International Journal for Multidisciplinary Research (IJFMR)
E-ISSN: 2582-2160 ● Website: www.ijfmr.com ● Email: editor@ijfmr.com

Enhancing Cancer Classification Through Ensemble Machine Learning and Gene Selection Approaches

Omar Gheni Abdulateef

College Of Literature, University of Samarra, Iraq

Abstract:
The vast dimensionality of gene expression data and the limited number of relevant genes necessitate the adoption of gene selection techniques. Furthermore, the choice of an efficient classifier plays a pivotal role in achieving accurate results. In this study, we employ the Minimum Redundancy Maximum Relevance (mRMR) method for gene selection, coupled with ensemble classifiers and individual classifiers like K-Nearest Neighbors (KNN) and Decision Trees (DT). A comparative analysis between two ensemble classifiers and two individual classifiers is conducted, revealing the superior performance of the ensemble classifiers. Our investigation utilizes four distinct cancer gene expression datasets to showcase the efficacy of employing ensemble classifiers and gene selection methods for cancer classification. The ensemble classifier (Bagging Classifier), in conjunction with the MRMR method selecting only the top 30 genes, achieved an impressive overall accuracy of 94% across all four employed datasets.

Keywords: Gene Expression, Machine Learning, Gene Selection, Cancer Classification.

I. INTRODUCTION

Cancer encompasses a range of conditions arising from the aberrant growth of cells within the human body, initially localized in a specific region and subsequently spreading to other areas [1]. The challenge with these diseases lies in their elusive nature, making early diagnosis a formidable task. To address this, researchers have pioneered various methodologies, leveraging technologies like MIR and CT scans coupled with machine learning and feature selection. These tools serve multifaceted purposes, including cancer classification and diagnosis.

In recent times, advanced techniques such as Microarray and RNA-seq have emerged, offering the means to measure expressed level of gene activity between healthy and unhealthy tissue [2]. These breakthroughs empower researchers to develop innovative approaches for gene expression analysis, facilitating the identification of a select set of biomarker genes. These biomarkers play a pivotal role in early detection, serving as distinctive identifiers. Additionally, they can be employed as training data for classifier methods, further enhancing our ability to detect cancer at its nascent stages [3].

Gene expression data presents unique challenges due to its high dimensionality (a small number of samples with a high number of genes), presence of noisy and duplicate data, and the fact that only a limited number of genes are relevant to the target [4],[5]. Thus, problems including the intricacy of classifiers, labor-intensive early procedures, and the difficulty of attaining exact accuracy impede the analysis of this data. To address these formidable obstacles, researchers have employed gene selection methods, aimed at

reducing dimensionality by identifying a subset of pertinent genes suitable for classifier training [6]. This strategy not only streamlines the process but also improves the accuracy of early-stage analysis.

Moreover, the choice of an appropriate classifier poses another significant challenge in this domain. To tackle this challenge head-on, this study undertook a comprehensive evaluation of various classifiers, encompassing both ensemble methods and individual classifiers. The goal was to ascertain their performance and discern how effectively they can handle the complexities inherent in gene expression data analysis.

The structure of this work is as follows: Section one presentes a comprehensive review of recent publications in the field, offering valuable insights and context. Section two delves into a detailed explanation of the methods employed for classifying cancer gene expression data, shedding light on the techniques that have been utilized. Section three centres on the experimental setup, outlining the parameters and conditions that guided the study. In section four, the experimental process and its results are presented and elucidated, showcasing the empirical findings. Section five is dedicated to a thorough discussion of the outcomes achieved through the application of gene selection and classifier approaches, offering an in-depth analysis. Finally, the study concludes with section six, summarizing key findings and charting a path for future research endeavours.

II. RELATED WORK.

Guillermo et al [15] presented a novel design of convolutional neural network (CNN) to classify 33 cancer types as well as breast cancer subtypes. The proposed model 2D-Hybrid-CNN was compared to two other structures of CNN including 1D-CNN, 2D-Vanilla-CNN to show the effectiveness of developed model. To evaluate the performance of the proposed model, research downloaded the datasets from TCGA repository for 34 classes (33 cancer, 1 normal). The developed model accomplished between (93.5 – 95) accuracy among the 34 classes. While it achieved an average accuracy 88% when employed to five subtypes of breast cancer. Although, the study achieved good results in terms of accuracy when employed 34 classes, however it was under the level when it employed to breast cancer types that may lead that the proposed model not efficient with cancer subtypes. Furthermore, the study was not employed to binary classification that may achieved less accuracy.

Hila et al. [16] presented a novel feature subset selection technique for gene expression classification that makes use of an adaptive neuro-fuzzy inference system. Four independent microarray gene expression datasets—Leukemia, Prostate Cancer, DLBC Stanford, and Colon Cancer—each linked to a distinct type of cancer were used in the analysis to evaluate this methodology. Classification accuracies for the Colon Cancer, Leukaemia, Prostate Cancer, and DLBC Stanford datasets were found to be 89.47%, 83.33%, 80.65%, and 73.33%, respectively, when compared to existing classifiers.

Saba et al. [17] introduced a novel strategy for feature selection aimed at identifying a subset of informative genes. They integrated this method with Support Vector Machine (SVM) techniques to evaluate its effectiveness. This unique approach amalgamates top feature selection techniques from three distinct methodologies: filter, wrapper, and embedded methods. To create a unified list of genes for SVM model training, they incorporated an intersection step. Their study yielded notable results, achieving accuracy rates of 94%, 78.25% for sensitivity, 83.56% for specificity, and 80.9% for the F-measure. It's noteworthy that, despite commendable accuracy rates, certain evaluation metrics, such as sensitivity, exhibited relatively lower values.

Ping et.al [18] developed a novel approach, the Differential Regulation Network Embedded Deep Neural Network (DRE-DNN), to predict the outcomes of liver cancer (hepatocellular carcinoma). This approach was applied to three distinct datasets (GEO GSE10143, GSE14520, and TCGA). The model achieved notable improvements over traditional DNN, as demonstrated by the average AUC values: 86% for GSE10143, 74% for GSE14520, and 72% for TCGA. To evluate the performance of proposed model a diverse array of data sources used. Notably, the study employed a substantial dataset to train the DRE-DNN model, which played a vital role in its effectiveness as a prediction tool. However, despite its advantages in addressing the issue of overfitting, the classification results were not entirely satisfactory. Nonetheless, the DRE-DNN model provides a valuable contribution by mitigating the overfitting challenges often encountered in such predictive models.

Jing et.al [19] introduced an innovative approach for the prediction of four distinct subtypes of breast cancer: Basal, Her2, Luminal A, and Luminal B. Their method, the Multi-Grained Cascade Forest (gcForest), was coupled with a feature selection strategy that depended on the identification of 30 informative genes. This gene selection process aimed to enhance classification accuracy while also reducing training time. The study employed TCGA RNA-Seq data, a valuable resource in the field, to fuel the analysis. To gauge the effectiveness of the gcForest classifier, the researchers compared it with three other machine learning approaches: KNN, Support Vector Machine (SVM), and Multilayer Perceptron (MLP). The results were promising, with the gcForest classifier outperforming the other methods and achieving an accuracy rate of 92%. However, it's worth noting some limitations in this research. The gcForest classifier relies on the decision tree principle, which may not be ideal for processing continuous gene expression data, necessitating data discretization and potentially resulting in information loss. Furthermore, the study did not incorporate external data for a comprehensive evaluation of the proposed model, leaving room for further exploration and refinement.

Yuan et al. [20] employed a combination of Random Forest (RF) and Support Vector Machine (SVM) algorithms to differentiate between two subtypes of lung cancer: Adenocarcinomas (AC) and Squamous Cell Carcinomas (SCC). Additionally, they harnessed Monte-Carlo (MCSF) and incremental feature selection methods to pinpoint genes of significance. The study involved the utilization of Affymetrix U133 arrays, which profiled 20,502 genes, in the analysis of 77 lung AC and 73 lung SCC samples sourced from the Gene Expression Omnibus (GEO GSE43580). Their findings revealed that, for optimal classification employing an SVM classifier, the selection of 1100 features (genes) yielded higher accuracy compared to using a mere 43 informative features (genes) identified by the MCSF method. The accuracy score demonstrated a decrease from 96% to 0.86% when using SVM and from 93% to 88% with RF.

Pineda et.al [21] used ReliefF as feature selection to select a small number of genes that would be used for training Naïve bayes classifier. This study aims to classify lung cancer subtypes that have been collected from TCGA repository. The research achieved 89% accuracy. While Kilicarslan et.al [22] integrated CNN and ReliefF to classify different types of cancer. This study accomplished 83% as an accuracy.

Manzalawy et.al [23], proposed a novel Multi view feature selection as gene selection method to select a subset of important genes therefore applying eXtreme gradient boosting classifier to classify between renal clear cell carcinoma that were collected from TCGA repository. The accuracy score was 76%. The study achieved poor accuracy even though used gene selection method that reduce the dimensionality of the employed dataset. Additionally, the study split the dataset randomly into 20% for testing and 80% for training that may not give real results. Many studies used alternative methods to split the datasets such

cross validation to ensure the result obtained more realistic. Another limitation with study, it has not used other evaluation metric such precision, recall, f1-score to evaluate the performance of the proposed model.

III. RESEARCH METHODOLOGY

This section describes the methods that have been used to meet the goal of this study. This study provides integration of gene selection and classifier approaches to enhance cancer classification and reduce the dimensionality of employed datasets. The study used individual and ensemble classifiers to compare the performance between them when MRMR applied. Additionally, the research used preprocessing stage before applying the data to the gene selection method such as handling missing values, duplication, and preparation of the data to be fitted with ML.

1. Pre-processing

Cancer gene expression data include noise, missing value, and duplication as well as only limited number of genes are associated with the target (disease). This study used pre-process stage to remove the duplication, preparing the data, and handle missing data. Pre-processing stage addressed three main issues, described as follows:

- Handle missing data: gene expression data usually missing the name of the gene symbol , based on that any row of the dataset was not identified with gene symbol will be removed from the data.
- Handle duplication: remove duplicate data from the original data.
- Preparing the data: preparing the data in machine learning format that can be used for classification.

2. Gene selection

Gene expression data is often characterized by high dimensionality, indicating a small number of samples and a large number of features (genes). Furthermore, only a fraction of these genes may be pertinent to a particular class or disease. To mitigate these challenges, the Minimum Redundancy Maximum Relevance (mRMR) method is employed as a gene selection technique. By reducing dimensionality, addressing overfitting concerns, and enhancing cancer classification accuracy, mRMR proves instrumental in this context [7].

The primary objective of the mRMR method is to pinpoint a subset of features that exhibit the strongest correlation with a specific class (relevance), while simultaneously minimizing correlations among the selected features (redundancy). This ranking process hinges on the principles of minimal redundancy and maximal relevance. Relevance is typically quantified using the F-statistic (for continuous features) or mutual information (for discrete features), while redundancy is evaluated through metrics like the Pearson correlation coefficient (for continuous features) or mutual information (for discrete features).

A. Classifier Approaches

1. Bagging Classifier (BC)

BC is a parallel ensemble learning technique where multiple base learners are simultaneously created. It used to improve the accuracy and robustness of classification models and reduce variance and prevent overfitting in machine learning models [8]. The main principle behind bagging is that you can lower the chance of overfitting and enhance the model's generalization ability by training many models on various subsets of data and then combining their predictions. The ensemble's resilience is increased by the variety of base models.

2. Gradient Boosting Classifier (GB)

GB is considered one of the machine learning approaches that employed for both classification and

regression challenges. It's also a kind of ensemble learning technique that implemented by a combination of multiple weaker models, typically decision trees. In general, it aims to reduce the prediction errors and increase the accuracy [9]. The "Gradient" in Gradient Boosting comes from the use of gradient descent optimization to minimize the loss function. By iteratively fitting models to the residuals and adjusting the model parameters, Gradient Boosting aims to make successive models focus on the examples that the previous models found challenging to predict accurately.

3. K-Nearest Neighbors (KNN)

A straightforward and user-friendly machine learning approach for classification and regression applications is K-Nearest Neighbors (KNN) [10]. This kind of algorithm is known as instance-based or lazy learning, and it bases its predictions on the data points, or instances, in the training set. Since KNN is a non-parametric method, it operates without making any implicit predictions about the data's distribution.

4. Decision Trees (DT)

A well-liked machine learning approach for classification and regression applications is the decision tree [11]. It is a supervised learning technique that may be applied to a range of tasks including data processing and judgment [12]. Decision trees are especially helpful for jobs that need a series of choices and have a structure like a tree. At the base of the tree, the algorithm begins with the full dataset. Decision trees' interpretability is its main benefit. The model's decisions and rules are simple to comprehend and illustrate. They are frequently employed in domains like fraud detection, credit risk assessment, and medical diagnosis where explainability and transparency are essential.

B. Evaluation Metrics

In this experiment four evaluation metrics have been used to evaluate how a classifier is. The primary motivation for employing a variety of evaluation metrics, as opposed to solely relying on accuracy, stems from the inherent limitations of accuracy when dealing with imbalanced datasets. The reason for this is quite clear: accuracy alone may not provide a true reflection of a model's performance. Let's illustrate this with a practical example. Consider a dataset where 90% of the cases represent a specific class, such as instances of lung cancer, while only 10% of the dataset contains normal lung cases. If a classifier was to predict all instances as lung cancer, it could still achieve a remarkable 90% accuracy. However, this is clearly not an effective or desirable outcome, as it ignores the 10% of the data that is correctly classified as normal. These metrics described as follows [13].

1. Accuracy (Ac): Ac measures the ratio of correctly predicted instances to the total number of instances in the dataset. It can be calculated as follows. Mathematically calculated as below [13].

$$Ac = \frac{TP + TN}{TP + FP + TN + FN} \qquad (1)$$

True Positive (TP) signifies instances where the model correctly identifies and predicts a positive class. True Negative (TN) represents instances in which the model accurately predicts a case as a member of the negative class. This means that non-cancerous cases, for instance, are correctly classified as such by the model. On the flip side, False Positive (FP) occurs when the model erroneously predicts a positive class. False Negative (FN) is the outcome when the model incorrectly identifies a case as a member of the negative class.

2. Precision (Pre): Pre is a statistical metric that is frequently applied to data analysis, statistics, and machine learning. It gauges how accurately a model or system makes favorable predictions. In binary classification tasks, where you have to discriminate between two classes, like "positive" and "negative," precision is especially important. Mathematically calculated as below [13].

$$Pre = \frac{TP}{TP+FP} \quad (2)$$

3. Recall (Rec): Re is the percentage of real positive occurrences that were accurately predicted to be positive. Mathematically calculated as below [13].

$$Rec = \frac{TP}{TP+FN} \quad (3)$$

4. F1-score (F1): F1 is combined precision and recall into a single value. It provides a balance between these two metrics and is especially useful when both false positives and false negatives in the evaluation of a model's performance are important. Mathematically calculated as below [13].

$$F1 = 2 * \frac{Precision * Recall}{Precision+Recall} \quad (4)$$

C. Cross Validation (CV)

In machine learning and data analysis, cross-validation is a statistical method applied to evaluate a predictive model's effectiveness. To evaluate a model's robustness and spot possible problems like overfitting, its main goal is to determine how well the model would generalize to an independent dataset. Several subsets of the dataset are divided into, and the model is trained and tested on these subsets before the results are aggregated in cross-validation. It evaluates a model on several subsets of the data, which yields a more reliable assessment of the model's performance [14]. It aids in overfitting detection. A model may be overfitted if it performs exceptionally well on training data but badly on validation data across several cross-validation folds. Since all data points are used for both training and testing in several rounds, it makes the most use of the given data. In this experiment , 5-fold cross validation used to split the datasets that have been employed for cancer gene expression as described in Fig 1.

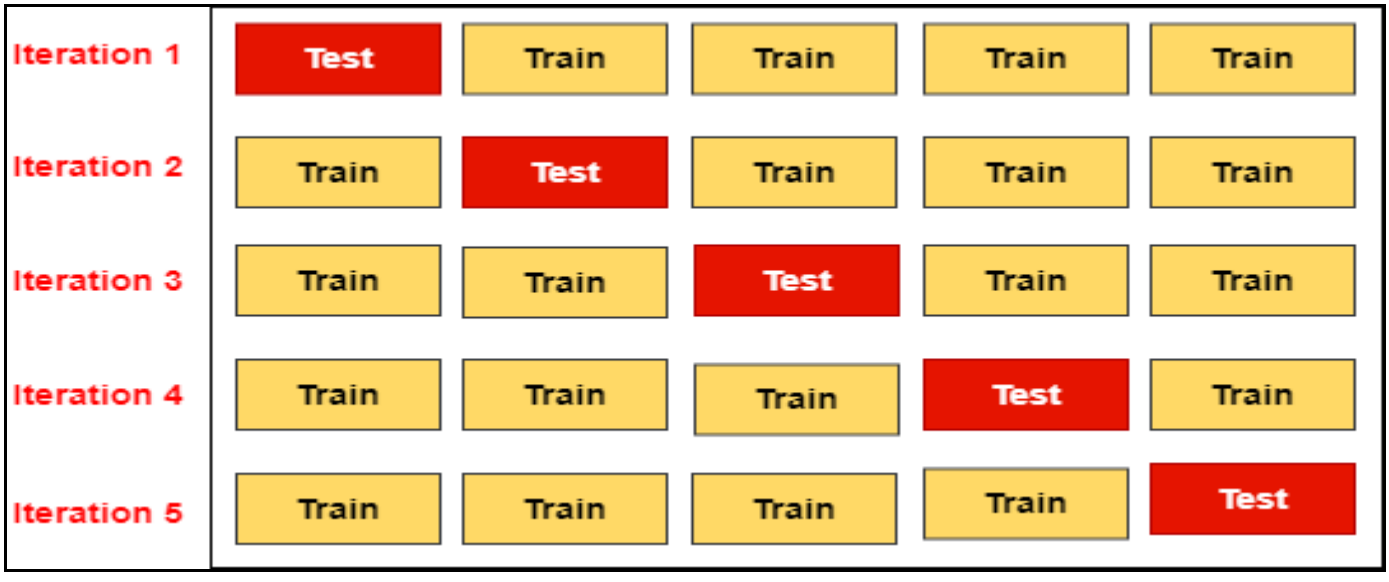

Fig 1. 5 k fold cross validation

IV. FRAMEWORK EXPERIMENTAL

In this experiment, a structured approach comprises four stages. The initial stage involves data collection from RNA-seq and Microarray datasets, addressing issues such as missing and duplicated data in a process referred to as pre-processing. The second stage utilizes the MRMR method to select the top 30 optimal genes, which are crucial for training the classifiers. The third stage applies four classifier methods, including ensemble and individual classifiers, to make predictions based on the selected genes. In the final stage, the performance of these classifiers is assessed using metrics like accuracy, precision, recall, and F1-score, allowing for a comparison of their effectiveness as illustrated in Fig 2. This systematic approach facilitates informed comparisons between the classifiers and meaningful conclusions about their performance.

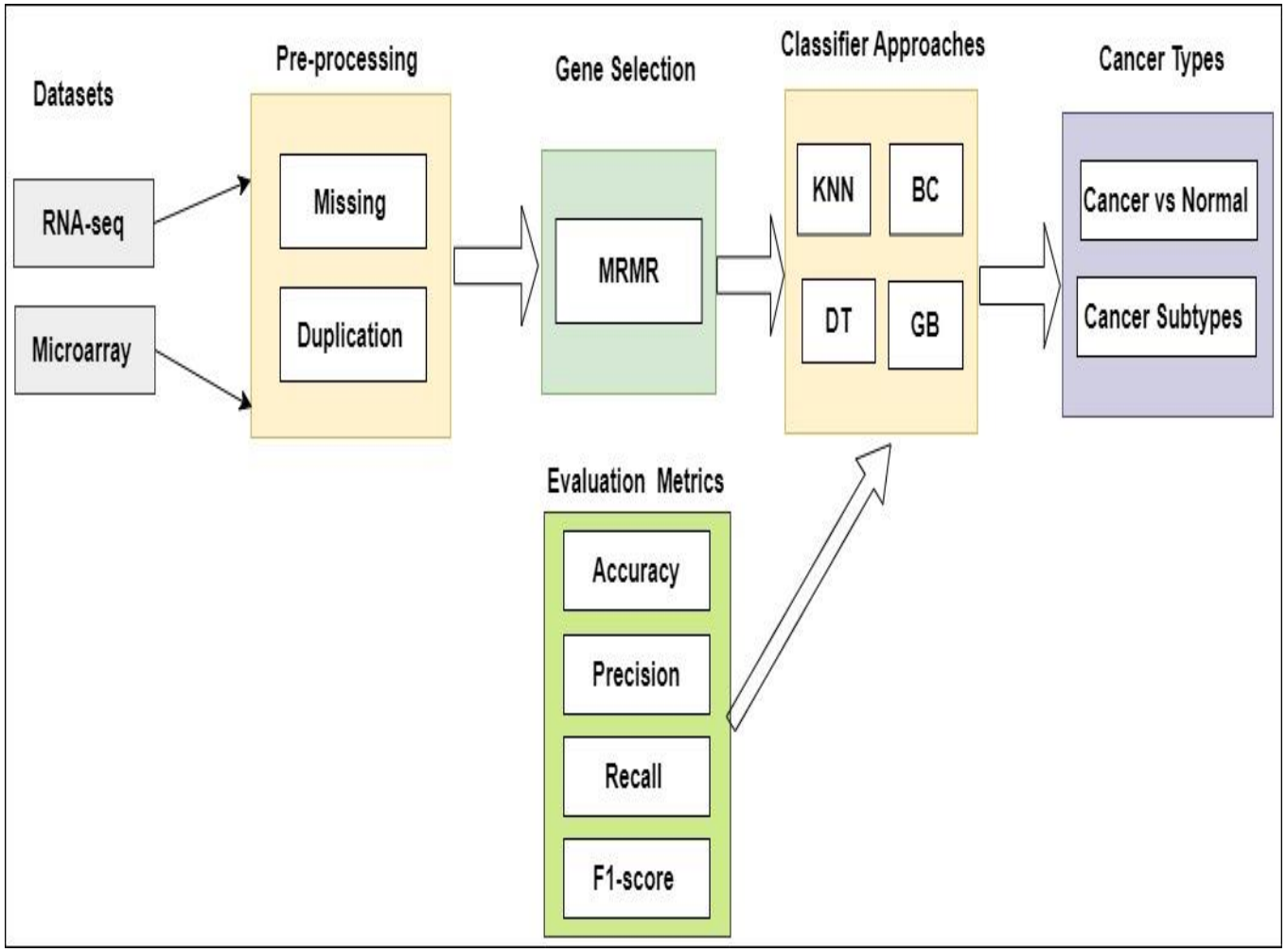

Fig 2. Experimentational Framework.

V. EXPERIMENTAL SET UP

A constructed the ensemble classifier model was implemented using Python software, taking advantage of the computational prowess offered by an Intel Core i7-8565U processor and a substantial 32 GB of RAM. To comprehensively evaluate the model's efficacy, we conducted a thorough analysis across a range of cancer types. This rigorous assessment involved the utilization of four datasets, consisting of two microarray and two RNA-seq datasets. Employing a cross-validation methodology, we meticulously partitioned the datasets into training and testing subsets, ensuring a robust evaluation of model generalization and the generation of dependable results.

VI. Experimental Results

1. Original Datasets Used.

The datasets that have been employed in this experiment were collected from The Cancer Genome Atlas (TCGA) and Gene Expression Omnibus (GEO) repositories. The datasets involve binary and multi classes as well as Microarray and RNA-seq datasets to show the performance of the used classifiers. Table1 describe in detail the datasets employed

Table 1: Original datasets description have been used for evaluating the proposed model.

Dataset Name	Measurement	No.Genes	No.class	Number of samples	Reference
Breast cancer subtypes	RNA-seq	20531	5	964	[24]
Lung cancer	Microarray	54675	2	150	[20]
Five cancer types	RNA-seq	972	5	2086	[24]
Leukmia	Microarray	22284	2	64	[25]

2. Finding Results

Table 2 presents the comparing performance of the four classifiers which integrated with gene selection (mRMR) method to select the top 30 genes for each dataset. The findings demonstrated that Bagging Classifier (BC) outperformed the other classifiers in all employed datasets. Consequently, ensemble classifier achieved better results than individual classifiers such as DT and KNN as described in the table below. 5K-fold cross validation method has been used to evaluate the classifier methods. While BC demonstrated the highest accuracy in most of the utilized datasets, it did not yield the best performance in the Leukemia dataset, despite achieving 100% accuracy. The remaining evaluation metrics ranged between 96.3 to 95.3. In contrast, KNN also achieved 100% accuracy and outperformed BC in the context of this dataset when considering other evaluation metrics. This underscores the importance of not relying solely on accuracy as the sole criterion for classifier evaluation.

Table 2: Results achieved by applying individual classifiers vs ensemble classifiers.

Datasets	**Classifier**	**Accuracy %**	**Precision %**	**Recall %**	**F1-score %**
Breast cancer subtypes	DT	77	76	72	73
	KNN	74.6	73.8	73	73
	GB	79	78	76	77
	BC	**83**	**82**	**80**	**80.5**
Lung cancer	DT	84.6	83	86	84
	KNN	84.6	94.7	72	81
	GB	86.6	82.7	93	87
	BC	**88.6**	**93.8**	**82**	**87**
Five cancer types	DT	92	88.7	87.5	87.5
	KNN	93	90	88	88.7

	GB	89	76.7	79	77.5
	BC	**95**	**92**	**92**	**92**
Leukemia	DT	75	77.8	80	75
	KNN	**100**	**100**	**100**	**100**
	GB	94	92.6	92	92
	BC	100	96.3	96.4	95.3

VII. DISCUSSION

This study offers a comparative analysis of integrating gene selection with both individual and ensemble classifiers. The results conclusively reveal the superior performance of ensemble classifier approaches, with a particular highlight on the Bagging Classifier (GB). In this section, we delve into the outcomes obtained from employing both ensemble and individual classifiers. The assessment is presented as follows:

As depicted in Fig 3, we present the accuracy scores achieved by the four classifiers across 5 different folds in classifying breast cancer subtypes. The findings are quite compelling, demonstrating that the Bagging Classifier (BC) outperformed the other classifiers, securing the highest accuracy rate of 88%. On the opposite end of the spectrum, the k-Nearest Neighbors (KNN) classifier recorded the lowest accuracy at 74%. These outcomes underscore the effectiveness of ensemble classifiers, particularly BC, in enhancing classification accuracy for breast cancer subtypes.

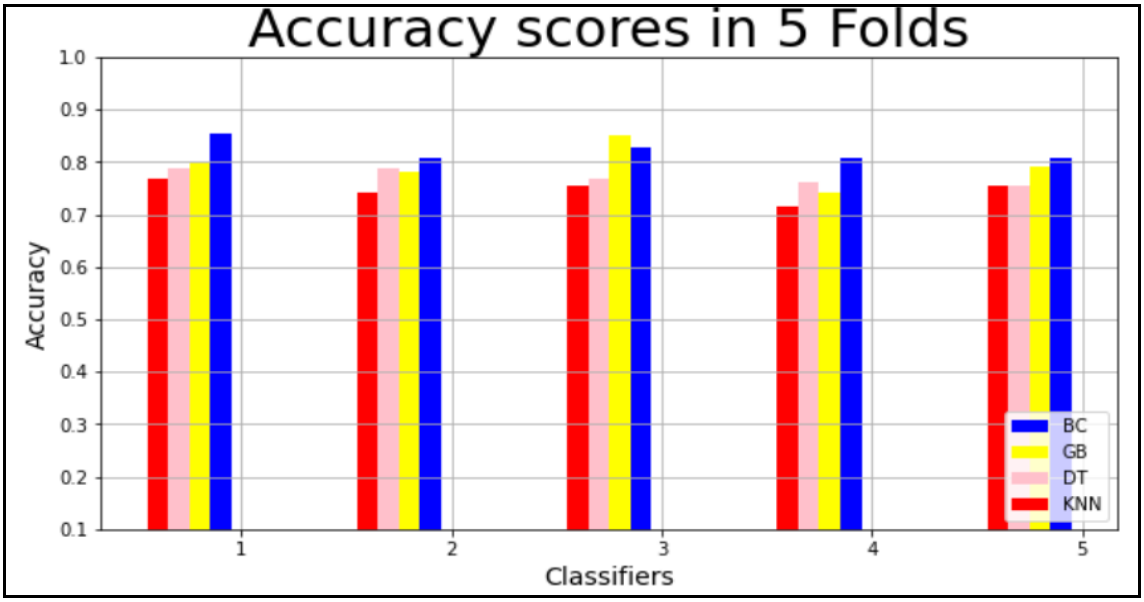

Fig 3: Accuracy scores for the four classifiers in 5 folds cross validation (subtypes of breast cancer)

Fig 4 provides a visual representation of the accuracy scores achieved by the four classifiers over five distinct folds for the classification of lung cancer. The insights uncovered in this analysis paint a vivid picture, reaffirming the Bagging Classifier (BC) as a standout performer with an impressive accuracy rate of 88.6%. In stark contrast, the Decision Tree (DT) classifier posted the lowest accuracy score at 83%. The comparative aspect of these findings is particularly noteworthy, underscoring the pivotal role of ensemble classifiers, and in this context, BC, in significantly enhancing classification accuracy for lung cancer.

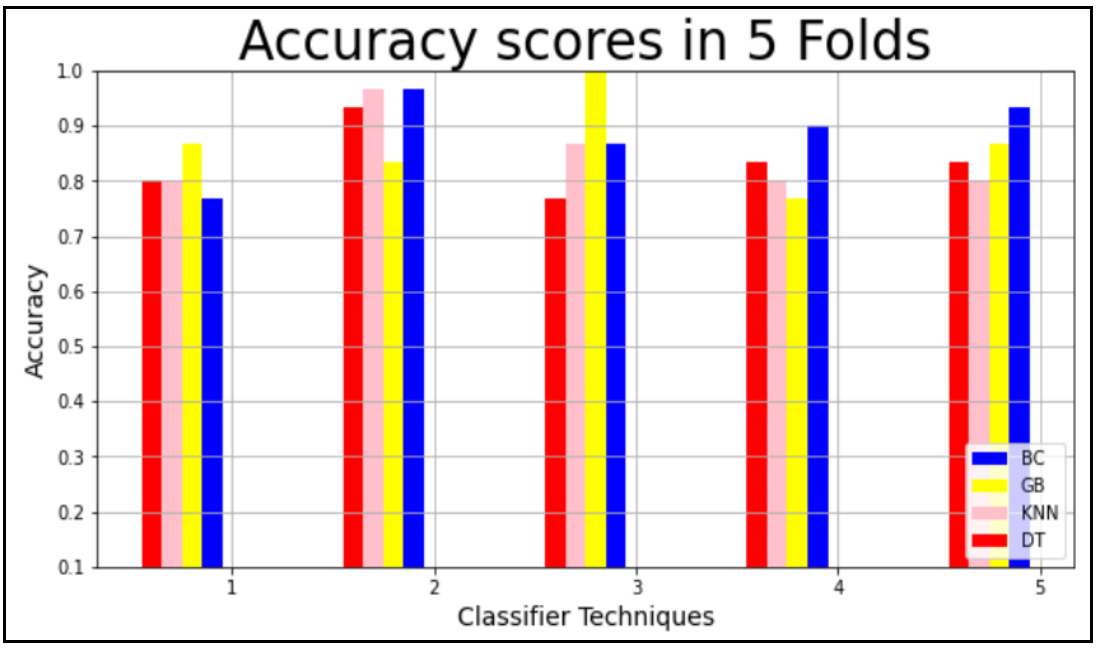

Fig 4: Accuracy scores for the four classifiers in 5 folds cross validation (Lung cancer)

Fig 5 presents a graphical overview of accuracy scores across five separate cross-validation folds when classifying five different cancer types. The analysis reveals clear trends. BC stands out as a top performer, boasting an impressive accuracy rate of 95%. In contrast, the GB classifier records the lowest accuracy score at 89%. Moreover, BC demonstrated notable enhancements in various evaluation metrics, achieving 92% for precision, recall, and F1-score, in sharp contrast to GB, which yielded scores of 76.7%, 79%, and 77.5% for precision, recall, and F1-score, respectively. These comparative results emphasize the substantial impact of ensemble classifiers, particularly BC, in elevating accuracy levels in five cancer types of classification.

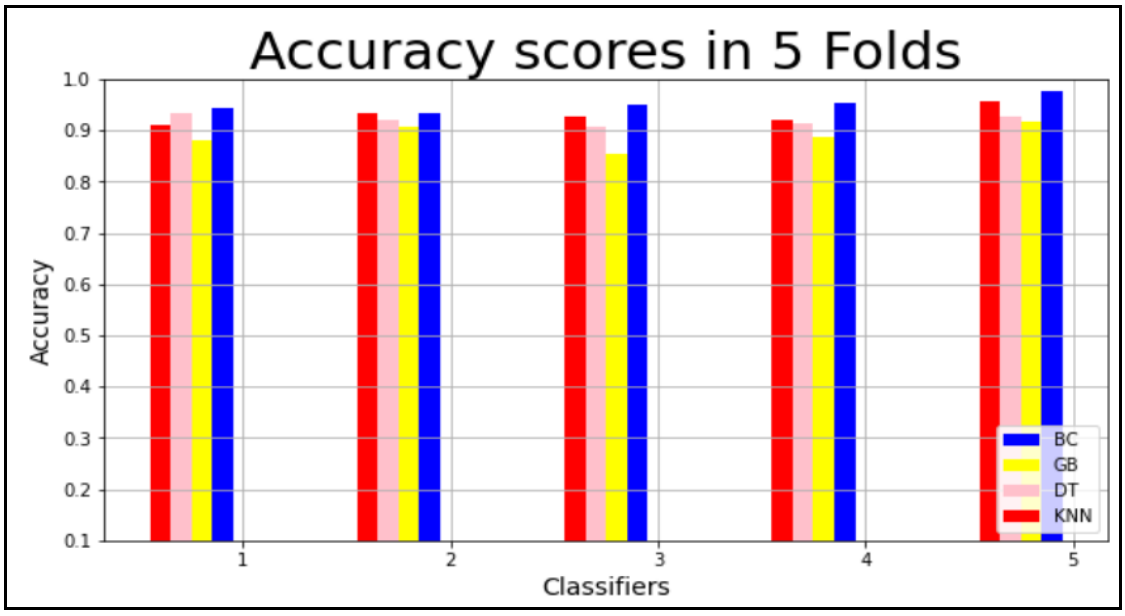

Fig 5: Accuracy scores for the four classifiers in 5 folds cross validation (Five cancer types)

Fig 6 presents a comprehensive view of the results pertaining to accuracy scores when employing four different classifier approaches across a five-fold classification of leukemia cancer expression data. The

results offer valuable insights into the performance of these classifiers. Notably, KNN emerged as the leading classifier in this scenario, surpassing the other approaches. While BC achieved a similar level of accuracy, a closer examination of additional evaluation metrics, including precision, recall, and F1-score, revealed that KNN outperformed BC. This difference in performance underscores the superiority of KNN in effectively classifying this dataset, as it demonstrated a more balanced and robust classification performance.

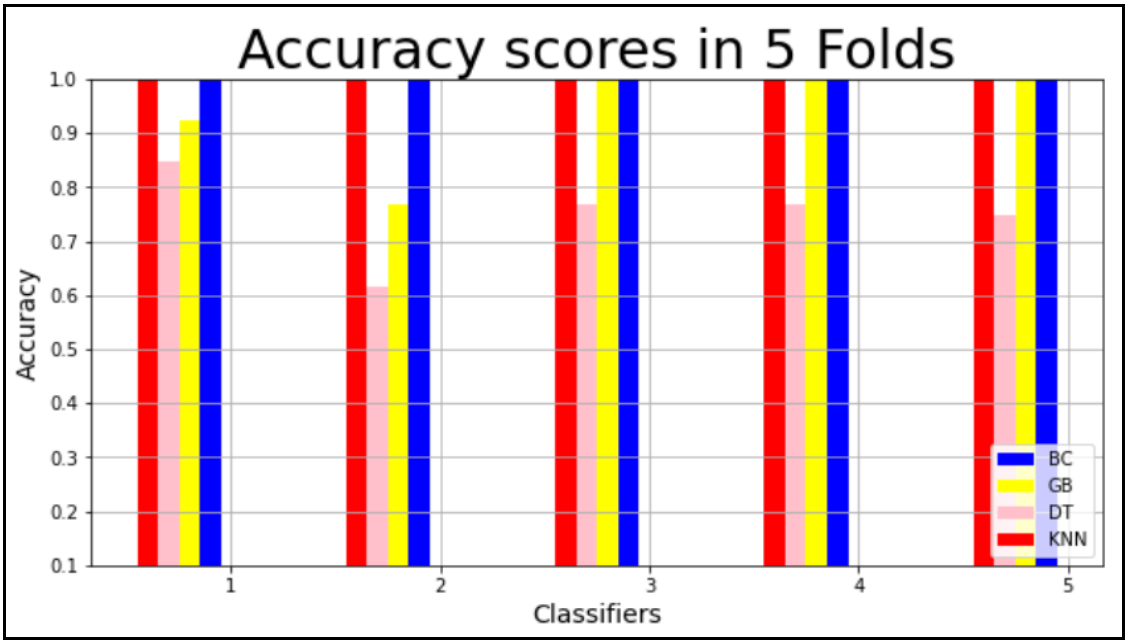

Fig 6: Accuracy scores for the four classifiers in 5 folds cross validation (Leukemia)

VIII. CONCLUSION

This study presents a comprehensive review of the recent work in cancer classification using gene expression data and machine learning. It also come up with comparative results between ensemble and individual classifiers. The findings demonstrated that ensemble classifiers surpassed the individual classifiers in all three employed datasets. The ensemble classifier (Bagging Classifier), integrated with MRMR method for selecting only the top 30 genes, accomplished overall accuracy of 94% across all four employed datasets.

While this study successfully achieved promising accuracy in the classification of cancer across four different expression datasets, it is essential to acknowledge certain limitations that can be addressed in future research. Firstly, the experiment utilized a restricted dataset for comparison. Future work should aim to encompass a more comprehensive set of 33 cancer types available through TCGA to enhance the scope of the study. Secondly, the study employed a limited number of genes by selecting the top 30. It is important to recognize that utilizing a fixed number of genes across various datasets may not yield optimal results. In future research, a more flexible feature selection method should be considered, allowing for the adaptation of the number of genes based on the characteristics of each dataset individually.

REFERENCES

1. S. Shandilya and C. Chandankhede, "Survey on recent cancer classification systems for cancer diagnosis", Proc. Int. Conf. Wireless Commun. Signal Process. Netw. (WiSPNET), pp. 2590-2594, Mar. 2017.

2. Md Maniruzzaman et al., "Statistical characterization and classification of colon microarray gene expression data using multiple machine learning paradigms", Computer methods and programs in biomedicine, vol. 176, pp. 173-193, 2019.
3. J. Pati, "Gene Expression Analysis for Early Lung Cancer Prediction Using Machine Learning Techniques: An Eco-Genomics Approach", IEEE Access, vol. 7, pp. 4232-4238, 2019.
4. Mahendran, N., Durai Raj Vincent, P., Srinivasan, K., and Chang, C.-Y. (2020). Machine learning based computational gene selection models: a survey, performance evaluation, open issues, and future research directions. Frontiers in genetics, 11:603808.
5. Yaping, Z. and Changyin, Z. (2021). Gene feature selection method based on relief and pearson correlation. In 2021 3rd International Conference on Applied Machine Learning (ICAML), pages 15–19.
6. M. Khalsan, M. Mu, E. S. Al-Shamery, L. Machado, M. O. Agyeman and S. Ajit, "Intersection Three Feature Selection and Machine Learning Approaches for Cancer Classification," 2023 International Conference on System Science and Engineering (ICSSE), Ho Chi Minh, Vietnam, 2023, pp. 427-433, doi: 10.1109/ICSSE58758.2023.10227163.
7. Radovic, M., Ghalwash, M., Filipovic, N. et al. Minimum redundancy maximum relevance feature selection approach for temporal gene expression data. BMC Bioinformatics 18, 9 (2017). https://doi.org/10.1186/s12859-016-1423-9.
8. Y. Lv et al., "A classifier using online bagging ensemble method for big data stream learning," in Tsinghua Science and Technology, vol. 24, no. 4, pp. 379-388, Aug. 2019, doi: 10.26599/TST.2018.9010119.
9. Guillen, Maria D., Juan Aparicio, and Miriam Esteve. "Gradient tree boosting and the estimation of production frontiers." Expert Systems with Applications 214 (2023): 119134.
10. S. Zhang, X. Li, M. Zong, X. Zhu and R. Wang, "Efficient kNN Classification With Different Numbers of Nearest Neighbors," in IEEE Transactions on Neural Networks and Learning Systems, vol. 29, no. 5, pp. 1774-1785, May 2018, doi: 10.1109/TNNLS.2017.2673241.
11. Khalsan, Mahmood, et al. "Fuzzy Gene Selection and Cancer Classification Based on Deep Learning Model." arXiv preprint arXiv:2305.04883 (2023).
12. Costa, Vinícius G., and Carlos E. Pedreira. "Recent advances in decision trees: An updated survey." Artificial Intelligence Review 56.5 (2023): 4765-4800.
13. M. Khalsan et al., "A Survey of Machine Learning Approaches Applied to Gene Expression Analysis for Cancer Prediction," in IEEE Access, vol. 10, pp. 27522-27534, 2022, doi: 10.1109/ACCESS.2022.3146312.
14. M. Khalsan, M. Mu, E. S. Al-Shamery, S. Ajit, L. Machado and M. O. Agyeman, "A Novel Fuzzy Classifier Model for Cancer Classification Using Gene Expression Data," in IEEE Access, doi: 10.1109/ACCESS.2023.3325381.
15. Mostavi, M., Chiu, YC., Huang, Y. et al. Convolutional neural network models for cancer type prediction based on gene expression. BMC Med Genomics 13 (Suppl 5), 44 (2020). https://doi.org/10.1186/s12920-020-0677-2.
16. Hilal, A. M., Malibari, A. A., Obayya, M., Alzahrani, J. S., Alamgeer, M., Mohamed, A., Motwakel, A., Yaseen, I., Hamza, M. A., Zamani, A. S., et al. (2022). Feature subset selection with optimal adaptive neuro-fuzzy systems for bioinformatics gene expression classification. Computational Intelligence and Neuroscience, 2022.

17. Bashir, S., Khattak, I. U., Khan, A., Khan, F. H., Gani, A., and Shiraz, M. A novel feature selection method for classification of medical data using filters, wrappers, and embedded approaches. Complexity, 2022.
18. J. Li, Y. Ping, H. Li, H. Li, Y. Liu, B. Liu, and Y. Wang, "Prognostic prediction of carcinoma by a differential-regulatory-network-embedded deep neural network," Comput. Biol. Chem., vol. 88, Oct. 2020, Art. no.107317.
19. J. Xu, P. Wu, Y. Chen and L. Zhang, "Comparison of Different Classification Methods for Breast Cancer Subtypes Prediction," 2018 International Conference on Security, Pattern Analysis, and Cybernetics (SPAC), Jinan, China, 2018, pp. 91-96, doi: 10.1109/SPAC46244.2018.8965553.
20. F. Yuan, L. Lu and Q. Zou, "Analysis of gene expression profiles of lung cancer subtypes with machine learning algorithms", Biochimica et Biophysica Acta (BBA) Mol. Basis Disease, vol. 1866, no. 8, Aug. 2020.
21. L. Pineda, H. A. Ogoe, J. B. Balasubramanian, C. R. Escareño, S. Visweswaran, J. G. Herman, and V. Gopalakrishnan, ''On predicting lung cancer subtypes using 'omic' data from tumor and tumoradjacent histologically-normal tissue,'' BMC Cancer, vol. 16, no. 1, p. 184, Dec. 2016.
22. S. Kilicarslan, K. Adem, and M. Celik, ``Diagnosis and classification of cancer using hybrid model based on ReliefF and convolutional neural network," Med. Hypotheses, vol. 137, Apr. 2020, Art. no. 109577.
23. Y. El-Manzalawy, ``CCA based multi-view feature selection for multiomics data integration," in Proc. IEEE Conf. Comput. Intell. Bioinf. Com- put. Biol. (CIBCB), Jun. 2018, pp. 1-8.
24. Weinstein, J.N., Collisson, E.A., Mills, G.B., Shaw, K.R.M., Ozenberger, B.A., Ellrott, K., Shmulevich, I., Sander, C., Stuart, J.M. and Cancer Genome Atlas Research Network, 2013. The cancer genome atlas pan-cancer analysis project. Nature genetics, 45(10), p.1113.
25. Edgar R, Domrachev M, Lash AE. Gene Expression Omnibus: NCBI gene expression and hybridization array data repository Nucleic Acids Res. 2002 Jan 1;30(1):207-10