features selected by each one. This gives very important information that reduced feature set size can include most relevant features responsible for class objective study or classification. The results of the feature selection and stability analysis using ensemble techniques indicate that the selected features have a significant impact on the performance of the machine learning models.

Table 5: Stability Analysis for Colon Dataset

SN.	Algorithms Implemented	#Total Selected Features by the Algorithm	Stability
1.	Recursive-Feature-Elimination (RFE) by Using Random Forest Tree-Based Estimator	361	0.71
2.	Recursive Feature Elimination (RFE) by Using Gradient Boost Based Estimator	250	0.79
3.	Sequential Forward Floating Selection (SFFS)	300	0.63
4.	Exhaustive Feature Selection	286	0.81
5.	Voting Ensemble with Recursive Feature Elimination (RFE)	356	0.86
6.	Ensemble With Fixed Number of Features	350	0.88
7.	Ensemble With Contiguous Number of Features	293	0.83

From table V stability index for all ensemble techniques is observed high for Colon dataset. Ensemble technique have been shown to be effective at handling high dimensionality and complexity of data, and thus have the potential to be stable in feature selection. The main reason for the stability of ensemble techniques is the averaging effect of multiple models. By combining the

predictions of multiple models, ensemble techniques can reduce the variance and increase the robustness of the predictions. This makes ensemble learning a robust technique for feature selection gaining high confidence of more stable feature selection.

Table 6: Stability Analysis for Leukemia Dataset

SN.	Algorithms Implemented	#Total Selected Features by the Algorithm	Stability
1.	Recursive-Feature-Elimination (RFE) by Using Random Forest Tree-Based Estimator	465	0.72
2.	Recursive Feature Elimination (RFE) by Using Gradient Boost Based Estimator	320	0.68
3.	Sequential Forward Floating Selection (SFFS)	400	0.53
4.	Exhaustive Feature Selection	356	0.82
5.	Voting Ensemble with Recursive Feature Elimination (RFE)	420	0.88
6.	Ensemble With Fixed Number of Features	451	0.89
7.	Ensemble With Contiguous Number of Features	391	0.83

Stability analysis for leukemia dataset presented in table IV. Here also ensemble techniques have shown high stability for the selected features compared to the rest of the algorithms. Highest stability of 0.89 is achieved by ensemble with fixed number of features for leukemia dataset. The results of stability analysis would show that the selected genes are relatively stable when compared to other feature selection methods and the performance of the classifier trained using the selected genes is better than those trained using other feature selection methods or using all the genes.

4. Conclusion

The ensemble techniques used in this analysis effectively identified the essential features in the dataset, and the stability analysis revealed that these features were consistently important across different model configurations. These findings demonstrate the utility of ensemble techniques for feature selection and stability analysis and suggest that incorporating these techniques into the model-building process can lead to more robust and accurate models. Our manuscript has presented the use of the tree-based algorithm for ensemble feature selection. The implementation of two high-dimensional datasets from the biomedical area is presented. Ensemble techniques have shown good accuracy compared to other algorithms studied.

Feature selection is a highly effective technique for ensemble learning. Specifically in the research presented, two tree-based algorithms were ensemble into an ensemble feature selection methodology. The results showed that this approach was extremely stable and robust in both of the data sets tested. It was shown from the research into feature selection and stability analysis with these ensemble methods that they are superior to their ability to detect important features within a dataset. Additionally, the features selected by ensemble methods had higher stability levels compared to those selected by other techniques of feature selection.

The study also demonstrates that the performance of the models trained using the selected features is better than those taught using other feature selection methods or using all the features. Our observation from the results is that including a tree-based algorithm has built a more robust feature selection, which can be noted from the high stability index of the selected features. Tree-based ensemble learning can be further studied for other tree-based algorithms. Future work can be done for specific tree-based algorithm ensembles for more robust feature selection.

References

[1] A. Wang, H. Liu, J. Yang, and G. Chen, "Ensemble feature selection for stable biomarker identification and cancer classification from microarray expression data," *Computers in Biology and Medicine*, vol. 142, p. 105208, 2022.

[2] R. Salman, A. Alzaatreh, and H. Sulieman, "The stability of different aggregation techniques in ensemble feature selection," *Journal of Big Data*, vol. 9, no. 1, pp. 1–23, 2022.

[3] B. Pes, N. Dessì, and M. Angioni, "Exploiting the ensemble paradigm for stable feature selection: a case study on high-dimensional genomic data," *Information Fusion*, vol. 35, pp. 132–147, 2017.

[4] D. Guan, W. Yuan, Y. K. Lee, K. Najeebullah, and M. K. Rasel, "A review of ensemble learning-based feature selection," *IETE Technical Review*, vol. 31, no. 3, pp. 190–198, 2014.

[5] A. Kalousis, J. Prados, and M. Hilario, "Stability of feature selection algorithms," in *Proc. 5th IEEE Int. Conf. Data Mining (ICDM'05)*, Nov. 2005, pp. 8–pp.

[6] S. Nogueira, K. Sechidis, and G. Brown, "On the stability of feature selection algorithms," *J. Mach. Learn. Res.*, vol. 18, no. 1, pp. 6345–6398, 2017.

[7] J. Racicot, *Dynamiques de connectivité cérébrale fonctionnelle associées aux fluctuations journalières des états affectifs*, 2024.

[8] B. Pes, "Ensemble feature selection for high-dimensional data: a stability analysis across multiple domains," *Neural Computing and Applications*, pp. 1–23, 2019.

[9] T. Al-Quraishi et al., "Breast cancer risk assessment prediction using an ensemble classifier," in *Proc. CAINE2017*, 2017.

[10] T. Gaudelet et al., "Unveiling new disease, pathway, and gene associations via multi-scale neural networks," *arXiv preprint arXiv: 1901.10005*, 2019.

[11] N. M. Abdulkareem and A. M. Abdulazeez, "Machine learning classification based on random forest algorithm: A review," *Int. J. Sci. Bus.*, vol. 5, no. 2, pp. 128–142, 2021.

[12] D. Baby, S. J. Devaraj, and J. Hemanth, "Leukocyte classification based on feature selection using extra trees classifier: A transfer learning approach," *Turkish J. Electr. Eng. Comput. Sci.*, vol. 29, no. 8, pp. 2742–2757, 2021.

[13] L. Azkue, J. Kerexeta, J. Sampedro, M. Espejo, and N. Larburu, "Predictive models of ward admissions from the emergency," *Age*, vol. 50, pp. 23–77.

[14] P. G. Asteris et al., "Slope stability classification under seismic conditions using several tree-based intelligent techniques," *Appl. Sci.*, vol. 12, no. 3, p. 1753, 2022.

[15] R. M. Mohana, C. K. K. Reddy, P. R. Anisha, and B. R. Murthy, "Random forest algorithms for the classification of tree-based ensemble," *Mater. Today: Proc.*, 2021.