Spring 2024

Advancing Cancer Classifcation through Machine Learning Analysis of RNA-Seq Gene Expression Data

Emil Agbemade
University of Central Florida, emil.agbemade@ucf.edu

Amina Issoufou Anaroua
University of Central Florida, amina.issoufouanaroua@ucf.edu

Dimitri Bamba
University of Central Florida, di714545@ucf.edu

STARS Citation

Agbemade, Emil; Anaroua, Amina Issoufou; and Bamba, Dimitri, "Advancing Cancer Classifcation through Machine Learning Analysis of RNA-Seq Gene Expression Data" (2024). *Data Science and Data Mining*. 16.
https://stars.library.ucf.edu/data-science-mining/16

Advancing Cancer Classification through Machine Learning Analysis of RNA-Seq Gene Expression Data

1st Emil Agbemade
Dept. of Statistics and Data Science
University of Central Florida
Orlando, USA
emil.agbemade@ucf.edu

***Abstract*—This study delves into the classification of various cancer types using the RNA-Seq (HiSeq) PANCAN dataset from the UCI Machine Learning Repository, which encompasses a rich collection of gene expression data across multiple tumor samples. To improve cancer diagnosis and treatment, our methodology confronts the challenges inherent in high-dimensional datasets, such as the Hughes Effect and the Curse of Dimensionality, through innovative feature selection methods and machine learning approaches. A key component of our strategy includes the use of tree-based algorithms, particularly Random Forest, to refine the dataset to seventy genes of utmost relevance for tumor classification, and the application of PCA and Kernel PCA for dimensional reduction, enabling the visualization of non-linear patterns in gene expression data. The research further investigates the gene interaction network through network analysis, employing modularity metrics to understand significant community structures linked to biological processes in cancer. Our model evaluation assesses various machine learning models, highlighting the precision and low-test error rates of SVM, Logistic Regression, and KNN, suggesting their effectiveness in exploiting the dataset's inherent separability. The study's comprehensive approach not only provides a systematic framework for analyzing gene expression data but also paves the way for advanced research into the genetic mechanisms of cancer, with implications for personalized medicine and treatment strategies.**

***Index Terms*—Cancer classification, gene expression data, RNA-Seq, machine learning, feature selection, dimensionality reduction, network analysis.**

I. Background and Motivation

Microarrays and RNA-Seq are two examples of the cutting-edge biotech tools that have allowed researchers to capture gene expression data from tissues. In order to improve diagnosis and treatment optimization, these methodologies allow for the separation between healthy and sick states, as well as between different types and subtypes of cancer [1], [2], [25]. But there are obstacles to cancer categorization, like the Hughes Effect and the Curse of Dimensionality [5], [6]. This problem occurs when there are more genes in DNA sequences than there are samples, which causes classification algorithms to be less effective because of irrelevant genes [7], [8].

To tackle this, a lot of research has focused on picking out a small number of important genes for categorization. Feature selection methods in machine learning have been used to try to identify these genetic markers [3], [7]–[12]. One example is the use of classification effectiveness to rank features in filter selection algorithms. When doing this, Pavithra et al. [10] made use of mutual information, whereas Guyon et al. [11] used a recursive Support Vector Machine (SVM). Wrapper approaches, on the other hand, are more computationally intensive, but they choose features according to classifier performance [8]. To identify effective feature groups, these methods are investigated: Genetic Algorithms (GA), Support Vector Machines (SVM) [14]–[16], K-Nearest Neighbors [16], Local Search [17], Tabu Search [18], Particle Swarm Optimization (PSO), and Artificial Bee Colony (ABC) [15], [27]. In terms of feature extraction, classification accuracy, and stability, Zhu et al. [15] discovered that GA was significantly more effective than GA-SVM, PSO-SVM, and ABC-SVM. The results obtained by Singh et al. [19], [24], [26] using GA were superior to those obtained by Local Search [17] and Tabu Search [18], [28].

When employing microarrays or RNA-Seq for cancer classification, GA has proven particularly helpful in selecting appropriate gene sets [1], [10], [20], [21]. Several gene expression cancer RNA-Seq datasets have been made available for research purposes through public databases such as GEO and the UCI Machine Learning Repository Repository [21], [23]. Using training and testing datasets of identical size, Zhang et al. [21] achieved an accuracy of 81.54% when classifying peripheral blood data using an SVM method. Using data from the Cancer Genome Atlas Pan-cancer Analysis Project, Salman et al. [22], [29] investigated a Hybridized Genetic Algorithm with an Artificial Neural Network with 4 hidden layers (ANN+GA) without first selecting any features. This method took into account the impact of all features when adjusting the weights of the neural network.

Our study reflects a culmination of rigorous methodologies applied to dissect the complexities of gene expression data in cancer classification. We distilled a comprehensive list of seventy genes, which are most indicative of various tumor types. This selection was carefully conducted using a tree-based method, specifically the Random Forest algorithm, underlining our methodical approach to identifying critical genetic markers for cancer typology. Employing Principal Component Analysis (PCA) and Kernel PCA (KPCA), we successfully navigated the hurdles posed by high-dimensional data, unearthing hidden non-linear relationships within the gene expression data. It was the application of Kernel PCA that highlighted the intricate non-linear dynamics at play, challenging the sufficiency of linear analysis methods for this dataset and advocating for a detailed approach that embraces its complexity. The modularity analysis of our gene interaction network revealed a meaningful community structure, transforming the network into a narrative-rich canvas. The quantitative findings are significant, demonstrating that the network's community structures are intricately linked to the underlying biological processes that drive cancer, far exceeding the likelihood of random occurrence. Our machine learning model evaluations revealed a clear stratification in performance. Notably, SVM, Logistic Regression, and KNN exhibited exceptional performance with remarkably low test errors, suggesting that the dataset possesses an inherent separability skillfully leveraged by these models. Conversely, Random Forest and XGBoost, while robust, indicate a potential for overfitting, necessitating further optimization. The effectiveness of these models ignites discussions on computational efficiency, adaptability to various data types, and the balance between model complexity and interpretability. In essence, our study has not only provided a blueprint for analyzing gene expression data but has also illuminated the path for further research into the genetic underpinnings of cancer. The findings advocate for targeted biological pathway analyses and experimental validations, potentially contributing to the advancement of personalized cancer treatment strategies and the broader field of precision medicine.

II. Dataset

The ”RNA-Seq (HiSeq) PANCAN Data” dataset from the UCI Machine Learning Repository is an invaluable asset in cancer research, particularly for the classification and study of various cancer types. Comprising gene expression data from a wide array of tumor samples, the dataset is substantial, featuring 801 samples and 16,384 features. Its high dimensionality presents both a rich field for discovery and a significant computational challenge. A notable aspect of this dataset is the highly imbalanced nature of the target variable, necessitating the use of downsampling techniques to balance the classes. This is crucial to avoid overfitting issues that typically arise when working with highly imbalanced data. After implementing downsampling, each class is represented with 78 points, providing a more balanced framework for analysis as shown in Figure 4. Researchers can use machine learning algorithms on this balanced and detailed genomic dataset to try out different feature selection methods and make accurate models for classifying cancer. This makes it an important tool for learning more about cancer genomics and finding more personalized ways to treat it.

III. Exploratory Analysis

We looked at how four carefully picked genes were spread across different types of cancer to learn more about genetic trends in different types of cancer. Our data showed that the high number of genes in a BRCA type shows that there is a strong link between those genes and the development or progression of BRCA.

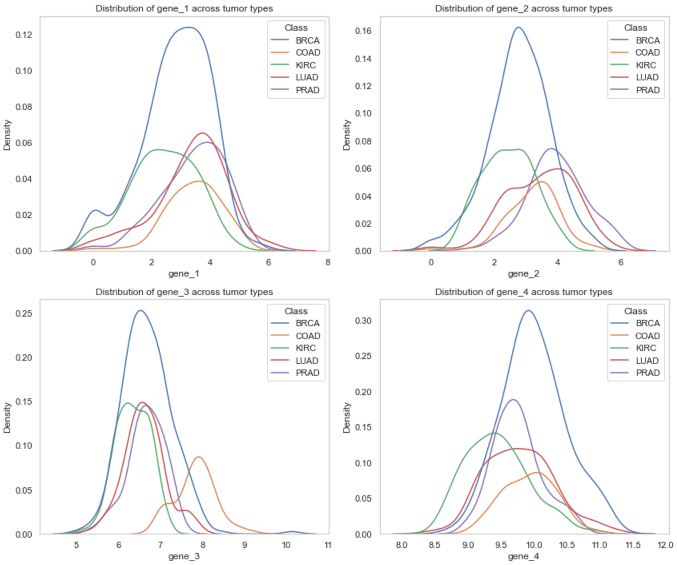


Fig. 1. Distribution of Sample genes across Tumor types

This could mean that these genes are very important to the biology of BRCA. This also gives us a better understanding of how BRCA works at the molecular level. Figuring out why the genes are overrepresented could help us learn important things about the tumor's biology, like how cells grow, survive, or spread.

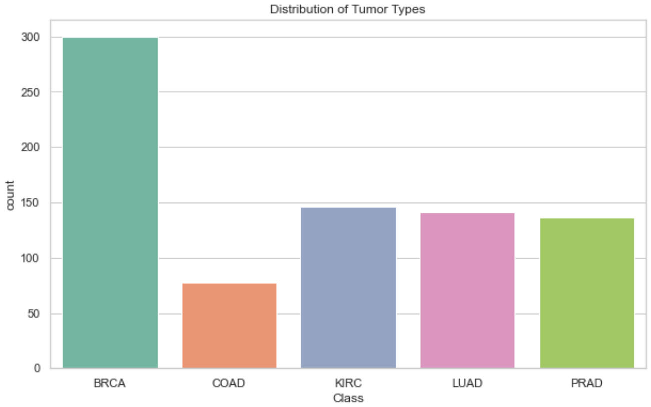


Fig. 2. Distribution of tumor types

The preliminary examination of the types of tumors has shown that the BRCA tumor has the highest count, while the COAD tumor has the lowest count. According to our findings, KIRK, LUAD, and PRAD are all dispersed in an equal manner. Since this is the case, there is a large imbalance between the various types of tumors. First and foremost, it is essential to treat the imbalance issue to avoid overfitting the model.

Figure 3 provide an in-depth explanation of a correlation heatmap, specifically for gene expression analysis. The diagonal line in the heatmap shows each gene correlates perfectly with itself, denoted by a uniform color, usually indicating the highest correlation value of 1.0. This feature is standard in correlation heatmaps and acts as a reference for interpreting other data. The paragraph then explains the adjacent color scale, crucial for interpretation, with a gradient from blue to red and a midpoint often in white. Blue represents negative correlations, possibly as strong as -1.0, indicating an inverse relationship between gene expressions. White signifies no correlation, while red indicates positive correlations, up to a perfect positive correlation of 1.0.

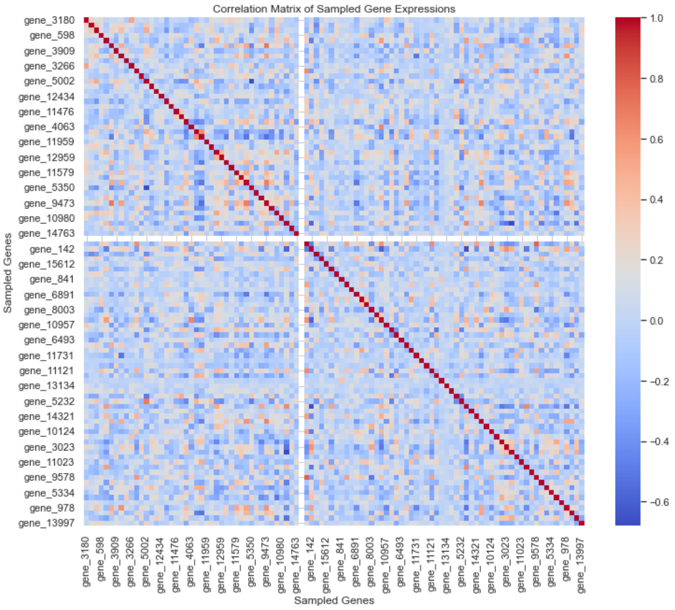


Fig. 3. Correlation Analysis

Furthermore, each square in the heatmap represents the correlation coefficient between gene expressions, positioned at the intersection of genes on the x and y axes. The color intensity within each square shows the strength and direction of the relationship. However, Clusters of red suggest groups of positively correlated genes, possibly functioning together in biological pathways. Blue clusters may indicate a regulatory relationship or involvement in different biological processes. A closer examination of color intensities can uncover genes with stronger relationships, essential for understanding genetic interactions and their roles in biological functions. It is important to note outliers and anomalies, which could reveal unique gene interactions and warrant further investigation.

IV. Methodology

- **Dataset Preparation**: The methodology begins with the RNA-Seq (HiSeq) PANCAN data from the UCI Machine Learning Repository, focusing on the classification of various cancer types. The dataset contains a high-dimensional array of gene expression data from tumor samples, comprising 801 samples and 16,384 features. To address the issue of class imbalance in the target variable, downsampling techniques are employed to ensure an equal representation of classes, which is critical to prevent model overfitting.
- **Exploratory Data Analysis (EDA)**: The next phase involves EDA, where the distribution of a selection of genes across different cancer types is examined. This investigation aims to identify potential links between gene expression and specific cancer types, such as BRCA, and to gain insights into the molecular mechanisms of cancer.
- **Feature Selection and Dimensionality Reduction**: A tree-based method is used for feature selection to narrow down the most relevant genes for tumor type classification. The Random Forest algorithm identifies 70 key genes. To manage the high dimensionality of the data, PCA and Kernel PCA are applied to reduce the dataset to a two-dimensional space. This step is crucial for visualizing and identifying non-linear patterns that are not apparent in higher dimensions.
- **Network Analysis**: The methodology also includes the construction and analysis of a gene interaction network. The network's community structure is analyzed, and metrics such as modularity are used to determine the significance and non-randomness of the network division.
- **Model Evaluation**: Finally, machine learning models are evaluated using metrics like Precision, Recall, F1 Score, Accuracy, and Test Error. Models assessed include Random Forest, Decision Tree, XGBoost, SVM, Logistic Regression, KNN, and Naive Bayes. The model evaluation focuses on the performance of these algorithms in accurately classifying tumor types based on the selected features.

This comprehensive methodology integrates data preparation, exploratory analysis, feature selection, dimensionality reduction, network analysis, and rigorous model evaluation to ensure a robust approach to classifying cancer types using gene expression data. The outcomes of each methodological step inform the subsequent stages, leading to a detailed understanding of the dataset and the development of predictive models with the potential to contribute to personalized cancer treatment strategies.

V. Results and Discussion

The bar chart illustrates the class distribution after a downsampling process applied to a dataset containing different cancer types. Each bar represents one of the five cancer types: PRAD, LUAD, BRCA, KIRC, and COAD. All bars appear to reach the same height, indicating that each cancer type now has an equal number of samples, specifically 78, as a result of the balancing technique. The variety of colors for each bar enhances visual distinction, making it clear that the dataset has been evenly balanced across the different classes, which is typically done to prevent model bias towards more frequently occurring classes in machine learning tasks.

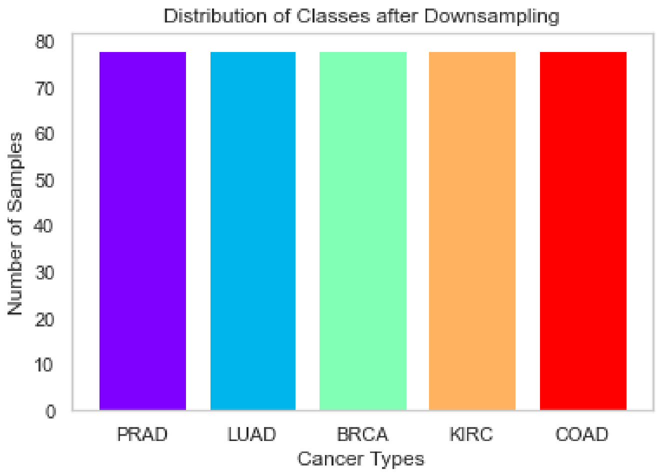


Fig. 4. Distribution of genes across Tumor types After Downsampling.

We choose the desired attributes using a tree-based method. By doing this, we hope to reduce the likely set of genes or characteristics that will be most useful in accurately determining which tumor types to focus on. For instance, the genes thought to be most significant were selected using a random forest. This particular strategy resulted in the selection of seventy genes as the most pertinent for our categorization task. Figure 5 shows the top twenty genes that were chosen, and it's shown below.

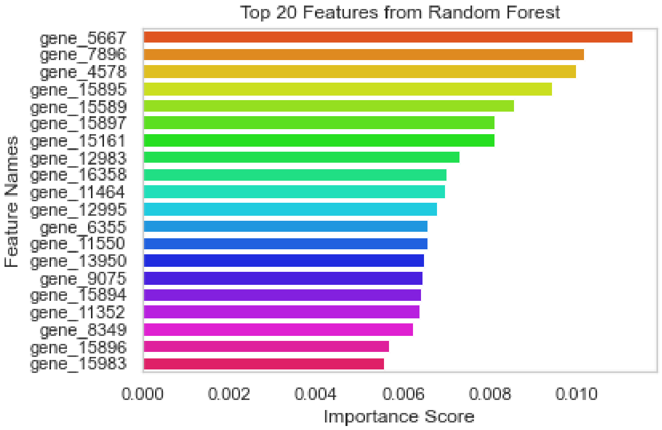


Fig. 5. Top 20 Feature Importance

The gene expression dataset's complex structure is further examined. We use Principal Component Analysis (PCA) and Kernel PCA to overcome the limits of high-dimensional data visualization. These methods help condense the complex dataset into a two-dimensional structure. This dimensional reduction aims to reveal nonlinear patterns in our gene expression data that might be hidden in higher dimensions.

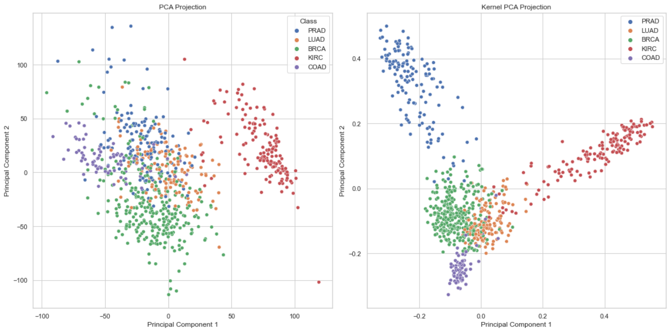


Fig. 6. Visualization of the Data Structure.

PCA is our first step. The original data is transformed into orthogonal principal components via PCA to conserve as much variation as feasible in these lower dimensions. However, this linear transformation has drawbacks. The two-dimensional PCA projection does not reveal gene expression data correlations in our scenario. PCA plots do not reveal whether gene interactions are linear, nonlinear, or complicated.

We use Kernel PCA to overcome this constraint. This sophisticated method uses kernel methods to find and depict nonlinear correlations that PCA may miss. Kernel PCA can reveal structures and patterns that linear approaches cannot by transferring our data into a higher-dimensional feature space and applying PCA there.

Kernel PCA projections give interesting insights. Kernel PCA shows non-linear patterns in our dataset, unlike PCA. This discovery matters. It suggests that gene expression interactions may include complicated, nonlinear dynamics.

Thus, this realization affects our analysis strategy. Due to the dataset's non-linear properties, linear analysis methods may be insufficient or misleading. Instead, we must use methods that capture and analyze these non-linear correlations to solve problems. To effectively study and understand our gene expression dataset's biological events, this change is necessary.

Metric	**Value**
Original Modularity	0.5451
Avg Randomized Modularity	0.1603
p-value	0.0000

TABLE I
SUMMARY OF NETWORK MODULARITY ANALYSIS

The graphical representation of the network from Fig 7, with nodes color-coded to denote distinct communities, visually confirms the quantitative findings. The nodes representing genes are sized proportionally to the number of connections, revealing the existence of hub genes within each community. These hubs are critical, as they may play pivotal roles in the regulation of cancer-related biological processes. The interconnectivity within communities and the presence of

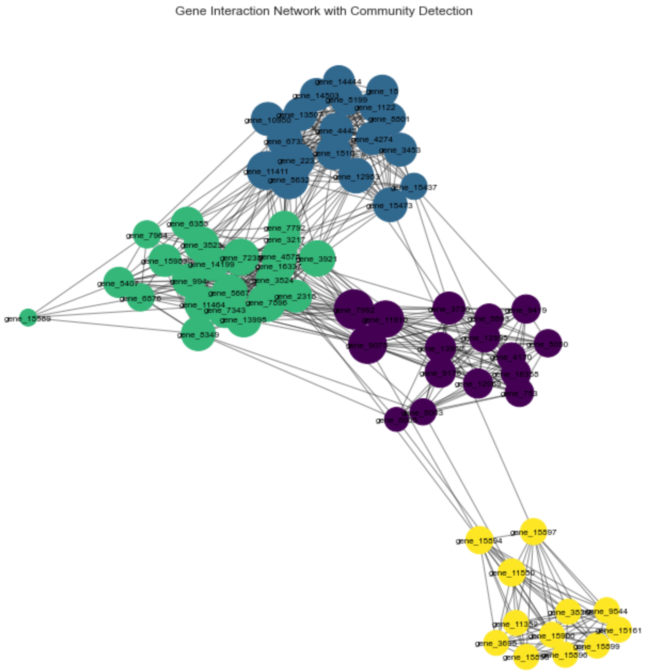


Fig. 7. Gene Interaction Network

Community 1	Community 2	Community 3	Community 4
gene_18	gene_753	gene_3695	gene_994
gene_223	gene_3730	gene_3836	gene_2318
gene_1122	gene_4170	gene_9544	gene_3217
gene_1510	gene_5050	gene_11352	gene_3523
gene_3453	gene_5693	gene_11550	gene_3524
gene_4274	gene_7992	gene_15161	gene_3921
gene_4442	gene_8003	gene_15894	gene_4578
gene_5199	gene_8005	gene_15895	gene_5407
gene_5632	gene_9075	gene_15896	gene_5667
gene_6733	gene_9176	gene_15897	gene_6355
gene_8801	gene_9419	gene_15899	gene_6876
gene_10950	gene_11910	gene_15900	gene_7238
gene_11411	gene_12069		gene_7343
gene_12983	gene_12995		gene_7792
gene_13507	gene_13950		gene_7896
gene_14444	gene_16358		gene_7964
gene_14503			gene_8349
gene_15437			gene_11464
gene_15473			gene_13998
			gene_14199
			gene_15589
			gene_15983
			gene_16337

TABLE II
DISTRIBUTION OF GENES IN DIFFERENT COMMUNITIES

such hub genes provide a starting point for hypothesizing about gene function and potential pathways involved in cancer pathogenesis and progression.

The exploration of gene interaction networks has yielded substantial insights into the community structure inherent within the data representing five distinct cancer types. The modularity analysis performed on the network revealed an original modularity score of 0.5451 from Table 1. This substantial score indicates a strong community structure and suggests that the network's divisions are indeed meaningful and not the result of random fluctuations in gene expression data.

Further reinforcing this conclusion is the comparison with the average modularity score obtained from a thousand randomized networks, which stands at a significantly lower 0.1603. The stark contrast between the original and randomized modularity scores highlights the presence of non-random structures within the network. This is quantitatively supported by a p-value of 0.0, suggesting with high confidence that the observed community structures are statistically significant and unlikely to have arisen by chance.

The visual and statistical analyses together tell a compelling story: the gene interaction network has important structures that are linked to the biological processes that cause cancer. These findings pave the way for targeted biological pathway analyses and experimental validation to elucidate the roles of these gene communities in cancer biology.

The table presents a distribution of genes across four distinct genetic communities, revealing patterns of gene clustering that could be indicative of shared functions or regulatory mechanisms. Community 1, with 19 listed genes, and Community 2, with 17, demonstrate a diverse range of genes, suggesting a possible broad spectrum of genetic functions or interactions within these communities. In contrast, Community 3, consisting of only 12 genes, and Community 4, with the most, 24 genes, might indicate more specialized or tightly regulated gene groups. The absence of certain genes in specific communities could hint at unique genetic characteristics or roles that differentiate these communities from one another. This distribution pattern can provide valuable insights for further genetic research, especially in understanding gene interactions, functional genomics, and potential implications in fields like disease research or biotechnology.

Figure 8 presents a rich visual representation of the data, where the color scale on the left denotes the intensity of the measured variable. Yellow signifies the highest values, suggesting robust gene expression, whereas blue corresponds to the lowest values, indicative of low or absent expression. Hierarchical clustering is demonstrated through dendrograms branching across the top and left margins, organizing columns and rows—potentially experimental conditions and individual genes—into clusters based on similarity in expression patterns. The proximity of branches within these dendrograms reflects the degree of resemblance, with shorter branches denoting higher similarity. Examining the heatmap reveals conspicuous color blocks, such as a striking yellow stripe signifying high gene expression in specific conditions, contrasting with expansive green areas that indicate moderate expression levels prevalent across numerous conditions and genes. This visualization unveils co-regulated genes or conditions eliciting comparable gene expression responses, with distinct color variations highlighting genes that are upregulated or down-

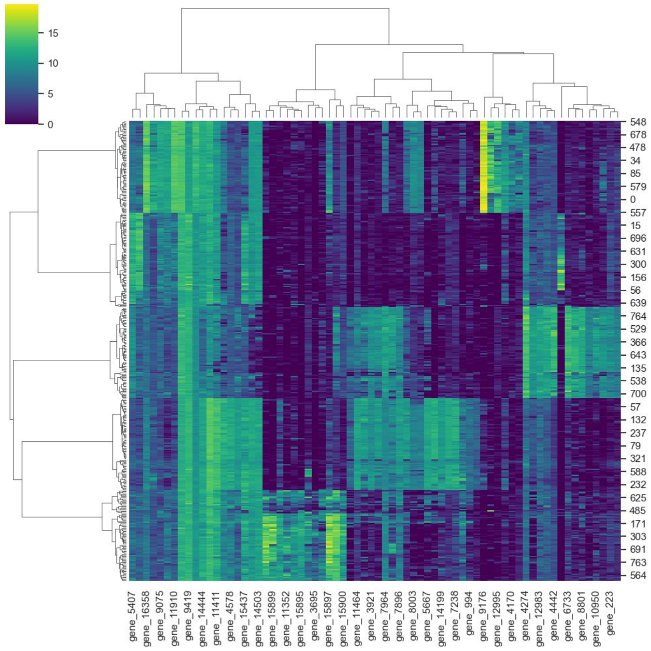
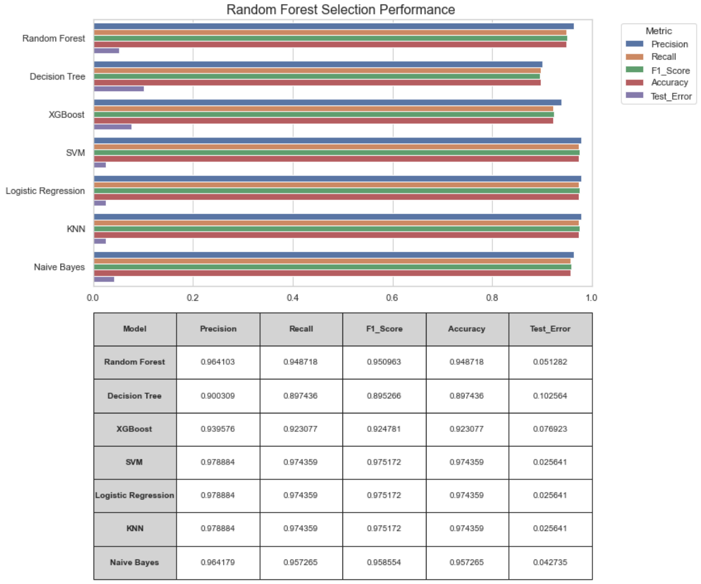


Fig. 8. Heatmap with row and column-wise clustering

regulated under certain circumstances. Furthermore, anomalies are spotlighted by intense yellow spots against a dark blue backdrop, pinpointing outlier genes or those of particular interest that exhibit high expression levels in an environment where other genes display minimal expression. This detailed array of data provides a foundation for deeper biological insights and further investigations into gene functionality and regulatory mechanisms.



Model	Precision	Recall	F1_Score	Accuracy	Test_Error
Random Forest	0.964103	0.948718	0.950963	0.948718	0.051282
Decision Tree	0.900309	0.897436	0.895266	0.897436	0.102564
XGBoost	0.939576	0.923077	0.924781	0.923077	0.076923
SVM	0.978884	0.974359	0.975172	0.974359	0.025641
Logistic Regression	0.978884	0.974359	0.975172	0.974359	0.025641
KNN	0.978884	0.974359	0.975172	0.974359	0.025641
Naive Bayes	0.964179	0.957265	0.958554	0.957265	0.042735

Fig. 9. Model Evaluation

The evaluation of machine learning models on the dataset reveals interesting insights into their performance. The Random Forest model exhibits a high degree of precision, recall, F1 score, and accuracy, all surpassing 94%, with a modest test error of approximately 5%. This indicates a balanced capacity for error avoidance and accurate prediction. In contrast, the Decision Tree, while robust, falls behind the Random Forest, as anticipated due to the latter being an ensemble technique known for outperforming single models; it shows roughly 90% across precision, recall, and accuracy, with a test error nearing 10%.

XGBoost, another ensemble approach, presents slightly reduced precision and recall when compared to Random Forest but surpasses the Decision Tree with its F1 score, albeit with a 7% test error. Meanwhile, the SVM model shines with exceptionally high precision, recall, and F1 scores—almost touching or exceeding 97%—matched with an accuracy of 97% and the lowest test error at about 2.6%, suggesting superior performance on the test data.

Logistic Regression mirrors SVM's impressive metrics, presenting equally low test error rates, which is noteworthy considering SVM's typically more complex nature. KNN matches SVM and Logistic Regression in precision, recall, and F1 score but slightly lags in accuracy, potentially due to its method of handling boundary cases. Naive Bayes stands its ground with good precision and recall, and an F1 score and accuracy over 95%, yet harbors a test error slightly higher than SVM and Logistic Regression, at around 4.2%.

In the broader discussion, the stellar performance of SVM, Logistic Regression, and KNN—demonstrated by high metric scores and low test errors—suggests a dataset with clear separability, which these models exploit effectively. The Random Forest model also delivers solid performance, potentially owing to its versatility with various data types. The higher test errors observed in Decision Tree and XGBoost could point to overfitting or a need for further parameter optimization. Interestingly, the simplicity of SVM and Logistic Regression seems to provide an edge over the more complex models like Random Forest and XGBoost, indicating the dataset might lack complex non-linearities that necessitate advanced modeling.

Test error emerges as a pivotal metric, projecting the expected performance of the models on unseen data; the lower the test error, the more reliable the predictions. Ultimately, while accuracy levels are high across all models, the final selection may hinge on factors like model complexity, computational efficiency, interpretability, and potential performance with expanded datasets or varied features.

VI. Conclusion

The study's result sums up all the hard work that went into figuring out gene expression data and how it affects how cancers are classified. A list of seventy genes, mostly the twenty most likely to be related to different types of tumors, has been made. A tree-based algorithm, especially Random Forest, was used to carefully choose which traits to include in this distillation. This targeted approach makes it easier to both analyze and find the genetic markers that are most linked to a certain type of cancer.

We were able to get around the problems with high-dimensional visualization by using principal component analysis (PCA) and kernel principal component analysis (KPCA) to find the gene expression data's hidden non-linear relationships. It was the Kernel PCA that showed the complicated nonlinear dynamics. It also made people question whether linear analysis methods would work with this dataset and pushed for a more advanced method that accepts and even welcomes its complexity.

The modularity study showed that the community structure is important in the gene interaction network, which has grown into a story-filled canvas. The results show that there is a strong link between the network's nodes and the basic biological processes that cause cancer, much stronger than what could be explained by chance alone.

There is a clear order to the way the machine learning models we tried did it. In particular, SVM, Logistic Regression, and KNN all did very well with very little test error. This may mean that the dataset naturally separates into different groups, which these models use to their advantage. On the other hand, more changes might be needed because Random Forest and XGBoost, even though they are strong, show signs of overfitting. Because of how well these models work, people are talking about how to make them more efficient on computers, how to make them work with different types of data, and the trade-offs between model complexity and interpretability.

References

[1] Alizadeh, A. A., Eisen, M. B., Davis, R. E., Ma, C., Lossos, I. S., Rosenwald, A., ... & Staudt, L. M. (2000). Distinct types of diffuse large B-cell lymphoma identified by gene expression profiling. Nature, 403(6769), 503-511.

[2] Golub, T. R., Slonim, D. K., Tamayo, P., Huard, C., Gaasenbeek, M., Mesirov, J. P., ... & Lander, E. S. (1999). Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. science, 286(5439), 531-537.

[3] Salem, H., Attiya, G., & El-Fishawy, N. (2017). Classification of human cancer diseases by gene expression profiles. Applied Soft Computing, 50, 124-134.

[4] Xu, R., Anagnostopoulos, G. C., & Wunsch, D. C. (2007). Multiclass cancer classification using semisupervised ellipsoid ARTMAP and particle swarm optimization with gene expression data. IEEE/ACM Transactions on Computational Biology and Bioinformatics, 4(1), 65-77.

[5] Hughes, G. (1968). On the mean accuracy of statistical pattern recognizers. IEEE transactions on information theory, 14(1), 55-63.

[6] Libbrecht, M. W., & Noble, W. S. (2015). Machine learning applications in genetics and genomics. Nature Reviews Genetics, 16(6), 321-332.

[7] Antoniadis, A., Lambert-Lacroix, S., & Leblanc, F. (2003). Effective dimension reduction methods for tumor classification using gene expression data. Bioinformatics, 19(5), 563-570.

[8] Zhao, G., & Wu, Y. (2016). Feature subset selection for cancer classification using weight local modularity. Scientific Reports, 6(1), 34759.

[9] J.C. Ang, A. Mirzal, H. Haron, H.N. Abdull Hamed, Supervised unsupervised and semi-supervised feature selection: a review on gene selection, IEEE/ACM Trans. Comput. Biol. Bioinforma. 13 (5) (2016) 971–989.

[10] D. Pavithra, B. Lakshmanan, Feature selection and classification in gene expression cancer data, 2017 International Conference on Computational Intelligence in Data Science, IEEE, 2017, pp. 1–6.

[11] I. Guyon, J. Weston, S. Barnhill, V. Vapnik, Gene selection for cancer classification using support vector machines, Mach. Learn. 46 (1–3) (2002) 389–422.

[12] Y. Piao, K.H. Ryu, Detection of differentially expressed genes using feature selection approach from RNA-seq, 2017 IEEE International Conference on Big Data and Smart Computing, IEEE, 2017, pp. 304–308.

[13] C. Ding, H. Peng, Minimum redundancy feature selection from microarray gene expression data, J. Bioinforma. Comput. Biol. 3 (02) (2005) 185–205.

[14] C.-H. Zheng, D.-S. Huang, L. Shang, Feature selection in independent component subspace for microarray data classification, Neurocomputing 69 (16–18) (2006) 2407–2410.

[15] X. Zhu, N. Li, Y. Pan, Optimization performance comparison of three different group intelligence algorithms on a SVM for hyperspectral imagery classification, Remote Sens. 11 (6) (2019) 734.

[16] P. Maji, C. Das, Relevant and significant supervised gene clusters for microarray cancer classification, IEEE Trans.Nanobiosci. 11 (2) (2012) 161–168.

[17] J. Brimberg, N. Mladenović, R. Todosijević, D. Urošević, Solving the capacitated clustering problem with variable neighborhood search, Ann. Oper. Res. 272 (1–2) (2019) 289–321.

[18] G. Palubeckis, A. Ostreika, D. Rubliauskas, Maximally diverse grouping: an iterated Tabu search approach, JORS 66 (4) (2015) 579–592.

[19] K. Singh, S. Sundar, A new hybrid genetic algorithm for the maximally diverse grouping problem, Int. J. Mach. Learn. Cybern. (2019) 1–20.

[20] E. Bonilla-Huerta, A. Hernandez-Montiel, R. Morales-Caporal, M. Arjona-López, Hybrid framework using multiple-filters and an embedded approach for an efficient selection and classification of microarray data, IEEE/ACM Trans. Comput. Biol. Bioinforma. 13 (1) (2016) 12–26.

[21] Y. Zhang, Q. Deng, W. Liang, X. Zou, An efficient feature selection strategy based on multiple support vector machine technology with gene expression data, Hindawi. BioMed Res. Int. (2018), https://doi.org/10.1155/2018/7538204.

[22] I. Salman, O.N. Ucan, O. Bayat, K. Shaker, Impact of metaheuristic iteration on artificial neural network structure in medical data, Processes 6 (5) (2018) 57, https://doi.org/10.3390/pr6050057.

[23] A. Feitosa Neto, A.M. Canuto, J.C. Xavier-Junior, Hybrid metaheuristics to the automatic selection of features and members of classifier ensembles, Information 9 (268) (2018) 1–25, https://doi.org/10.3390/info9110268.

[24] García-Díaz, P., Sanchez-Berriel, I., Martínez-Rojas, J. A., & Diez-Pascual, A. M. (2020). Unsupervised feature selection algorithm for multiclass cancer classification of gene expression RNA-Seq data. Genomics, 112(2), 1916-1925.

[25] Agbemade, Emil, ”Predicting Heart Disease using Tree-based Model” (2023). Data Science and Data Mining. 1. https://stars.library.ucf.edu/data-science-mining/1

[26] Alipour Yengejeh, Amir, ”Genome-Wide Association Study of The Maize Crop by The Lasso Regression Analysis” (2023). Data Science and Data Mining. 6. https://stars.library.ucf.edu/data-science-mining/6

[27] Seyedmonir, S., Bayrami, M., Jafarzadeh Ghoushchi, S., Alipour Yengejeh, A., Morabbi Heravi, H. (2021). Extended fully fuzzy linear regression to analyze a solid cantilever beam moment. Mathematical Problems in Engineering, 2021, 1-9.

[28] Wang, S., Zhu, X., Ding, W., Yengejeh, A. A. (2022). Cyberbullying and cyberviolence detection: A triangular user-activity-content view. IEEE/CAA Journal of Automatica Sinica, 9(8), 1384-1405.

[29] Alipour Yengejeh, Amir, ”A Linear Regression Model to Predict the Critical Temperature of a Superconductor” (2023). Data Science and Data Mining. 12. https://stars.library.ucf.edu/data-science-mining/12

APPENDIX

CODE

The code for this study is hosted on GitHub. Click here for your reference.