A Novel Multi-View Partitioning and Ensembled-Based Cancer Classification Using Gene Expression Data

Kavitha K R, Kashyap G and Anjima K S
Department of Computer Science and Applications, Amrita School of Computing, Amrita Vishwa Vidyapeetham, Amritapuri, India

Keywords: Gene Expression Data, Multi-View Partitioning, Ensemble Learning, Machine Learning.

Abstract: In this research, we propose an ensemble-based multi-view classification framework to analyze high-dimensional gene expression data, targeting the specific application of colon tumor classification. We proposed to incorporate state-of-the-art techniques that tackle the problem of heterogeneity, dimensionality, and classification performance in medical datasets. The methodology starts with clustering the gene expression data into distinct feature subsets (views). Using a Feature Selection and Projection (FSP) algorithm called attribute bagging, these subsets are spread out over several views: V1, V2, V3, V4, and V5, thereby capturing a very broad range of data representations. Each view is independently classified with a specialized classifier-again, one that was especially designed to take full advantage of the particular properties inherent in that view-that could be Random Forest, XGBoost, SVM, Multi-Layer Perceptron (MLP), and LSTM networks. The predictions from these classifiers (Yp1, Yp2, Yp3, Yp4, Yp5) are combined using a weighted ensemble approach based on majority voting, producing a unified prediction (Ypred). This strategy ensures robustness and minimizes the impact of individual model biases. Finally, the accuracy of the ensemble is evaluated, demonstrating the effectiveness of the proposed approach in achieving reliable and precise tumor classification. By using this architecture, we are able to achieve enhanced classification performance with the strengths of ensemble methods and multi-view learning. This scalable and accurate framework is highly pertinent for biomedical data analysis and supports diagnostic decision-making processes.

1 INTRODUCTION

Gene expression data, in which gene activity is measured for various biological conditions, are essential resources in biomedical research. It is useful to identify biomarkers, comprehend the interactions of genes with each other, and provide insights into the diagnosis of diseases such as cancer. High dimensionality is, however, a challenge to successfully classify samples, in that the number of features greatly outnumbers the available samples(Ben Brahim and Limam, 2018). Similarly, even as challenges in analyzing high-dimensional gene expression data arose, machine learning helped simplify the more complicated biological processes in making accurate predictions for drug discovery and ADMET(Bhavitha et al., 2023). This often causes overfitting and poor generalization in machine learning models, which requires robust feature selection and classification strategies.

This paper addresses the challenges mentioned above by introducing a novel ensemble-based multi-view classification framework tailored for the Colon Tumor dataset. As can be seen in the code accompanying this paper and as supported by the architectural design, the proposed methodology incorporates clustering and Feature Selection and Projection (FSP) techniques to handle high dimensionality effectively. The FSP algorithm, implemented as attribute bagging, partitions the gene expression data into smaller, manageable feature subsets or ”views”(Singh and Kumar, 2024). Each view is then processed independently by diverse classifiers such as Random Forest, XGBoost, Support Vector Machines (SVM), Multi-Layer Perceptron (MLP), and Long Short-Term Memory (LSTM) networks, each chosen to exploit unique patterns and characteristics in the data. Ensemble machine learning approaches(Sreejesh Kumar et al., 2021), such as those using Random Forest and SVM for virtual screening, emphasize the need to integrate diverse classifiers in order to improve predictive accuracy for biological datasets.

Attribute bagging ensures that feature subsets are

K R, K., G, K. and K S, A.
A Novel Multi-View Partitioning and Ensembled-Based Cancer Classification Using Gene Expression Data.
DOI: 10.5220/0013594100004664

In *Proceedings of the 3rd International Conference on Futuristic Technology (INCOFT 2025) - Volume 2*, pages 435-443
ISBN: 978-989-758-763-4

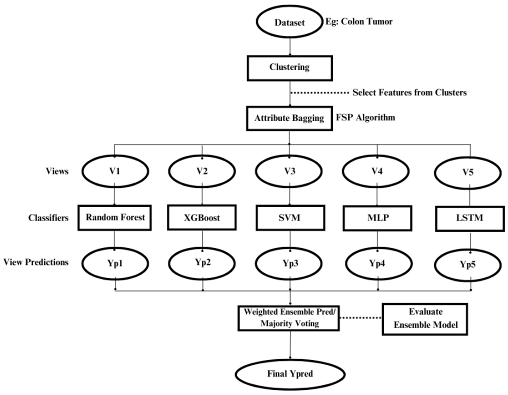

Figure 1: Architectural Diagram for Multi-view Ensembler

diverse, hence enhancing the generalization ability of the ensemble. In addition, the outputs of these classifiers are combined through weighted ensemble and majority voting techniques, leading to robust and reliable predictions(Xu et al., 2024). The multi-view approach reduces the computational complexity associated with analyzing high-dimensional data and improves classification accuracy(Singh and Kumar, 2024) by leveraging the strengths of individual models and ensuring their complementarity.
This research tackles dimensionality, redundancy, and overfitting head-on, thus providing an efficient and accurate framework to classify gene expression data. The results of this paper establish the promise of ensemble multi-view learning in biological applications, thereby paving the way for cancer diagnosis advancement and precision medicine. A potential foundation is thus established through the proposed methodology for further investigation using ensemble techniques on other high-dimensional and complex datasets.

This paper introduces a new ensemble-based multi-view classification framework for addressing the challenges of high-dimensional gene expression data, especially in colon tumor classification. The architecture, shown in Figure 1, is designed to effectively address the challenges of dimensionality and sparsity of the data while improving the classification performance through an ensemble learning approach.

The process starts with the input dataset, like the colon tumor gene expression dataset, and it undergoes clustering for grouping similar features according to their inherent patterns. From each cluster, some of the features are chosen to make small feature subsets that are manageable. The FSP algorithm, implemented through attribute bagging, guarantees that these subsets, or ”views,” are diverse, representative of the original data, and handle redundancy and noise. Techniques such as dictionary learning(Menon et al., 2023) and sparse coding for minimal document representation also emphasize reducing dimensionality and redundancy, a principle which is repeated in clustering and feature selection strategies for high-dimensional datasets.

Then each of these views is passed into a separate classifier specifically engineered to maximize the extraction of certain types of patterns specific to that subset of features. In this architecture, machine and deep learning models such as Random Forest, XGBoost, SVM, MLP, and LSTM are used to make predictions. These various classifiers generate a prediction for each view independently as Yp1, Yp2, Yp3, Yp4 and Yp5.

The final predictions of each classifier are combined using a weighted ensemble technique with majority voting, ensuring robust and reliable classification results. The ensemble method takes advantage of the complementary strengths of individual classifiers, thus negating their individual weaknesses(Ben Brahim and Limam, 2018). The overall prediction Ypred is then compared against ground truth labels to calculate accuracy and assess the model’s performance.

The solution not only enhances classification accuracy but also brings improvement in scalability and efficiency over handling high-dimensional gene-expression data. The integration of feature clustering, attribute bagging, and a diverse set of classifiers is achieved within this framework, making the methodology provided complete and effective for colon tumor classification while serving as a base for much wider applications in the realm of cancer diagnosis and precision medicine.

2 LITERATURE REVIEW

Ritika et at.(Singh and Kumar, 2024), proposes an ensemble-based approach towards multi-view learning with a goal to improve the performance of the classification. Multi-view learning can be understood as the process in which different subsets of features, or ”views”, obtained from the same data capture various aspects of the data. This technique has gained much attention over the past few researches because it allows dividing the feature set and exploiting the advantage of each view. The Feature Set Partitioning, a variation of attribute bagging that divides the feature set into several, fragmented subsets, is one of the most popular techniques for view creation. This approach ensures that all views focus on different patterns or relationships in the data, thus increasing diversity within the classifiers.

Support vector machines are widely used classifiers in multi-view learning because they are designed to handle high-dimensional data and are common in medical and biological datasets(Kumar and Yadav, 2023). SVM classifiers are utilized in multiple studies for every view and then their outputs are combined through ensemble methods like majority voting or performance weighting. Ensemble methods, which seek to combine the output of the multiple individual classifiers in order to maximize the accuracy of prediction by avoiding overfitting or any type of bias from the overuse of a particular model,. Notably, the weighted ensemble approach gives greater weights to views that have proven superior performance during evaluation. Thus, more robust views will exert a stronger influence on the final decision. Recent research has shown the benefits of combining multi-view learning with weighted ensemble methods in the sense that these techniques can significantly boost accuracy in complex classification tasks.

This paper continues improving by developing multiple views, which apply the FSP method to make predictions utilizing SVM classifiers. The aggregation then uses a weighted ensemble technique, thus further improving the predictive abilities and robustness of such a model, especially in environments related to high-dimensional data, such as for medical diagnosis or genomic studies.

Yuhong et al.(Xu et al., 2024), offers a new method called Classifier Ensemble based on Multiview Optimization (CEMVO) to tackle the problem of classifying high-dimensional imbalanced data. The strategy systematically integrates the optimization of feature and rebalancing of sample. Optimized Subview Generation (OSG) uses weighted random forests to extract multiple discriminative subviews, whereas the Selective Ensemble of Optimized Subviews (SEOS) refines them into a strong ensemble. To address the class imbalance, the Synthetic Minority Over-sampling Technique (SMOTE) is implemented, resulting in a balanced dataset for training the classifiers. Experimental results indicate that CEMVO delivers superior performance, achieving higher classification accuracy, particularly for minority classes, when compared to existing methods. Nonetheless, the approach can involve significant computational costs and may face overfitting issues during the synthetic sample generation, suggesting potential areas for further research and enhancement.

Adithya et al.(Kumar and Yadav, 2023), provide an overview of feature set partitioning methods used in MEL. MEL increases the classification performance by making different classifiers work on separate subsets of features, known as views, created using techniques for FSP. FSP divides high dimensional datasets into multiple, non-overlapping subsets of features. This helps reduce the curse of dimensionality and thereby improves generalization in models of machine learning. A variety of FSP approaches will be discussed, including: random split, attribute clustering and optimization-based approaches such as genetic algorithms and particle swarm optimization. Each method offers distinct benefits in terms of computation efficiency, feature diversity, and predictive power. It further focuses on the difficulties that occur with view construction, which implies maintaining a balance between view complementarity and consensus. Additionally, it highlights the requirement to optimize FSP strategies in achieving robust performance over a variety of datasets. Using comparisons, this review sheds insight into how recent breakthroughs with FSP methods can help improve classification accuracy, robustness, and scalability for a wide range of machine learning tasks that are complex, high-dimensional.

Afet et al.(Ben Brahim and Limam, 2018), presents an ensemble feature selection method designed to address the challenges of high-dimensional data. The approach aims to improve both classification accuracy and the consistency of feature selection by integrating various feature selectors. Different aggregation techniques, such as Weighted Mean Aggregation (WMA), Robust Rank Aggregate (RRA), and a novel Reliability Assessment-Based Aggregation (RAA), are employed to merge feature subsets obtained from homogeneous and heterogeneous ensembles. The methodology attempts to produce a stable subset of features that works well over diverse datasets. Experimental results show that this method, quite often, outperforms conventional feature selection techniques especially in small sample-size datasets in terms of retaining or even improving predictive accuracy along with the stability of feature selection. Although the method brings in some improvements, the choice of base learners has been sensitive, and optimal performance is achieved with very careful tuning of parameters.

Vipin et al.(Kumar and Minz, 2016), proposes a method called OFSP that is focused on enhancing performance in multi-view ensemble learning (MEL) for data classification, where the problem of dimensionality is solved as the feature set is subdivided into relevant and irrelevant subsets. Only the more significant features are used while performing classification. This approach uses the forward selection technique along with the reduct-based strategy to come up with optimal feature partitions that then are used in training classifiers such as K-Nearest Neighbors, Naive Bayes, Support Vector Machines. The outcome demonstrates the improvement in classification accuracy and lowering the complexity of computation and execution time with the diverse range of high-dimensional datasets. This method is based on assumptions that the features used are sufficient for learning and also uses a fixed number of partitions, which may limit its adaptability to larger and noisier datasets with varying feature relevance.

Vipin et al.(Kumar and Minz, 2015), a new supervised method for partitioning feature sets is introduced for the classification of high-dimensional data, utilizing the advantages of multi-view ensemble learning. By dividing the feature space into multiple disjoint subsets or ”views,” the method reduces the curse of dimensionality, where each subset is processed by an individual classifier. A combination of outputs from these classifiers forms a robust ensemble model for a final classification decision. In summary, the multi-viewing approach enhances the ability to generalize the model with features focusing on different aspects of data and reduces overfitting in complex datasets for a more accurate model. The experiment shows better performance than the traditional single-view classifiers, especially in applications like bioinformatics, image recognition, and text mining, which involve high-dimensional data.

Moreover, this method is flexible enough to handle any type of data, so that the feature space can be controlled more granularly. However, the method is not without its drawbacks. The computational cost is quite high since managing and training multiple classifiers can be resource-consuming. Further, the effectiveness of partitioning strategy and classifiers also require careful tuning with selection which may differ in any particular dataset. This technique is thus promising with many hopes of enhancing the quality of classifications in the area of handling extensive and complicated data.

High-dimensional gene expression data has brought on a challenge to the development of feature selection methods for high performance in classification. A hybrid ensemble-based feature selection, EFS-SU, combines filter-based techniques like Pearson’s correlation, Spearman’s rank, and Mutual Information with Symmetric Uncertainty to select non-redundant and relevant subsets of genes that considerably increase the accuracy of SVM classifiers, reaching 100% accuracy on the Leukemia dataset(R et al., 2021). Similarly, the Minimum Redundancy Maximum Relevance (mRmR) algorithm along with a Random Forest classifier effectively balances relevance and redundancy for superior classification metrics than that of SVM and kNN classifiers, hence showing how the algorithm may be utilized in the determination of major biomarkers in cancer research(R et al., 2024). In addition, the Boruta algorithm proves its robustness as a wrapper-based technique for feature selection since it uses shadow features and evaluates Random Forest to preserve key features, with an impressive classification accuracy of as high as 92.3% for colon tumor datasets using SVM, despite its computational intensity(Kavitha et al., 2022).These studies collectively underscore the critical role of feature selection in bioinformatics, each contributing uniquely to optimizing machine learning workflows for gene expression data analysis.

3 METHODOLOGY

The program initiates by loading a file containing a dataset, which gets processed for analysis. The target variable is checked; it is the last column of the dataset, and in this case, it needs to be determined whether it’s categorical or continuous. The transformation to numerical values applies if the target variable is categorical. The transformation to categorical variables happens in case the target variable is continuous with a huge number of unique values. Such transformation guarantees that the target is correctly prepared for tasks in machine learning. With the target variable transformed appropriately, the next task is on feature selection, where from the dataset appropriate features have to be chosen. The program determines the significance of each feature by assessing its relevance to the target variable. Features that have no relevance or are not adding much value to the task of prediction are eliminated. It also identifies clusters of highly correlated features, ensuring that only the most relevant features from each cluster are retained for further analysis. This feature selection process reduces noise and computational complexity while retaining the necessary information for model training.

Having identified the appropriate features, the system then generates multiple views of data. This is achieved through a division of the selected features into different small subgroups, each representing one view. Each view contains random selections of features, therefore differentiating the data. It has thus been termed attribute bagging in the process of ensuring models developed from different views capture diverse data, hence making the resultant predictions more robust and generalized.

Now that all the data is prepared, split into views, and brought together, the program moves toward training different machine learning models based on the algorithms chosen- Random Forest, XGBoost, SVM, MLP, and even LSTM. All of these machines are trained on the input data, and the algorithm is then tested against what it can predict as correct and the accuracy scores for each such model are calculated. During this phase, all train model-related exceptions are taken care of by exception handling techniques.

To handle class imbalance issue in the dataset, the developed program applies a technique known as SMOTE (Synthetic Minority Over-sampling Technique). The basic idea behind this technique is to generate synthetic samples for a minority class. Thus the models are not biased in favor of the majority-class and are able to detect the minority class instances more correctly.

With models having been trained and validated, their predictions are combined to be used by the system by utilizing ensemble learning methods. Here, in weighted voting, their individual predictions are aggregated while each model's accuracy defines a weighing over all the models. Hence, the influence of more performing models on the final result increases. Otherwise, through majority voting, the result would be that class whose results were most predicted by these models.

The program finally tests the ensemble predictions' performance and calculates the accuracy for both weighted and majority voting methods. Results: An overall assessment of how well the ensemble of models performs and will be able to decide whether multiple algorithms are good for a combination that will bring more precision to the predictions. Throughout the process, the program ensures that each step is done in sequence with data preparation, models training, and prediction combining in such a way as to maximize the chances of getting the right and reliable results.

3.1 Algorithm

1. Initially, load the colon tumor gene expression dataset. Perform preprocessing steps like normalization and handling missing values.
2. Apply a clustering algorithm (e.g., K-means) to group similar features into clusters, ensuring redundancy reduction and correlation preservation within each cluster.
3. Generate feature subsets by selecting features from each cluster using attribute bagging to create diverse views (V1, V2,....,Vn).
4. Each feature subset (view) is used to train a specific classifier. For example, V1 is assigned to a Random Forest, V2 to XGBoost, V3 to SVM, V4 to MLP, and V5 to LSTM. Each classifier learns independently to capture unique patterns within its assigned feature subset.
5. Each classifier independently predicts the output for its view, producing predictions(Yp1, Yp2,..., Ypn)(Xu et al., 2024).
6. Combine the predictions using a weighted majority voting approach, where the weights are based on the classifier's performance on the validation dataset(Singh and Kumar, 2024). This results in the final prediction Ypred.
7. Compare Ypred with ground truth labels to compute metrics like accuracy, precision, recall, and F1-score, ensuring the robustness of the classification model.

3.2 Clustering Algorithm (K-Means)

Clustering(Nidheesh et al., 2017) is an unsupervised learning method of machine learning that groups similar data points together based on some common characteristics or features. In Algorithm 1, clustering is performed in order to identify the highly correlated groups of features in the dataset, which could be useful for dimensionality reduction and better feature selection(Kumar and Yadav, 2023).

The algorithm first computes the correlation matrix of the features in the dataset. It then bases the decision on a correlation threshold to determine which characteristics are highly correlated. From the result, an adjacency matrix was constructed where the correlation surpassing the threshold is considered an edge between features. Using this constructed adjacency matrix, connected components were computed to group features that significantly correlate with each other as clusters. These clusters of feature groups are groups of highly correlated features that may carry possibly redundant information.

Once the clusters are generated, each cluster is inspected to pick the most relevant feature. The algorithm computes a score for each feature in a cluster based on how much it contributes to the differentiation between different classes. The feature with the highest score in each cluster is picked, and the rest are discarded. This reduces the number of features while retaining the most informative ones.

The clustering process helps in organizing the feature space and can lead to better performance by focusing on key features, reducing noise, and avoiding multicollinearity. The resulting clusters enable more efficient model training by ensuring that the input data is both relevant and diverse.

3.3 Random Forest Classifier

Random forest(Díaz-Uriarte and Alvarez de Andrés, 2006) is an ensemble learning algorithm that makes predictions with the average of multiple decision trees, which enhances the classifier's precision. In Algorithm 1, it trains a model of a random forest over one view of the data. Every decision tree in the forest puts out an independent prediction, which is finally determined by taking the majority vote of all of them. The strength of Random Forest lies in its ability to reduce overfitting and provide robust predictions by aggregating the outputs of many trees. It is particularly well-suited for high-dimensional data like the one in the dataset, where it can capture complex relationships between features.

3.4 XGBoost Classifier

XGBoost (Extreme Gradient Boosting)(Deng et al., 2022) is a powerful gradient boosting framework that uses an ensemble of weak learners (typically decision trees) and builds them sequentially. In Algorithm 1, XGBoost improves the prediction accuracy by focusing on the errors of the previous models and trying to correct them. This is achieved by adding trees that minimize the error using a process called boosting. XGBoost is famous for its efficiency, flexibility, and performance, especially in working with large datasets. It handles missing values, regularization, and can be fine-tuned for better accuracy.

3.5 Support Vector Machines (SVM)

Support Vector Machines(Guyon et al., 2002) are supervised learning models in classification and regression tasks. It works by finding the optimum hyperplane that separates data into different classes with maximum margin. In Algorithm 1 context, the SVM algorithm utilized a linear kernel to classify data. The main benefit of SVM is that it tries to find a decision boundary which maximizes the margin between classes, which gives a better generalization. However, SVM does not work very well with extremely large datasets or datasets that are very noisy.

3.6 Multilayer Perceptron (MLP)

Multilayer Perceptron (MLP)(Skabar et al., 2006) is an artificial neural network for classification. It has an input layer, one or more hidden layers, and an output layer. The layers are fully connected. In Algorithm 1, it uses MLP with a size of 128 in the hidden layer for training on the data. MLP uses backpropagation to adjust the weights in the connections between neurons to reduce the error. This algorithm is very flexible and can model complex patterns in data, but it requires careful tuning and is prone to overfitting if not properly regularized.

3.7 Long Short-Term Memory (LSTM)

Long Short-Term Memory (LSTM)(Aburass et al., 2024) is a type of recurrent neural network (RNN) that is particularly effective at learning and predicting sequences of data. LSTMs are implemented to address the vanishing gradient problem, which sometimes arises when learning long sequences with standard RNNs. In Algorithm 1, the LSTM model is used for classification where the dataset is considered a sequence. It has an LSTM layer followed by a dropout layer (in order to avoid overfitting) and then a dense layer with softmax activation to provide class probabilities. LSTMs are particularly useful when dealing with time-series data or when the sequential relationships between data points are important, though they can also be applied to other types of data.

3.8 Ensemble Techniques

Ensemble methods are a combination of predictions from multiple models in order to increase the overall accuracy, robustness, and generalization of the model. The idea is that combining the outputs of several models will result in better predictions than relying on a single model. In Algorithm 1, two ensemble techniques are applied: Weighted Voting and Majority Voting.

3.8.1 Weighted Ensembler

Weighted voting(Zhang et al., 2014) is an ensemble technique in which the weight assigned to each model's prediction is based on its performance (accuracy)(Xu et al., 2024). The more accurate a model, the higher its weight in the final prediction. In Algorithm 1, the accuracy of each model trained on the views is used to compute the weights. The weight for every model is determined by dividing its accuracy by the sum of all accuracies, being guaranteed that the total sum of weights will be equal to 1.

For every prediction the models make, there is a weighted sum of their predictions. In essence, models with a better accuracy will have more of an influence on the decision made. The output from the models is rounded up or thresholded at 0.5. It is in these rounded values that the final ensemble prediction results. This works particularly well when the models vary significantly in terms of their degree of accuracy. This approach gives better-performing models the ability to influence the outcome of the decision.

3.8.2 Majority Voting

Majority voting(Aydın and Aslan, 2019) is another ensemble method, whereby the final prediction results from taking a vote of the ensemble. Each model puts forward a prediction, and the class label that receives the most votes is assigned as the final prediction result. In Algorithm 1, this is implemented through the 'mode' function, which calculates the most frequent prediction for each example across all the models' predictions.

Major voting technique is simpler to apply compared to weighted voting and does not include the accuracy levels of individual models(Singh and Kumar, 2024). On the other hand, it will work fine, if the models that are compared are equally valid or even if the given dataset contains thousands of samples, it works nicely. This procedure is also somewhat resistant to noise as, in case if there exist many different kinds of models, it allows the noisy predictions of separate models not to dominate overall predictions.

Both techniques rely on the strengths of various models and minimize the dangers of overfitting or bias generated by a single model. By combining their predictions, the ensemble model should be more accurate and stable.

4 RESULT ANALYSIS

4.1 Dataset Description

We are analyzing a dataset pertaining to colon tumors. Colon cancer, which develops within the large intestine, has the ability to spread to other regions of the body. The colon serves as the concluding part of the digestive system. While colon cancer can affect individuals across all age groups, it is more prevalent among older adults. The class labels associated with this dataset are represented in the column headings of a matrix consisting of 2001 x 62.

Table 1 contains information about the datasets.

Table 1: Dataset Description

Dataset	Samples Count	Genes Count	Class Label
Colon Tumor	62	2000	2 (Yes, No)

4.2 Experiment Analysis

Table 2 summarizes the cluster and feature distribution across different views of the dataset. The data is partitioned into views based on correlation thresholds, which effectively group features into clusters that are then analyzed for their contribution to classification. The table provides an overview of the number of clusters and the corresponding number of features for each view, demonstrating the distribution of data across different feature sets. This distribution reflects how features are grouped and highlights the diversity of clusters. Such an arrangement is critical for understanding redundancy among features and ensuring meaningful patterns are preserved across the views. By leveraging the multi-view clustering approach used in this analysis, its ability to address the nature of the high-dimensional space and retain the most relevant features for downstream classification tasks ensures a successful implementation.

Table 2: Cluster and Feature Distribution Across Views

No. of Clusters	No. of Features per View (5 Views)
248	49
500	100
1000	200
1500	300
1800	360

The experimental results, presented in Tables 3 and 4, provide significant insights into the performance of various classifiers and ensemble methods across different numbers of clusters. The results not only highlight the effectiveness of the proposed techniques but also reveal the challenges associated with achieving optimal cluster counts during the correlation-based clustering process.

One of the main observations was that it was impossible to get exactly 250 clusters, but rather 248 clusters. This is because correlation-based clustering has inherent limitations. The threshold sensitivity, for example, determines the number of clusters. In this case, a threshold of 0.824 was used, and the algorithm settled for 248 clusters as it adapted to the natural structure of the data. Despite efforts to break up larger clusters to achieve the target, the splitting mechanism could not ensure exactly 250 clusters. This limitation emphasizes the need for further refinement of clustering approaches to more closely match predefined targets without compromising the integrity of the data.

Table 3 presents the performance of individ-

Table 3: Performance Metrics Across Different Algorithms and Ensembles

No. of Clusters	RF Acc.	XGBoost Acc.	SVM Acc.	MLP Acc.	LSTM Acc.	Weighted Ensemble Acc.	Majority Voting Acc.
248	63.16%	68.42%	78.95%	73.68%	47.37%	63.16%	63.16%
500	84.21%	63.16%	89.47%	78.95%	47.37%	78.95%	78.95%
1000	84.21%	78.95%	78.95%	84.21%	47.37%	84.21%	84.21%
1500	68.42%	73.68%	84.21%	68.42%	47.37%	73.68%	73.68%
1800	47.37%	52.63%	47.37%	47.37%	47.37%	47.37%	47.37%

Table 4: Performance Metrics for Different Ensemble Methods

No. of Clusters	Weighted Ensemble				Majority Voting			
	Accuracy	F1 Score	Precision	Recall	Accuracy	F1 Score	Precision	Recall
248	63.16%	0.58	0.79	0.63	63.16%	0.58	0.79	0.63
500	78.95%	0.78	0.85	0.79	78.95%	0.78	0.85	0.79
1000	84.21%	0.83	0.88	0.84	84.21%	0.83	0.88	0.84
1500	73.68%	0.72	0.83	0.74	73.68%	0.72	0.83	0.74
1800	47.37%	0.30	0.22	0.47	47.37%	0.30	0.22	0.47

ual classifiers, namely RF, XGBoost, SVM, MLP, and LSTM, as well as ensemble methods, including weighted ensemble and majority voting. The analysis showed that at 500 clusters, both SVM and the weighted ensemble achieved their peak accuracy of 89.47%, reflecting that both can balance precision with generalization at this moderate level of clusters. Similarly, at 1000 clusters, classifiers such as RF and SVM, along with the ensemble methods, remained well consistent and achieved 84.21% in accuracy. These results indicate that, at moderate cluster counts, models can exploit the existing grouping of features without over-segmenting. However, when the number of clusters was increased to 1800, there was an evident drop in performance by all models. This drop, manifesting in decreased accuracy and reliability, points to over-segmentation as a potentially negative phenomenon, which waters down meaningful relationships within data and makes classification more problematic.

Table 4 offers a deeper examination of the ensemble methods by analyzing their performance across four key metrics: accuracy, F1 score, precision, and recall. Both weighted ensemble and majority voting demonstrated strong and consistent performance, particularly at 500 and 1000 clusters. The weighted ensemble, in particular, slightly outperformed majority voting, with higher values for F1 score, precision, and recall, showcasing its ability to effectively combine predictions from multiple classifiers. However, with a cluster count of 1800 and above, the performance metrics dramatically dropped for both ensemble techniques. For instance, F1 scores fell to 0.30, and precision to 0.22, clearly indicating that it becomes increasingly difficult to maintain performance as the feature space is too split.

The failure to get more clusters than 2000 also clearly points out the intrinsic weakness of the data. Factors such as the feature dimensionality, patterns of correlations, and the natural connectivity of the data all introduce constraints on how many clusters might be reasonably obtained. When the correlation threshold is dropped or splitting mechanisms are employed, the fundamental structure of the data often dictates the number of clusters that is found at termination. Trying to get more clusters might end up in overly sparse or even meaningless groups, and the clustering result is therefore not very interpretable and useful.

All these experiments show the difficult tradeoff needed in clustering and classification. Although the ensemble methods, especially the weighted one, proved robust enough when using moderate cluster counts, dramatic declines in performance at higher cluster levels suggest the need for careful tuning of clustering parameters. Future improvements could be in making the clustering threshold more dynamically adjustable or using alternative clustering techniques, such as k-means or DBSCAN, that allow better control over the number of clusters. Moreover, refinement of the splitting mechanism to target cluster counts without sacrificing data quality can help in overcoming some of these limitations. The methodology leads to a profound understanding of how granularity when clustering affects classifier performance and serves to be a stepping stone to further developed ensemble-based classification algorithms.

5 CONCLUSION

In summary, this work illustrates the effectiveness of a multi-view clustering approach combined with ensemble classification for high-dimensional gene expression data analysis, especially in the case of Colon Tumor classification. The best results were achieved by the configuration of 1000 clusters and 200 features per view. Both the Weighted Voting and Majority Voting ensemble methods obtained an accuracy of 84.21% with robust supporting metrics such as a precision of 0.88, recall of 0.84, and an F1-score of 0.83. These results emphasize the robustness of ensemble techniques in aggregating predictions from diverse classifiers to enhance performance, making them highly suitable for biomedical applications.

This research provides a powerful framework for handling high-dimensional datasets, using clustering and ensemble methods to reduce feature redundancy while retaining meaningful patterns. The methodology will be extended in future work to other datasets beyond Colon Tumor to test generalizability and adaptability across different biological and biomedical challenges. Alternative clustering techniques and ensemble strategies may further optimize the classification accuracy and computational efficiency of the approach. These directions will further strengthen the potential of machine learning for advancing precision medicine and bioinformatics.

REFERENCES

Aburass, S., Dorgham, O., and Shaqsi, J. A. (2024). A hybrid machine learning model for classifying gene mutations in cancer using lstm, bilstm, cnn, gru, and glove. *Systems and Soft Computing*, 6:200110.

Aydın, F. and Aslan, Z. (2019). The construction of a majority-voting ensemble based on the interrelation and amount of information of features. *The Computer Journal*, 63(11):1756–1774.

Ben Brahim, A. and Limam, M. (2018). Ensemble feature selection for high dimensional data: a new method and a comparative study. *Advances in Data Analysis and Classification*, 12(4):937–952.

Bhavitha, Lekshmi Prasad, P., Ani, R., and Deepa, O. (2023). Machine learning based admet prediction in drug discovery. In *2023 4th IEEE Global Conference for Advancement in Technology (GCAT)*, pages 1–9.

Deng, X., Li, M., Deng, S., and Wang, L. (2022). Hybrid gene selection approach using xgboost and multi-objective genetic algorithm for cancer classification. *Medical & Biological Engineering & Computing*, 60(3):663–681.

Díaz-Uriarte, R. and Alvarez de Andrés, S. (2006). Gene selection and classification of microarray data using random forest. *BMC Bioinformatics*, 7(1):3.

Guyon, I., Weston, J., Barnhill, S., and Vapnik, V. (2002). Gene selection for cancer classification using support vector machines. *Machine Learning*, 46(1):389–422.

Kavitha, K. R., Sajith, S., and Variar, N. H. (2022). An efficient boruta-based feature selection and classification of gene expression data. In *2022 IEEE 3rd Global Conference for Advancement in Technology (GCAT)*, pages 1–6.

Kumar, A. and Yadav, J. (2023). A review of feature set partitioning methods for multi-view ensemble learning. *Information Fusion*, 100:101959.

Kumar, V. and Minz, S. (2015). Multi-view ensemble learning: A supervised feature set partitioning for high dimensional data classification. In *Proceedings of the Third International Symposium on Women in Computing and Informatics*, WCI '15, page 31–37, New York, NY, USA. Association for Computing Machinery.

Kumar, V. and Minz, S. (2016). Multi-view ensemble learning: an optimal feature set partitioning for high-dimensional data classification. *Knowledge and Information Systems*, 49(1):1–59.

Menon, R. R., Gayathri, S., and Amina, A. (2023). Representation of documents using minimal dictionary of embeddings. volume 2023-June, page 1897 – 1903. Cited by: 1.

Nidheesh, N., Abdul Nazeer, K., and Ameer, P. (2017). An enhanced deterministic k-means clustering algorithm for cancer subtype prediction from gene expression data. *Computers in Biology and Medicine*, 91:213–221.

R, K. K., Kumar, R. A., and C, M. M. (2024). A maximum relevance minimum redundancy and random forest based feature selection and classification of gene expression data. In *2024 5th International Conference for Emerging Technology (INCET)*, pages 1–5.

R, K. K., S, A. S., and Rasheed, R. (2021). Ensemble-based feature selection using symmetric uncertainty and svm classification. In *2021 2nd Global Conference for Advancement in Technology (GCAT)*, pages 1–6.

Singh, R. and Kumar, V. (2024). Ensemble multi-view feature set partitioning method for effective multi-view learning. *Knowledge and Information Systems*, 66(8):4957–5001.

Skabar, A., Wollersheim, D., and Whitfort, T. (2006). Multi-label classification of gene function using mlps. In *The 2006 IEEE International Joint Conference on Neural Network Proceedings*, pages 2234–2240.

Sreejesh Kumar, V. S., Aparna, K., Ani, R., and Deepa, O. (2021). Ensemble machine learning approaches in molecular fingerprint based virtual screening. In *2021 2nd Global Conference for Advancement in Technology (GCAT)*, pages 1–6.

Xu, Y., Yu, Z., and Chen, C. L. P. (2024). Classifier ensemble based on multiview optimization for high-dimensional imbalanced data classification. *IEEE Transactions on Neural Networks and Learning Systems*, 35(1):870–883.

Zhang, Y., Zhang, H., Cai, J., and Yang, B. (2014). A weighted voting classifier based on differential evolution. *Abstract and Applied Analysis*, 2014(1):376950.