IJCRT.ORG ISSN : 2320-2882

INTERNATIONAL JOURNAL OF CREATIVE RESEARCH THOUGHTS (IJCRT)

An International Open Access, Peer-reviewed, Refereed Journal

A Classifier Based Comparative Approach On Gene Expression

Kumari Pritee[1], Malavika Samanthapudi[2]

[1]Assistant Professor, Information System Management, IIM Sambalpur

[2]Scholar, Adamas University,Kolkata

Abstract

Gene expression classification has emerged as a powerful method to classify patients with Acute Myeloid Leukemia (AML) and Acute Lymphoblastic Leukemia (ALL). Classification of cancer based on gene expression has provided insight into possible treatment strategies. By analyzing gene expression profiles through DNA microarray technology, new cases of cancer can be accurately classified, providing a general approach for identifying new cancer classes and assigning tumors to known classes.

In this study, Feature Selection techniques were employed to identify the most informative genes, which served as input for the classification models and also Dimension Reduction techniques like Principle Component Analysis (PCA) were applied. The Machine Learning algorithms such as Support Vector Machine (SVM) and Naïve Bayes were applied to the dataset, enabling the classification of cancer cases based on their gene expression profiles.

The results indicated that both SVM and Naive Bayes classifiers can effectively classify gene expression data into the ALL and AML categories. SVM demonstrated higher accuracy, precision, and recall compared to Naive Bayes, making it more suitable for overall classification tasks. However, Naive Bayes is slightly slower than SVM but exhibited competitive performance particularly in terms of recall, indicating its proficiency in correctly identifying positive instances.

List of Abbreviation

Abbreviation	Full form
DNA	Deoxyribonucleic acid
RNA	Ribonucleic acid
mRNA	Messenger Ribonucleic acid
AML	Acute Myeloid Leukemia
ALL	Acute Lymphoblastic Leukemia
SVM	Support Vector Machine
PCA	Principal Component Analysis

Keywords:- DNA, RNA, mRNA, AML, ALL, SVM & PCA

1. Introduction

Using DNA microarray technology to monitor gene expression, gene expression classification has become a potent tool in cancer diagnosis[1], offering a general method for categorising new cancer cases and discovering both established and novel cancer classes. Researchers have been able to create more focused and efficient treatment plans by studying the patterns of gene expression in cancer cells. This analysis has provided important insights into the underlying molecular mechanisms of the disease[2]. Tumours can be categorised according to their gene expression profiles according to this method, which offers a thorough understanding of the gene activity within cancer cells. The most prevalent form of blood cancer across all age categories, especially in youngsters, is leukaemia. This results from immature growth and excessive blood cell proliferation, which can damage the immune system, brain tissue, and red blood cells[2][3]. The genetic code dictates to a cell when it should divide and when it should expire. Gene expression variations may result in faulty instructions, which can cause cancer. Leukaemia, myeloma, and lymphoma are the three primary categories of blood malignancies. Leukaemias, which start in the bone marrow's blood-forming tissue, are malignancies of the blood and bone marrow[4][5].

The categorization of patients with acute myeloid leukaemia (AML) and acute lymphoblastic leukaemia (ALL), two severe types of blood cancer, has benefited greatly from this method. An aggressive blood malignancy known as AML (Acute Myeloid Leukaemia) is typified by the fast proliferation of aberrant myeloid cells in the blood and bone marrow[6][7][8] Adults are typically affected, and symptoms include anaemia, easy bruising or bleeding, infections, and exhaustion. On the other hand, ALL (Acute Lymphoblastic Leukaemia) is a malignancy that affects lymphoid cells and grows quickly; it is more common in children. It

is characterized by an overproduction of immature lymphoblast, leading to symptoms such as fatigue, pale skin, recurrent infections, and bone or joint pain.

In order to find the most informative genes to use as input for the classification models, feature selection approaches were used in the work [9][10]. Additionally, dimension reduction techniques such as Principle Component Analysis (PCA) were implemented. The dataset was subjected to Machine Learning methods, including Support Vector Machine (SVM) and Naïve Bayes, which allowed for the classification of cancer patients according to their gene expression profiles. Furthermore, an accuracy comparison between the Support Vector Machine (SVM) and Naive Bayes classification models was carried out. These machine learning algorithms are able to discriminate between the two kinds of leukaemia because they are trained on gene expression data from known cases of ALL and AM[11][12]13]L. It is possible to assess the performance features of the Support Vector Machine (SVM) and Naive Bayes models by comparing their accuracy. The usefulness of each model for correctly categorising AML and ALL patients according to their gene expression profiles is ascertained by this comparison. Comprehending the advantages and disadvantages of these approaches advances the categorization of gene expression in leukaemia instances.

1.1 Gene Expression

The process via which the data contained in a gene is transformed into a useful product, like a protein or RNA molecule, is known as gene expression. The instructions needed to produce particular proteins, which are necessary for a variety of biological processes and activities within an organism, are found in genes. Transcription and translation are the two primary steps in the expression of genes. An enzyme known as RNA polymerase copies a gene's DNA sequence during transcription to create a corresponding RNA molecule known as messenger RNA (mRNA)[14]. The genetic information from the gene is transferred by this mRNA molecule to the cellular machinery in charge of protein synthesis.

The mRNA molecule interacts with ribosomes, which are cellular organelles in charge of protein synthesis, during the translation stage[15]. The mRNA molecule's nucleotide sequence is "read" by ribosomes, which then utilise this information to put amino acids together in a particular order to produce polypeptide chains. After that, this chain folds into a useful protein that performs its designated function within the cell [16].

A number of variables, including as developmental phases, cellular signals, and environmental cues, can affect the carefully controlled process of gene expression. It is essential for defining an organism's traits, capabilities, and reactions to its surroundings. When genes are expressed inappropriately or mutatedly, interrupting normal cellular functions[17], the result can be diseases such as cancer. Deciphering the patterns and regulation of gene expression is crucial in order to identify the molecular mechanisms that underlie illnesses and biological processes.

1.2 Gene Expression Classification

The practice of classifying samples or patients according to their gene expression profiles is known as "gene expression classification." It entails examining the patterns of gene expression in biological samples, such as tumour tissues or cells, and applying computational techniques to discern unique patterns of gene expression associated with various classes or categories of interest, including subtypes of disease or various forms of cancer[18][19]. There are normally two primary processes in the classification process. To create a classification model[20][21], gene expression data from a set of samples with predetermined classifications or categories is first used in a training phase. This model gains the ability to identify and distinguish between the gene expression patterns linked to various classes. During this training phase, a variety of statistical algorithms, machine learning approaches[22], and pattern recognition techniques can be used to create an appropriate classification model.

After training, the model can be used to categorise previously unobserved samples into the pre-established classes. A prediction is formed about the class or category to which the samples belong based on an analysis of their gene expression patterns against the model's learned patterns. Insights into disease diagnosis, prognosis prediction, therapy response prediction, and even the discovery of new subclasses within an illness can all be gained via this categorization method. Classifying gene expression has been particularly effective in cases of acute lymphoblastic leukaemia (ALL) and acute myeloid leukaemia (AML)[23]. Researchers have discovered distinct gene expression patterns particular to each subtype of leukaemia by examining the gene expression profiles of patients with the disease. These patterns are useful indicators that help differentiate AML from ALL, enabling precise categorization and guiding treatment choices[24][25].

1.3 Significance

The classification of gene expression is important for cancer research and therapeutic treatment. The following are some main arguments for the importance of classifying gene expression:

- Precise Disease Classification: Different cancer kinds can be precisely categorised and classified by the use of gene expression classification. Through the examination of gene expression patterns, scientists can pinpoint distinct molecular profiles that are particular to particular cancer subtypes or kinds[26]. This data is essential for a precise diagnosis and allows for customised treatment plans depending on the unique features of the cancer.
- Tailored Treatment Choices: The categorization of gene expression offers important insights into the aggressiveness and behaviour of tumours. Clinicians can make educated decisions about the best course of treatment for individual patients by having a thorough understanding of the gene expression patterns linked to various cancer types. With the use of this approach, personalised medicine is made

possible[27], allowing for the customisation of medicines to specifically target molecular changes that improve patient outcomes.

- Predictive and prognostic Biomarkers: Classifying gene expression can provide biomarkers linked to prognosis and response to treatment. Through the examination of gene expression patterns, scientists can pinpoint distinct molecular signatures that signify the advancement of a disease, the likelihood of survival[28], or the reaction to a given treatment. By using these indicators, medical professionals can better determine a patient's prognosis and choose the best therapeutic approaches, increasing the likelihood that the patient will respond to therapy.
- Discovery of Novel Cancer Subtypes: Classification of gene expression can reveal uncommon or until unidentified cancer subtypes. Through the examination of extensive gene expression profile databases, scientists are able to pinpoint unique molecular subgroups that belong to a specific form of cancer. This finding may contribute to our comprehension of the illness, the discovery of novel therapeutic targets, and the creation of specialised treatment plans for these particular subtypes.[29][30]
- Promoting Cancer Research: Classifying gene expressions advances our knowledge of cancer biology. Through the examination of gene expression patterns linked to distinct cancer types and subtypes, scientists can acquire a deeper understanding of the molecular processes that underlie the initiation, advancement, and reaction to therapeutic interventions in cancer. This information contributes to the advancement of cancer research overall and to the creation of innovative therapeutic approaches [31][32].

2. Literature Review

Six simulation-based classifiers have been chosen by [33] algorithms, and the tree random classifier in one of them achieves an accuracy gain of 98 percent, a significantly substantial classification accuracy. The optimal classifier methods were predicted by comparing them with different supervised algorithms. Experiments validated the practicality of the suggested approach. Tests of the template's recall and accuracy were conducted. All other high-accuracy algorithms were found to perform better than random trees of various categorization strategies.

By combining the predictions of many classifiers, a convergent learning-based model for leukaemia classification from gene expression by 2021 was proposed, which would increase classification accuracy [34]. The performance measure increases the handling of ionosphere data in terms of precision, accuracy, specificity, and sensitivity by employing the CART, CHAID, and QUEST classifications. According to the experimental findings, 93.84 percent of the test data selection for the ensemble model with the function choice attained perfect accuracy.

An Application of Classification Framework has been proposed to Cancer Gene Expression Profiles and the various supervised learning methods used to categorise tumours [35]. These results suggest that another individual classifier gets the best performance from the voting categories. Early and accurate cancer diagnosis boosts survival from 56% to more than 57%, resulting in an 86 percent reduction in death rates from cancer.

Using a clustering algorithm to analyse tumours and normal colon tissues, the main goal was to predict the initial leukaemia disease using various machine learning algorithms, such as SVM, Naïve Bayes, Decision Tree, and Linear Regression [36]. The goal was to successfully predict leukaemia in patients and enhance prediction accuracy in the shortest amount of time. Classifier algorithms with the ability to diagnose and forecast various conditions were trained using microarray data.

3. Dataset

The Curated Microarray Database (CuMiDa) (Feltes et al., 2019) is a repository that houses 78 carefully chosen and cross-checked cancer Microarray datasets from 30,000 Gene Expression Omnibus (GEO) research. This is where the data used in this paper were taken from. In order to produce a more dependable data source, CuMiDa provides a more recent dataset that has been manually and meticulously chosen, with sample quality, undesired probe extraction, background correction, and normalisation. You can access these statistics at https://sbcb.inf.ufrgs.br/cumida

It includes the test and training sets used in Golub et al.'s work (Golub et al., 1999). Measurements pertaining to bone marrow and peripheral blood samples from ALL and AML are included in these databases. To make the overall intensities of each chip equal, the intensity values have been scaled. The dataset contains three files:

1. Actual.csv: This file contains the identification of all 72 patients in the study and their labels (type of cancer, 47 ALL and 25 AML).
2. Data_set_ALL_AML_train.csv: This file contains the subset with training data (38 bone marrow samples, 7129 genes).
3. Data_set_ALL_AML_independent.csv: This file contains the subset with the test data (34 peripheral blood samples, 7129 genes).

1	patient	cancer
2	1	ALL
3	2	ALL
4	3	ALL
5	4	ALL
6	5	ALL
7	6	ALL
8	7	ALL
9	8	ALL
10	9	ALL

Fig 1. Actual file

1	Gene Description	Gene Accession Number	1	2	3	4
2	AFFX-BioB-5_at (endogenous control)	AFFX-BioB-5_at	-214	-139	-76	-135
3	AFFX-BioB-M_at (endogenous control)	AFFX-BioB-M_at	-153	-73	-49	-114
4	AFFX-BioB-3_at (endogenous control)	AFFX-BioB-3_at	-58	-1	-307	265
5	AFFX-BioC-5_at (endogenous control)	AFFX-BioC-5_at	88	283	309	12
6	AFFX-BioC-3_at (endogenous control)	AFFX-BioC-3_at	-295	-264	-376	-419
7	AFFX-BioDn-5_at (endogenous control)	AFFX-BioDn-5_at	-558	-400	-650	-585
8	AFFX-BioDn-3_at (endogenous control)	AFFX-BioDn-3_at	199	-330	33	158
9	AFFX-CreX-5_at (endogenous control)	AFFX-CreX-5_at	-176	-168	-367	-253
10	AFFX-CreX-3_at (endogenous control)	AFFX-CreX-3_at	252	101	206	49

1	Gene Description	Gene Accession Number	39	40	42	47
2	AFFX-BioB-5_at (endogenous control)	AFFX-BioB-5_at	-342	-87	22	-243
3	AFFX-BioB-M_at (endogenous control)	AFFX-BioB-M_at	-200	-248	-153	-218
4	AFFX-BioB-3_at (endogenous control)	AFFX-BioB-3_at	41	262	17	-163
5	AFFX-BioC-5_at (endogenous control)	AFFX-BioC-5_at	328	295	276	182
6	AFFX-BioC-3_at (endogenous control)	AFFX-BioC-3_at	-224	-226	-211	-289
7	AFFX-BioDn-5_at (endogenous control)	AFFX-BioDn-5_at	-427	-493	-250	-268
8	AFFX-BioDn-3_at (endogenous control)	AFFX-BioDn-3_at	-656	367	55	-285
9	AFFX-CreX-5_at (endogenous control)	AFFX-CreX-5_at	-292	-452	-141	-172
10	AFFX-CreX-3_at (endogenous control)	AFFX-CreX-3_at	137	194	0	52

Fig 3. Data_set_ALL_AML_independent file

4. Methodology

4.1 Machine Learning Pipeline

- Data Collection: Assemble a dataset for every sample that contains information on gene expression and the related class labels. Make sure the dataset includes a variety of interest classes or categories.
- Data preprocessing: Take care of missing values and eliminate duplicates from the data. To make comparisons between samples possible, normalise the gene expression data.
- Feature Selection: Use feature selection methods to find the most informative genes in a dataset with a large number of genes. Choose a selection of genes that are most important for categorization to reduce dimensionality.

- Data Split: Create training and testing subsets from the preprocessed dataset. The testing set will be utilised for evaluation, while the training set will be used to train the SVM model.
- Model Training: Using the training dataset, train an SVM model. The supervised learning algorithm SVM determines the best hyperplane to divide samples of various classes.
- Model Optimisation: To enhance the performance of the SVM model, adjust its hyperparameters. Finding the ideal hyperparameter values can be accomplished by methods such as grid search or random search.
- Principal Component Analysis: A popular dimensionality reduction method in data analysis and machine learning is principal component analysis, or PCA. Its goal is to retain the most important information while converting high-dimensional data into a lower-dimensional representation. Finding the principal components—the paths along which the data fluctuates most—is the fundamental idea of principal component analysis (PCA).
- Model Evaluation: Make use of the testing dataset to assess the trained SVM model. Determine performance parameters including recall, accuracy, precision, and F1-score to evaluate the classification performance of the model.
- Support Vector Machine, or SVM, is used to solve Regression and Classification problems. But it's mostly applied to machine learning classification challenges. In order to make it simple to classify fresh data points in the future, the SVM method seeks to identify the optimal line or decision boundary that can divide n-dimensional space into classes. We refer to this optimal decision boundary as a hyperplane.
- One of the most straightforward and efficient classification algorithms, the Naïve Bayes classifier aids in the rapid development of machine learning models with rapid prediction capabilities. Being a probabilistic classifier, it makes predictions based on the likelihood that an object will occur. Sentimental analysis and spam filtration are two well-known applications of the Naïve Bayes algorithm, where the feature independence assumption frequently holds up rather well. The algorithm is named "naive" because, given the class label, it presumes that every feature is independent of every other feature.
- Performance Comparison: Examine how well the SVM model performs in comparison to alternative classification methods or algorithms. Evaluate the benefits and drawbacks of the SVM model for classifying gene expression

.

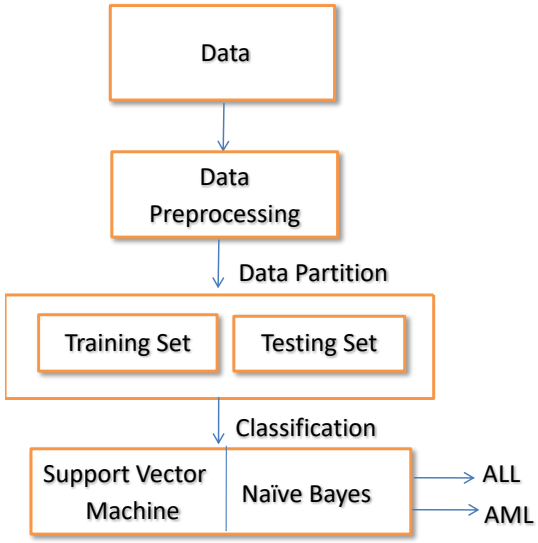

Fig 4. Architecture

4. Result and Discussions

The results of the Gene Expression Classification are presented in this section. Our goal in the project work was to use Support Vector Machine (SVM) and Naive Bayes algorithms to categorise patients with AML and ALL based on gene expression data. Using a confusion matrix, Support Vector Machine (SVM) and Naïve Baiyes were compared. By comparing the expected labels with the actual labels, this matrix enabled a thorough study that gave important insights into how well each model performed. These algorithms' performance was assessed in terms of F1 score, recall, accuracy, and precision.

When compared to Naive Bayes, SVM performed better at reliably classifying AML and ALL. With an accuracy of 0.941, SVM was able to classify 94.1% of the instances correctly; additionally, the precision of 0.875 indicates that 87.5% of the instances classified as AML or ALL were true positive predictions; and the recall of 1.0 indicates that SVM correctly identified all instances of AML. The harmonic mean of recall and precision is represented by the F1 score of 0.933, which denotes overall strong performance. Naive Bayes, on the other hand, performed somewhat worse, with accuracy of 0.912, precision of 0.867, F1 score of 0.897, and recall of 0.929. These findings imply that, when using gene expression data to discriminate between AML and ALL, Naive Bayes is less reliable compared to SVM.

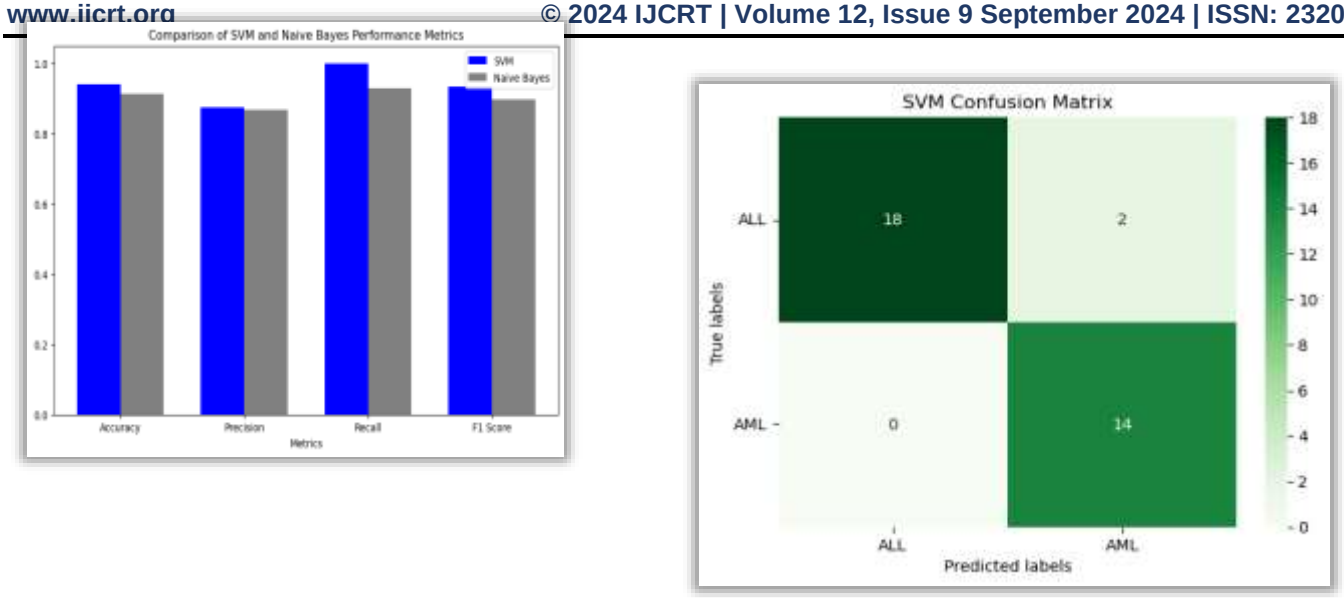

Fig 5. Confusion Matrix of Support Vector Machine and Naïve Bayes

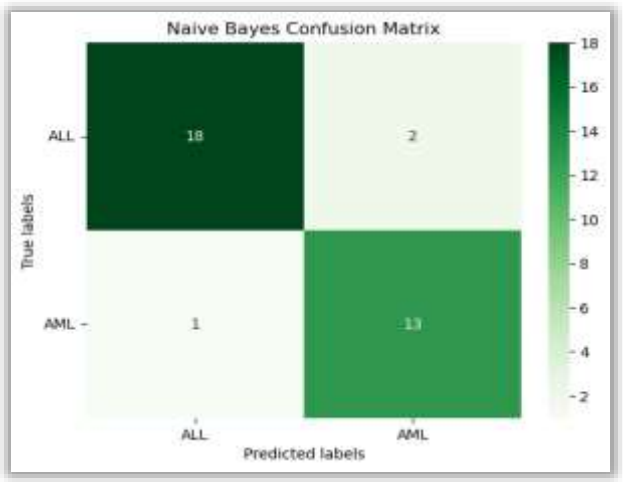

Fig 6. Comparison of SVM and Naïve Bayes Performance Metrics

6. Conclusion

In the project work, we used two supervised learning algorithms, a Support Vector Machine (SVM) and Naïve Bayes, to classify the type of cancer in each patient to ALL and AML. Based on their gene expression data, we applied two machine learning techniques to classify the two types of leukaemia (ALL and AML). Various performance indicators were used to compare Naïve Baiyes and Support Vector Machine (SVM). Based on gene expression data, SVM outperformed Naïve Bayes in effectively diagnosing AML and ALL patients, with an accuracy of 0.941, meaning that 94.1% of the cases were correctly identified. Naive Bayes, on the other hand, performed marginally worse, with an accuracy of 0.912. Higher recall values, F1 score, accuracy, and precision were all displayed by SVM, demonstrating its superiority for this specific classification task.

7. Future Work

In the future, we can investigate and apply cutting-edge machine learning methods and algorithms, such as ensemble learning and deep learning models, to enhance the performance and classification accuracy of gene expression data.

Expand the categorization models to encompass additional disease categories and investigate their efficacy across various datasets. This will facilitate the evaluation of the models' adaptability and generalizability in various biomedical settings.

8. References

[1]. Thakur, T.; Batra, I.; Luthra, M.; Vimal, S.; Dhiman, G.; Malik, A.; Shabaz, M. Gene expression-assisted cancer prediction techniques. J. Healthc. Eng. 2021, 2021, 4242646. [Google Scholar] [CrossRef] [PubMed]

[2]. Ahluwalia, P.; Kolhe, R.; Gahlay, G.K. The clinical relevance of gene expression based prognostic signatures in colorectal cancer. Biochim. Biophys. Acta Rev. Cancer 2021, 1875, 188513. [Google Scholar] [CrossRef] [PubMed]

[3]. Schaafsma, E.; Fugle, C.M.; Wang, X.; Cheng, C. Pan-cancer association of HLA gene expression with cancer prognosis and immunotherapy efficacy. Br. J. Cancer 2021, 125, 422–432. [Google Scholar] [CrossRef]

[4]. Tourang, M.; Fang, L.; Zhong, Y.; Suthar, R.C. Association between Human Endogenous Retrovirus K gene expression and breast cancer. Cell. Mol. Biomed. Rep. 2021, 1, 7–13. [Google Scholar] [CrossRef]

[5]. Satyananda, V.; Oshi, M.; Endo, I.; Takabe, K. High BRCA2 gene expression is associated with aggressive and highly proliferative breast cancer. Ann. Surg. Oncol. 2021, 28, 7356–7365. [Google Scholar] [CrossRef]

[6]. Qian, Y.; Daza, J.; Itzel, T.; Betge, J.; Zhan, T.; Marmé, F.; Teufel, A. Prognostic cancer gene expression signatures: Current status and challenges. Cells 2021, 10, 648. [Google Scholar] [CrossRef] [PubMed]

[7]. Munkácsy, G.; Santarpia, L.; Győrffy, B. Gene Expression Profiling in Early Breast Cancer—Patient Stratification Based on Molecular and Tumor Microenvironment Features. Biomedicines 2022, 10, 248. [Google Scholar] [CrossRef]

[8]. Oliveira, L.J.C.; Amorim, L.C.; Megid, T.B.C.; De Resende, C.A.A.; Mano, M.S. Gene expression signatures in early Breast Cancer: Better together with clinicopathological features. Crit. Rev. Oncol. Hematol. 2022, 175, 103708. [Google Scholar] [CrossRef]

[9]. Schettini, F.; Chic, N.; Brasó-Maristany, F.; Paré, L.; Pascual, T.; Conte, B.; Martínez-Sáez, O.; Adamo, B.; Vidal, M.; Barnadas, E.; et al. Clinical, pathological, and PAM50 gene expression features of HER2-low breast cancer. NPJ Breast Cancer 2021, 7, 1. [Google Scholar] [CrossRef]

[10]. Zhong, Y.; Chalise, P.; He, J. Nested cross-validation with ensemble feature selection and classification model for high-dimensional biological data. Commun. Stat. Simul. Comput. 2023, 52, 110–125. [Google Scholar] [CrossRef]

[11]. Petinrin, O.O.; Saeed, F.; Li, X.; Ghabban, F.; Wong, K.C. Reactions' descriptors selection and yield estimation using metaheuristic algorithms and voting ensemble. Comput. Mater. Contin. 2022, 70, 4745–4762. [Google Scholar]

[12]. Hameed, S.S.; Petinrin, O.O.; Hashi, A.O.; Saeed, F. Filter-wrapper combination and embedded feature selection for gene expression data. Int. J. Adv. Soft Compu. Appl. 2018, 10, 90–105. [Google Scholar]

[13]. Townes, F.W.; Hicks, S.C.; Aryee, M.J.; Irizarry, R.A. Feature selection and dimension reduction for single-cell RNA-Seq based on a multinomial model. Genome Biol. 2019, 20, 295. [Google Scholar] [CrossRef] [PubMed] [Green Version]

[14]. Jain, I.; Jain, V.K.; Jain, R. Correlation feature selection based improved-binary particle swarm optimization for gene selection and cancer classification. Appl. Soft Comput. 2018, 62, 203–215. [Google Scholar] [CrossRef]

[15]. Kabir, M.F.; Chen, T.; Ludwig, S.A. A performance analysis of dimensionality reduction algorithms in machine learning models for cancer prediction. Healthc. Anal. 2023, 3, 100125. [Google Scholar] [CrossRef]

[16]. Prasad, Y.; Biswas, K.; Hanmandlu, M. A recursive PSO scheme for gene selection in microarray data. Appl. Soft Comput. 2018, 71, 213–225. [Google Scholar] [CrossRef]

[17]. Sharbaf, F.V.; Mosafer, S.; Moattar, M.H. A hybrid gene selection approach for microarray data classification using cellular learning automata and ant colony optimization. Genomics 2016, 107, 231–238. [Google Scholar] [CrossRef]

[18]. Alhenawi, E.A.; Al-Sayyed, R.; Hudaib, A.; Mirjalili, S. Improved intelligent water drop-based hybrid feature selection method for microarray data processing. Comput. Biol. Chem. 2023, 103, 107809. [Google Scholar] [CrossRef]

[19]. Keshta, I.; Deshpande, P.S.; Shabaz, M.; Soni, M.; Bhadla, M.K.; Muhammed, Y. Multi-stage biomedical feature selection extraction algorithm for cancer detection. SN Appl. Sci. 2023, 5, 131. [Google Scholar] [CrossRef]

[20]. Sayed, S.; Nassef, M.; Badr, A.; Farag, I. A nested genetic algorithm for feature selection in high-dimensional cancer microarray datasets. Expert Syst. Appl. 2019, 121, 233–243. [Google Scholar] [CrossRef]

[21]. Li, X.; Wang, H. On Mean-Optimal Robust Linear Discriminant Analysis. In Proceedings of the 2022 IEEE International Conference on Data Mining (ICDM), Orlando, FL, USA, 30 November–3 December 2022; pp. 1047–1052. [Google Scholar]

[22]. Li, X.; Wang, H. Adaptive Principal Component Analysis. In Proceedings of the 2022 SIAM International Conference on Data Mining (SDM), Alexandria, VA, USA, 28–30 April 2022; Society for Industrial and Applied Mathematics: Philadelphia, PA, USA, 2022; pp. 486–494. [Google Scholar]

[23]. Jiang, J.; Xu, J.; Liu, Y.; Song, B.; Guo, X.; Zeng, X.; Zou, Q. Dimensionality reduction and visualization of single-cell RNA-seq data with an improved deep variational autoencoder. Briefings Bioinform. 2023, 24, bbad152. [Google Scholar] [CrossRef] [PubMed]

[24]. Hameed, S.S.; Muhammad, F.F.; Hassan, R.; Saeed, F. Gene Selection and Classification in Microarray Datasets using a Hybrid Approach of PCC-BPSO/GA with Multi Classifiers. J. Comput. Sci. 2018, 14, 868–880. [Google Scholar] [CrossRef] [Green Version]

[25]. Dettling, M.; Bühlmann, P. Supervised clustering of genes. Genome Biol. 2002, 3, research0069.1. [Google Scholar] [CrossRef]

[26]. Alon, U.; Barkai, N.; Notterman, D.A.; Gish, K.; Ybarra, S.; Mack, D.; Levine, A.J. Broad patterns of gene expression revealed by clustering analysis of tumor and normal colon tissues probed by oligonucleotide arrays. Proc. Natl. Acad. Sci. USA 1999, 96, 6745–6750. [Google Scholar] [CrossRef] [PubMed] [Green Version]

[27]. Zhu, Z.; Ong, Y.S.; Dash, M. Markov Blanket-Embedded Genetic Algorithm for Gene Selection. Pattern Recognit. 2007, 49, 3236–3248. [Google Scholar] [CrossRef]

[28]. Microarray Datasets. Available online: https://csse.szu.edu.cn/staff/zhuzx/Datasets.html (accessed on 8 June 2023).

[29]. Golub, T.R.; Slonim, D.K.; Tamayo, P.; Huard, C.; Gaasenbeek, M.; Mesirov, J.P.; Coller, H.; Loh, M.L.; Downing, J.R.; Caligiuri, M.A.; et al. Molecular classification of cancer: Class discovery and class prediction by gene expression monitoring. Science 1999, 286, 531–537. [Google Scholar] [CrossRef] [PubMed] [Green Version]

[30]. Dudoit, S.; Fridlyand, J.; Speed, T.P. Comparison of discrimination methods for the classification of tumors using gene expression data. J. Am. Stat. Assoc. 2002, 97, 77–87. [Google Scholar] [CrossRef] [Green Version]

[31]. Díaz-Uriarte, R.; De Andres, S.A. Gene selection and classification of microarray data using random forest. BMC Bioinform. 2006, 7, 3. [Google Scholar] [CrossRef] [Green Version]

[32]. Nguyen, H.M.; Cooper, E.W.; Kamei, K. Borderline over-sampling for imbalanced data classification. In Proceedings of the Fifth International Workshop on Computational Intelligence & Applications, IEEE SMC Hiroshima Chapter, Hiroshima, Japan, 10–12 November 2009. [Google Scholar]

[33]. Alharbi, F.; Vakanski, A. Machine Learning Methods for Cancer Classification Using Gene Expression Data: A Review. Bioengineering 2023, 10, 173. https://doi.org/10.3390/bioengineering10020173

[34]. Mallick PK, Mohapatra SK, Chae GS, Mohanty MN. Convergent learning-based model for leukemia classification from gene expression. Pers Ubiquitous Comput. 2023;27(3):1103-1110. doi: 10.1007/s00779-020-01467-3. Epub 2020 Oct 16. PMID: 33100943; PMCID: PMC7567412.

[35]. Hijazi H, Chan C. A classification framework applied to cancer gene expression profiles. J Healthc Eng. 2013;4(2):255-83. doi: 10.1260/2040-2295.4.2.255. PMID: 23778014; PMCID: PMC3873740.

[36]. A. Karim, A. Azhari, M. Shahroz, S. Brahim Belhaouri and K. Mustofa, "Ldsvm: leukemia cancer classification using machine learning," Computers, Materials & Continua, vol. 71, no.2, pp. 3887–3903, 2022.