A comparative analysis of different gene selection methods on microarray expression data for cancer classification

Frank Britto-Bisso, *Student Member, IEEE,* and Lucero Gomez

***Abstract*—Microarray gene expression data is high-dimensional and low-sample by nature. Then, since it´s interpretation its highly dependant on the processing step in the data analysis pipeline, robust gene selection techniques are needed. Despite their high performance, myopic feature selection techniques fail to capture the real interactions between genes, resulting in very poor biological interpretability of machine learning models trained with these features. In this work, we explore another feature selection category named "non-myopic" that captures, either indirectly or directly, gene-gene interactions. We compared the performance of Logistic Regression, KNN, SVM, Random Forest and XGBoost classifiers trained with genes selected via the myopic feature selection techniques Mimimun redundancy - Maximun Relevance and Recursive feature elimination, with the same models trained with non-myopic feature selection techniques: Relief-based method called SURF*, as a representative example of indirect gene interactions; and a convolutional neural network based on the resnet pre-trained network, trained with an image representation of the microarray gene expression data.**

***Index Terms*—Microarray, Feature selection, Machine learning, CNN, High dimensionality**

I. Introduction

MICROARRAY, also known as a gene array or DNA chip, is a technological platform that contains thousands of immobilized DNA or RNA probes on a solid surface such as a glass slide or silicon chip. These probes are designed to detect and quantify the abundance of specific gene sequences or nucleic acids in a biological sample. For cancer research, microarrays are used to analyze the gene expression of tumor cells and compare it with that of normal cells. This allows the identification of genes that are overexpressed, providing crucial information about the molecular mechanisms involved in the disease. The microarray technology is known for its capacity to identify distinctive gene expression patterns among different types of cancer aiding in tumor classification and predicting patient prognosis [1]. Another important approach is the use of genotyping microarrays which detect genetic variants associated with the risk of developing cancer. These microarrays analyze the presence or absence of single nucleotide polymorphisms (SNPs) in an individual's genome revealing information about genetic predisposition to cancer.

The analysis of microarray-derived data exhibits significant challenges: the high dimensionality and low-sample size (see Supplementary Material 1 for an example), represented by the large number of analyzed features or genes, generates difficulties in data processing, result interpretability and further selection of relevant features. This can lead to the overfitting of machine learning classifiers, limmiting its generalization to new samples. Additionally, the low number of samples hinders the identification of subtle but statistically significant expression differences that might represent cancer subtypes or sample heterogeneity. Addressing these challenges is crucial for the development of personalized therapies, as well as biomarker and new pharmacological target´s discovery [2].

Regarding the issues of high dimensionality and low sample size, there are feature selection methods that can be divided into myopic and non-myopic methods. Myopic feature selection methods make decisions based on local criteria evaluating features individually without considering their interaction. These methods tend to be computationally efficient and relatively easy to implement. Non-myopic methods consider the interaction between features and make decisions globally. These methods may have higher computational costs as they evaluate multiple feature combinations. However their global approach can provide better feature selection.

Most literature is focused on myopic feature selection technique, particularly Principal Component Analysis (PCA) [3], [4], Miminum Redundancy - Maximun Relevance (MRMR) [5], [6] , and Recursive Feature Selection [7]–[9]. However, recent evidence [3], [10] suggests that this methods, despite their high performance for classification tasks, fail to capture real interactions between features, which translates to the functional and physical interactions between genes and proteins. In this context, in this work, we will explore the implementation of feature selection algorithms able to address, directly or indirectly, the relationship between genes, and analyze their impact in the performance metrics of a supervised learning binary classification task between tumor and non-tumor samples.

II. Materials and Methods

Each subsection of our study can be found in our Github repository. Data processing was done primarily in Rstudio (R v.4.3.1), while the feature selection methods and machine learning models were implemented in the Jupyter Notebook interface (Python v. 3.10).

A. Data collection

Breast cancer microarray expresion datasets were screened from the NIH - Gene Expression Omnibus database, considering the following exclusion criteria [11]: (i) only studies in Homo Sapiens, (ii) studies that employed pharmacological

interventions or treatments with small molecules, (iii) studies that include samples with induced mutations (e.g knockdown experiments) and xenograft techniques, (iv) studies that were correctly labeled and exhibit a reproducible protocol description, and (v) studies were the raw data was available (not only the BioProject processed by the author). The Series GSE70947 was finally chosen for analysis based on class balance (see Supplementary Material 1).

B. Data preprocessing

Data preprocessing was done following the pipeline proposed for the CuMiDa dataset construction [12] using R. Raw files were downloaded from the GEO webpage. Since data retrieval was done using an Agilent microarray platform [13], we used the *limma* package (v. 3.56.2) for background correction and quantile normalization. The *arrayQualityMetrics* package (v. 3.56) was then used to address data quality, from which samples that were tagged as "outliers" in ≥ 2 of the criterions were excluded. Ultimately, the *limma* and *Biobase* (v. 2.6) packages were used for differential gene expression analysis. Low expression genes were filtered based on the median expression value. Differentially expressed genes (DGEs) were finally obtained by applying a filter of $|\log_2 \text{FC}| \geq 1$ (FC for fold-change) with the Benjamini-Hochberg for the false discovery rate (FDR) correction of $p < 0.05$ [11]. Ultimately, gene enrichment analysis (that is, to correlate the DGEs with molecular and cellular biological functions) was done using the Gene Ontology and DisGENet resources, through the *clusterProfiler* package (v. 4.8.1).

C. Myopic feature selection methods

The Principal Component Analysis (PCA) was done in Python using the function *sklearn.decomposition.PCA* of the *scikit-learn* package (v. 1.2.2). Likewise, the recursive feature elimination (RFE) was done using the function *sklearn.feature_selection.RFECV* of the same package. The cross-validation was done implementing the a Logistic Regression model, as the documentation recommends. Ultimately, the minimun redundancy-maximun relevance (MRMR) was implemented using the *Pymrmre* package [14].

D. Non-myopic feature selection methods

The Relief-based feature selection method called SURF* was implemented in Python using the *scikit-rebate* package [15]. The DeepInsight framework was implemented based on the Supplementary Material of the original article [16], while the convolutional neural network (CNN) was loaded from the open-source Pytorch Image Models library.

E. Machine learning models

We adopted five machine learning models: Logistic Regression, Support Vector Machine (SVM), Random Forest, K-neirest neighbors (KNN) and XGBoost; all of them implemented with their respective package of the *scikit-learn* library. Their performance for the classification task (differentiate healthy from breast cancer tissue samples) was evaluated under the accuracy, precision, recall, and F1-score.

III. Results

A. Data description and pre processing

The dataset consisted in 296 samples with an equal distribution of classes (see Supplementary Material 1), and a total of 62976 probes per sample. After filtering, normalization and outlier detection, a total of 290 samples were taken for further analysis, from which we identified a total of 2051 DEGs. Compared to the healthy controls, the breast tissue cancer samples exhibit 997 up-regulated genes and 1054 downregulated genes (Fig. 1). To visualize the relationship between the DEGs, the protein-protein interaction network was constructed using the STRING platform (see Supplementary Material 2).

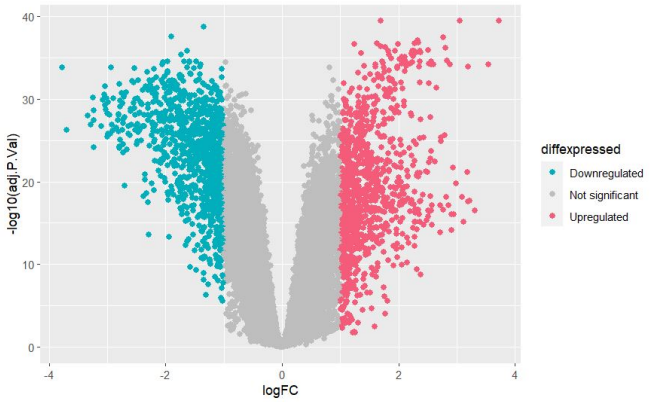

Fig. 1. Volcano plot of DEGs estimated from GSE70947 (cohort 2) dataset

Following the protocol detailed in [17], a gene enrichment analysis was conducted to explore the correlation between phenotype (that is, the cancerous or healthy state of each sample) and genotype. Fig. 2A shows that the DGEs represent key hallmarks of breast adenocarcinoma physiopathology, but are also representative to other diseases (see Fig. 2B), remarking the need for further gene selection techniques.

B. Myopic feature extraction

To set a standard baseline, we explored the performance of the aforementioned classifiers trained with the PCA components that explained 95% of the data variance, since this method was the most common approach for dimensionality reduction. The results are shown in Table I.

Model	Acc.	Pre.	Rec.	F1
Logistic Regression	0.844	0.865	0.849	0.857
KNN	0.854	0.831	0.925	0.875
SVM	0.833	0.863	0.830	0.868
Random Forest	0.854	0.868	0.868	0.868
XGBoost	0.865	0.857	0.906	0.881

TABLE I
PERFORMANCE METRICS OF DIFFERENT CLASSIFIERS TRAINED WITH PCA COMPONENTS

Furthermore, we implemented the recursive feature elimination method with cross-validation as a representative example of the wrapper class of feature selection algorithms, since it´s the most adopted strategy for microarray analysis [18]. The

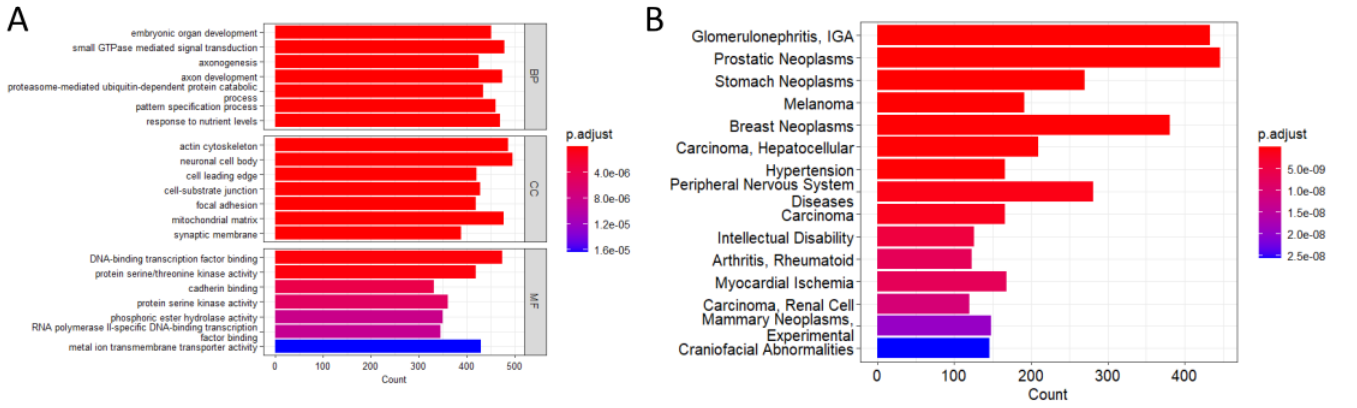

Fig. 2. Gene enrichment analysis results using the Gene Ontology (A) and DisGENet (B) resources

implementation is straightforward: on each iteration, once the Random Forest model performance is calculated, a "feature importance" metric is used to eliminate features. In our study, at the expense of time complexity, we defined a step size of 1 for each iteration, so that the least important feature is eliminated on each iteration. Table II shows the performance metric of the 5 classifiers trained with the 57 gene signature defined by this method (see Supplementary Material 3). The highest performance metrics ($\approx$ 98 %) were obtained with this methodology and a SVM classifier.

Finally, we implemented the minimun redundancy maximun relevance (MRMR) as a representative example of the filter class of feature selection algorithms, since it was specifically designed for working with microarray data [19]. The MRMR selection criteria ranks as "most important" those genes that are highly correlated with their class (defined by a metric called relevance) and poorly correlated between other genes (defined by a metric called redundancy). In our study, both relevance and redundancy are calculated based on mutual information [20]. To define the optimal number of features and the type of mutual information operator, we performed a 5-fold cross validation test with a Logistic Regression classifier. We tried 2, 3, 5, 10, 15 and 20 features, from which the 10 gene signature was selected as the best (see Supplementary Material 3). Table II shows its performance metrics, which are both the lowest among the included gene selection methods in our study, and lower than the PCA baseline results.

C. Non-myopic feature extraction

Empirical evidence suggests that the RELIEF-based feature selection methods (RBAs) are able to indirectly capture feature dependencies (in our context, gene-gene interactions) by assigning weights to features based on nearest-neighbors instances [15], [21]. At their core, for each iteration, RBAs select a random gene and search for the nearest neighbor of the same class and from another class, named nearest hit and miss, respectively. Then, an "importance" index is calculated, with which genes that are adjacent and belong to the same class receive a positive weight, while the negative weight is assign to genes that are close to each other and belong to the other class [22] Therefore, the genes with the highest ranking are the ones that represent the best class separation. Since there is no mechanistic way of choosing the best RBA for our dataset, we did a 5-fold cross-validation test using a Random Forest classifier [23]. Since every classifier exhibited similar performance metrics (see Supplementary Material 4), we choose the SURF* method because it used less features (209), which could be further be analyzed to address their potential as biomarkers. According to Table II, 4 out of 5 classifiers have a performance similar to the classifiers trained with the PCA components, except for the XGBoost model, which have a similar performance as the same model trained with the RFECV gene signature.

On the other hand, to directly capture the non-linear relationships between genes, we applied the DeepInsight framework [16] to transform our high-dimensional, tabular data into an image to implement a Convolutional Neural Network (CNN), which have already showed promising results for high-dimensionality biological data classification [24]. This technique executes a dimensionality reduction operation using Uniform Manifold Approximation and Projection (UMAP) technique, from $>$ 15 000 dimensiones to a 2D-plane visualization, in which genes with high correlation between each have close cartesian coordinates. Then, the smallest square that captures the UMAP plot is found, and the image is cropped, rotated and re-scaled for easier manipulation. Ultimately, the algorithm maps every gene expression value with a unique pixel.

We selected this image transformer over other state-of-the-art methodologies (e.g REFINED [25] or IGTD [26]) due to its compatibility with a reverse operation. Since the DeepInsight transformer maps every gene to a single pixel (see Supplementary Material 4), we can extract the section of each image that the CNN considered as representative for each class, and then reverse-engineer the image section into the tensor of genes from which it was build. The reverse mapping of each pixel to its corresponding gene was done using the GradCAM technique [27]. A visualization of the reverse operation is showed in Fig. 3.

For the CNN implementation, since we had a very low number of samples, we aimed to use a transfer learning pipeline. We loaded the *resnet26d* architecture [28] and train

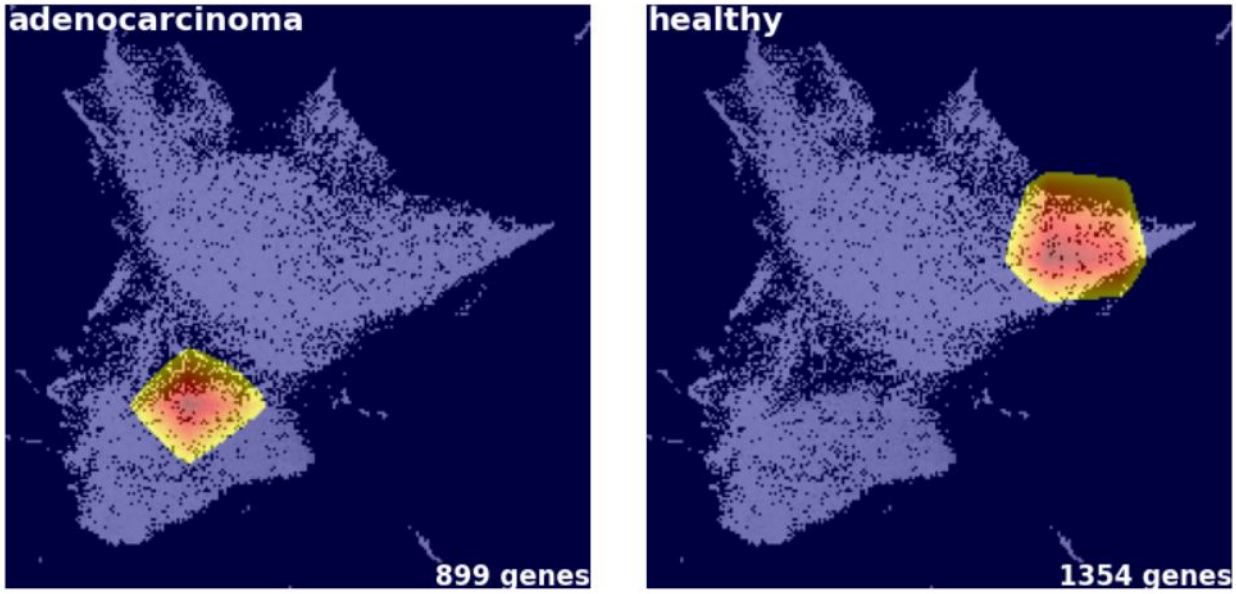

Fig. 3. GradCAM-based reverse mapping: from pixel to gene

it directly using the DeepInsight images. By trial and error, we defined an optimal pixel size of (224,224) for the images, and a mean flatten operation and a manual threshold of 0.6 were defined for the GradCAM-based reverse operation.

IV. Discussion

Microarray data (in general, every biological high-throughput technique) is naturally high-dimensional and low-sample size, as well as heterogeneous and sparse [29], [30]. Hence, its interpretation is highly dependant on the processing step in the data analysis pipeline, as one might be dealing with the curse of dimensionality. Hence the importance of dimensionality reduction techniques, which include unsupervised methods, feature extraction methods, in which a new space of features is generated based on raw data (very common for temporal data and image processing [31]); and feature selection methods, in which "important features" are selected [10]. In this work, we attempt to find the most optimal feature selection technique that defines the concept of "important feature" by addressing dependencies between features, which are not included in the most commonly used feature selection techniques [18] (hence the name "myopic").

The most common approach for microarray feature selection and further binary or multi classification task is the PCA operation [10]. In this context, we implemented the 5 classifiers trained with the PCA components that explain the 95% of out dataset´s variance after differential gene expression analysis to set a reference for comparison. Our results are consistent with previous literature [2], [4], [10]. Furthermore, we implemented two representative examples of the wrapper and filter methods [18]: RFECV and MRMR, respectively. The former exhibited the highest performance metrics with the SVM classifier. However, it´s important to consider that this technique yields a subset of features based on a subset ranking criterion [32]. Hence, the combination of features might result in an optimal performance (the best, in our case), but each individual feature could not necessarily be the most important. Therefore, considering its "myopic" nature, despite the high performance, the biological importance of the selected genes might theoretically not be high. Of course, gene enrichment analysis to validate this hypothesis is needed. Also, working with multiple classifiers for the cross-validation step is recommended to address reproducibility, for which a consistency index is needed [33]. Regarding the latter, the nature of the MRMR operation itself suggests low biological relevance. The "redundancy" metric discards features so that the final subset contains the ones with the lowest correlation between then, which, in theory, directly ignores gene-gene interactions; hence the very poor scores obtained in our study.

At a molecular level, genes and proteins, which are represented by each feature in the microarray dataset, exhibit a wide range of physical and functional interactions [34]. For the GSE70947 dataset, those interactions are summarized in the protein-protein network showed in Supplementary Material 2. Therefore, feature selection methods should operate in a way that captures these interactions (they shoyld be "non-myopic"). The RBAs were implemented in this study as a representative example of an indirect detection of these gene-gene interactions, based on a nearest-neighbors based operation. Recently developed RBAs, such as the MultiSURF and TuRF, are able to detect high-order interactions (up to 5 inter-gene dependencies [35]), However, its performance gets worst for a larger number of "irrelevant" or redundant features [15]. Therefore, we executed the SURF* technique only for the DGEs to reduce both the time complexity and the number of irrelevant features (i.e those with $\log_2 \rightarrow 0$). Following the ranking, the overall performance of the SURF* method can be categorized as the second best feature selection method of the study. Further optimization strategies such as hybrid techniques [36], [37] and new feature weighting strategies

Classifier	Accuracy	Precision	Recall	F1-score
Logistic Regression - RFECV	0.947917	0.957447	0.937500	0.947368
SVM - RFECV	0.979167	0.979167	0.979167	0.979167
KNN - RFECV	0.854167	0.803571	0.937500	0.865385
Random Forest - RFECV	0.875000	0.891304	0.854167	0.872340
XGBoost - RFECV	0.906250	0.882353	0.937500	0.909091
Logistic Regression - MRMR	0.822917	0.846154	0.830189	0.838095
SVM - MRMR	0.833333	0.862745	0.830189	0.846154
KNN - MRMR	0.802083	0.826923	0.811321	0.819048
Random Forest - MRMR	0.843750	0.880000	0.830189	0.854369
XGBoost - MRMR	0.864583	0.870370	0.886792	0.878505
Logistic Regression - SURF*	0.875000	0.867925	0.901961	0.884615
SVM - SURF*	0.864583	0.900000	0.849057	0.873786
KNN - SURF*	0.833333	0.793651	0.943396	0.862069
Random Forest - SURF*	0.885417	0.888889	0.905660	0.897196
XGBoost - SURF*	0.906250	0.882353	0.937500	0.909091
CNN - DeepInsight	0.828	0.966	0.757	0.848

TABLE II
COMPARATIVE ANALYSIS OF MYOPIC AND NON-MYOPIC GENE SELECTION METHODS

for the neirest-neighbors instances [38] could be explored. Nevertheless, it´s important to recognize that RBAs-based feature selection techniques do not offer an explanation about the nature of the dependencies between features [15]. In other words, just with the output features, one can´t know if those features reflect a real interaction or are just a linear effect.

The last implemented feature selection algorithm aimed to detect directly the gene-gene interactions based on the transformation of tabular data into images, mapping each pixel with the expression intensity of each gene. Deep learning algorithms have been extensively applied for cancer classification based on microarray data [39], with remarkable results specially for graph-based neural networks [40]–[42], which are able to directly capture topological features of gene networks (e.g see Supplementary Material 2). However, those networks need a large number of samples for training. Synthetic data construction [43] is not commonly employed since it needs further validation with independent datasets (which are also low-sample); so the most common strategy is merging different datasets and employing batch normalization techniques to reduce variability [44]. We decided for a more complex deep learning architecture with a low-sample size rather than a simple architecture with a larger sample size due to inconsistencies in those inter-dataset normalization techniques [45]. Following the ranking, the overall performance of the DeepInsight method can be categorized as the third best feature selection method of the study. Interestingly, some transfer learning techniques from Resnet-based CNNs for radar image classification [46] could be extrapolated for our application, since they deal with the very low sample problem. Similar pre-trained models could be used, too, as the one originally proposed by the authors of the DeepInsight transformation [47].

In this study, we systematically compared different gene selection methods and found that their application is case-specific. If the goal is to maximize predictive power, wrapper, non-myopic feature selection such as the recursive feature elimination technique, are recommended, at the expense of biological interpretation. On the other hand, if the goal is biomarker discovery or basic cancer research, non-myopic feature selection algorithms should be used, at the expense of the performance metric, which fortunately can be enhanced with further optimization steps. Finally, some limitations in our study should be addressed. We didn´t used an external validation set since we didn´t find another dataset that satisfies our inclusion criteria, used the same type of samples (biopsies) and conducted the microarray screening using an Agilent platform. More datasets such as the EMBL should be explored. Similarly, more resources such as the KEEG should be addressed for a more complete gene enrichment analysis.

REFERENCES

[1] S. B. Bhonde and J. R. Prasad, "Deep learning techniques in cancer prediction using genomic profiles," in *2021 6th International Conference for Convergence in Technology (I2CT)*. IEEE, Apr. 2021. [Online]. Available: https://doi.org/10.1109/i2ct51068.2021.9417985

[2] R. Shen, A. B. Olshen, and M. Ladanyi, "Integrative clustering of multiple genomic data types using a joint latent variable model with application to breast and lung cancer subtype analysis," *Bioinformatics*, vol. 25, no. 22, pp. 2906–2912, Sep. 2009. [Online]. Available: https://doi.org/10.1093/bioinformatics/btp543

[3] F. Alharbi and A. Vakanski, "Machine learning methods for cancer classification using gene expression data: A review," *Bioengineering*, vol. 10, no. 2, p. 173, Jan. 2023. [Online]. Available: https://doi.org/10.3390/bioengineering10020173

[4] Y. Cui, C.-H. Zheng, and J. Yang, "Dimensionality reduction for microarray data using local mean based discriminant analysis," *Biotechnology Letters*, vol. 35, no. 3, pp. 331–336, Nov. 2012. [Online]. Available: https://doi.org/10.1007/s10529-012-1092-3

[5] H. Alshamlan, G. Badr, and Y. Alohali, "mRMR-ABC: A hybrid gene selection algorithm for cancer classification using microarray gene expression profiling," *BioMed Research International*, vol. 2015, pp. 1–15, 2015. [Online]. Available: https://doi.org/10.1155/2015/604910

[6] G. Zhang, J. Hou, J. Wang, C. Yan, and J. Luo, "Feature selection for microarray data classification using hybrid information gain and a modified binary krill herd algorithm," *Interdisciplinary Sciences: Computational Life Sciences*, vol. 12, no. 3, pp. 288–301, May 2020. [Online]. Available: https://doi.org/10.1007/s12539-020-00372-w

[7] I. Guyon, J. Weston, S. Barnhill, and V. Vapnik, *Machine Learning*, vol. 46, no. 1/3, pp. 389–422, 2002. [Online]. Available: https://doi.org/10.1023/a:1012487302797

[8] H. Ravishankar, R. Madhavan, R. Mullick, T. Shetty, L. Marinelli, and S. E. Joel, "Recursive feature elimination for biomarker discovery in resting-state functional connectivity," in *2016 38th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*. IEEE, Aug. 2016. [Online]. Available: https://doi.org/10.1109/embc.2016.7591621

[9] X. Huang, L. Zhang, B. Wang, F. Li, and Z. Zhang, "Feature clustering based support vector machine recursive feature elimination for gene selection," *Applied Intelligence*, vol. 48, no. 3, pp. 594–607, Jul. 2017. [Online]. Available: https://doi.org/10.1007/s10489-017-0992-2

[10] S. Osama, H. Shaban, and A. A. Ali, "Gene reduction and machine learning algorithms for cancer classification based on microarray gene expression data: A comprehensive review," *Expert Systems with Applications*, vol. 213, p. 118946, Mar. 2023. [Online]. Available: https://doi.org/10.1016/j.eswa.2022.118946

[11] B. C. Feltes, J. de Faria Poloni, I. J. G. Nunes, S. S. Faria, and M. Dorn, "Multi-approach bioinformatics analysis of curated omics data provides a gene expression panorama for multiple cancer types," *Frontiers in Genetics*, vol. 11, Nov. 2020. [Online]. Available: https://doi.org/10.3389/fgene.2020.586602

[12] B. C. Feltes, E. B. Chandelier, B. I. Grisci, and M. Dorn, "CuMiDa: An extensively curated microarray database for benchmarking and testing of machine learning approaches in cancer research," *Journal of Computational Biology*, vol. 26, no. 4, pp. 376–386, Apr. 2019. [Online]. Available: https://doi.org/10.1089/cmb.2018.0238

[13] D. A. Quigley, A. Tahiri, T. Lüders, M. H. Riis, A. Balmain, A.-L. Børresen-Dale, I. Bukholm, and V. Kristensen, "Age, estrogen, and immune response in breast adenocarcinoma and adjacent normal tissue," *OncoImmunology*, vol. 6, no. 11, p. e1356142, Aug. 2017. [Online]. Available: https://doi.org/10.1080/2162402x.2017.1356142

[14] N. D. Jay, S. Papillon-Cavanagh, C. Olsen, N. El-Hachem, G. Bontempi, and B. Haibe-Kains, "mRMRe: an r package for parallelized mRMR ensemble feature selection," *Bioinformatics*, vol. 29, no. 18, pp. 2365–2368, Jul. 2013. [Online]. Available: https://doi.org/10.1093/bioinformatics/btt383

[15] R. J. Urbanowicz, R. S. Olson, P. Schmitt, M. Meeker, and J. H. Moore, "Benchmarking relief-based feature selection methods for bioinformatics data mining," *Journal of Biomedical Informatics*, vol. 85, pp. 168–188, Sep. 2018. [Online]. Available: https://doi.org/10.1016/j.jbi.2018.07.015

[16] A. Sharma, E. Vans, D. Shigemizu, K. A. Boroevich, and T. Tsunoda, "DeepInsight: A methodology to transform a non-image data to an image for convolution neural network architecture," *Scientific Reports*, vol. 9, no. 1, Aug. 2019. [Online]. Available: https://doi.org/10.1038/s41598-019-47765-6

[17] L. Zhang, R. Mao, C. T. Lau, W. C. Chung, J. C. P. Chan, F. Liang, C. Zhao, X. Zhang, and Z. Bian, "Identification of useful genes from multiple microarrays for ulcerative colitis diagnosis based on machine learning methods," *Scientific Reports*, vol. 12, no. 1, Jun. 2022. [Online]. Available: https://doi.org/10.1038/s41598-022-14048-6

[18] E. Alhenawi, R. Al-Sayyed, A. Hudaib, and S. Mirjalili, "Feature selection methods on gene expression microarray data for cancer classification: A systematic review," *Computers in Biology and Medicine*, vol. 140, p. 105051, Jan. 2022. [Online]. Available: https://doi.org/10.1016/j.compbiomed.2021.105051

[19] C. DING and H. PENG, "MINIMUM REDUNDANCY FEATURE SELECTION FROM MICROARRAY GENE EXPRESSION DATA," *Journal of Bioinformatics and Computational Biology*, vol. 03, no. 02, pp. 185–205, Apr. 2005. [Online]. Available: https://doi.org/10.1142/s0219720005001004

[20] M. Mandal and A. Mukhopadhyay, "An improved minimum redundancy maximum relevance approach for feature selection in gene expression data," *Procedia Technology*, vol. 10, pp. 20–27, 2013. [Online]. Available: https://doi.org/10.1016/j.protcy.2013.12.332

[21] A. Stief, J. R. Ottewill, and J. Baranowski, "Relief f-based feature ranking and feature selection for monitoring induction motors," in *2018 23rd International Conference on Methods &amp Models in Automation &amp Robotics (MMAR)*. IEEE, Aug. 2018. [Online]. Available: https://doi.org/10.1109/mmar.2018.8486097

[22] R. J. Urbanowicz, M. Meeker, W. L. Cava, R. S. Olson, and J. H. Moore, "Relief-based feature selection: Introduction and review," *Journal of Biomedical Informatics*, vol. 85, pp. 189–203, Sep. 2018. [Online]. Available: https://doi.org/10.1016/j.jbi.2018.07.014

[23] H. Aydadenta and Adiwijaya, "On the classification techniques in data mining for microarray data classification," *Journal of Physics: Conference Series*, vol. 971, p. 012004, Mar. 2018. [Online]. Available: https://doi.org/10.1088/1742-6596/971/1/012004

[24] J. Jumper, R. Evans, A. Pritzel, T. Green, M. Figurnov, O. Ronneberger, K. Tunyasuvunakool, R. Bates, A. Žídek, A. Potapenko, A. Bridgland, C. Meyer, S. A. A. Kohl, A. J. Ballard, A. Cowie, B. Romera-Paredes, S. Nikolov, R. Jain, J. Adler, T. Back, S. Petersen, D. Reiman, E. Clancy, M. Zielinski, M. Steinegger, M. Pacholska, T. Berghammer, S. Bodenstein, D. Silver, O. Vinyals, A. W. Senior, K. Kavukcuoglu, P. Kohli, and D. Hassabis, "Highly accurate protein structure prediction with AlphaFold," *Nature*, vol. 596, no. 7873, pp. 583–589, Jul. 2021. [Online]. Available: https://doi.org/10.1038/s41586-021-03819-2

[25] O. Bazgir, R. Zhang, S. R. Dhruba, R. Rahman, S. Ghosh, and R. Pal, "Representation of features as images with neighborhood dependencies for compatibility with convolutional neural networks," *Nature Communications*, vol. 11, no. 1, Sep. 2020. [Online]. Available: https://doi.org/10.1038/s41467-020-18197-y

[26] Y. Zhu, T. Brettin, F. Xia, A. Partin, M. Shukla, H. Yoo, Y. A. Evrard, J. H. Doroshow, and R. L. Stevens, "Converting tabular data into images for deep learning with convolutional neural networks," *Scientific Reports*, vol. 11, no. 1, May 2021. [Online]. Available: https://doi.org/10.1038/s41598-021-90923-y

[27] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, "Grad-cam: Visual explanations from deep networks via gradient-based localization," 2016. [Online]. Available: https://arxiv.org/abs/1610.02391

[28] T. He, Z. Zhang, H. Zhang, Z. Zhang, J. Xie, and M. Li, "Bag of tricks for image classification with convolutional neural networks," 2018. [Online]. Available: https://arxiv.org/abs/1812.01187

[29] Y. Imoto, T. Nakamura, E. G. Escolar, M. Yoshiwaki, Y. Kojima, Y. Yabuta, Y. Katou, T. Yamamoto, Y. Hiraoka, and M. Saitou, "Resolution of the curse of dimensionality in single-cell RNA sequencing data analysis," *Life Science Alliance*, vol. 5, no. 12, p. e202201591, Aug. 2022. [Online]. Available: https://doi.org/10.26508/lsa.202201591

[30] A. Kratz and P. Carninci, "The devil in the details of RNA-seq," *Nature Biotechnology*, vol. 32, no. 9, pp. 882–884, Sep. 2014. [Online]. Available: https://doi.org/10.1038/nbt.3015

[31] S. Krishnan and Y. Athavale, "Trends in biomedical signal feature extraction," *Biomedical Signal Processing and Control*, vol. 43, pp. 41–63, May 2018. [Online]. Available: https://doi.org/10.1016/j.bspc.2018.02.008

[32] J. M. Kernbach and V. E. Staartjes, "Foundations of machine learning-based clinical prediction modeling: Part II—generalization and overfitting," in *Acta Neurochirurgica Supplement*. Springer International Publishing, Dec. 2021, pp. 15–21. [Online]. Available: https://doi.org/10.1007/978-3-030-85292-4_3

[33] L. Li, W.-K. Ching, and Z.-P. Liu, "Robust biomarker screening from gene expression data by stable machine learning-recursive feature elimination methods," *Computational Biology and Chemistry*, vol. 100, p. 107747, Oct. 2022. [Online]. Available: https://doi.org/10.1016/j.compbiolchem.2022.107747

[34] B. Boucher and S. Jenna, "Genetic interaction networks: better understand to better predict," *Frontiers in Genetics*, vol. 4, 2013. [Online]. Available: https://doi.org/10.3389/fgene.2013.00290

[35] N. Pudjihartono, T. Fadason, A. W. Kempa-Liehr, and J. M. O'Sullivan, "A review of feature selection methods for machine learning-based disease risk prediction," *Frontiers in Bioinformatics*, vol. 2, Jun. 2022. [Online]. Available: https://doi.org/10.3389/fbinf.2022.927312

[36] H. Almazrua and H. Alshamlan, "A comprehensive survey of recent hybrid feature selection methods in cancer microarray gene expression data," *IEEE Access*, vol. 10, pp. 71 427–71 449, 2022. [Online]. Available: https://doi.org/10.1109/access.2022.3185226

[37] W. Jaisingh, S. C. B. Jaganathan, and A. Verma, "Gene selection by hybrid feature selection approaches and classification techniques in microarray dataset for cancer prediction," in *2022 IEEE 2nd International Symposium on Sustainable Energy, Signal Processing and Cyber Security (iSSSC)*. IEEE, Dec. 2022. [Online]. Available: https://doi.org/10.1109/isssc56467.2022.10051247

[38] D. M. D. Raj and R. Mohanasundaram, "An efficient filter-based feature selection model to identify significant features from high-dimensional microarray data," *Arabian Journal for Science and Engineering*, vol. 45, no. 4, pp. 2619–2630, Feb. 2020. [Online]. Available: https://doi.org/10.1007/s13369-020-04380-2

[39] K. Rezaee, G. Jeon, M. R. Khosravi, H. H. Attar, and A. Sabzevari, "Deep learning-based microarray cancer classification and ensemble gene selection approach," *IET Systems Biology*, vol. 16, no. 3-4, pp. 120–131, May 2022. [Online]. Available: https://doi.org/10.1049/syb2.12044

[40] S. Zhang, W. Xie, W. Li, L. Wang, and C. Feng, "GAMB-GNN: Graph neural networks learning from gene structure relations and markov blanket ranking for cancer classification in microarray data," *Chemometrics and Intelligent Laboratory Systems*, vol. 232, p. 104713, Jan. 2023. [Online]. Available: https://doi.org/10.1016/j.chemolab.2022.104713

[41] W. Xie, W. Li, S. Zhang, L. Wang, J. Yang, and D. Zhao, "A novel biomarker selection method combining graph neural network and gene relationships applied to microarray data," *BMC Bioinformatics*, vol. 23, no. 1, Jul. 2022. [Online]. Available: https://doi.org/10.1186/s12859-022-04848-y

[42] K. Yu, W. Xie, L. Wang, S. Zhang, and W. Li, "Determination of biomarkers from microarray data using graph neural network and

spectral clustering," *Scientific Reports*, vol. 11, no. 1, Dec. 2021. [Online]. Available: https://doi.org/10.1038/s41598-021-03316-6

[43] Y. Wang, D. J. Miller, and R. Clarke, "Approaches to working in high-dimensional data spaces: gene expression microarrays," *British Journal of Cancer*, vol. 98, no. 6, pp. 1023–1028, Feb. 2008. [Online]. Available: https://doi.org/10.1038/sj.bjc.6604207

[44] O. Fajarda, J. R. Almeida, S. Duarte-Pereira, R. M. Silva, and J. L. Oliveira, "Methodology to identify a gene expression signature by merging microarray datasets," *Computers in Biology and Medicine*, vol. 159, p. 106867, Jun. 2023. [Online]. Available: https://doi.org/10.1016/j.compbiomed.2023.106867

[45] M. Kenn, D. C. Castillo-Tong, C. F. Singer, M. Cibena, H. Kölbl, and W. Schreiner, "Microarray normalization revisited for reproducible breast cancer biomarkers," *BioMed Research International*, vol. 2020, pp. 1–27, Aug. 2020. [Online]. Available: https://doi.org/10.1155/2020/1363827

[46] Z. Fu, F. Zhang, Q. Yin, R. Li, W. Hu, and W. Li, "Small sample learning optimization for resnet based sar target recognition," in *IGARSS 2018 - 2018 IEEE International Geoscience and Remote Sensing Symposium*. IEEE, Jul. 2018. [Online]. Available: https://doi.org/10.1109/igarss.2018.8517574

[47] A. Sharma, A. Lysenko, K. A. Boroevich, E. Vans, and T. Tsunoda, "DeepFeature: feature selection in nonimage data using convolutional neural network," *Briefings in Bioinformatics*, vol. 22, no. 6, Aug. 2021. [Online]. Available: https://doi.org/10.1093/bib/bbab297