ISSN: 2230-9926

Available online at http://www.journalijdr.com

IJDR

International Journal of Development Research
Vol. 11, Issue, 09, pp. 50641-50646, September, 2021
https://doi.org/10.37118/ijdr.22977.09.2021

RESEARCH ARTICLE OPEN ACCESS

LEUKEMIA DIAGNOSIS WITH MACHINE LEARNING ENSEMBLE FROM GENE EXPRESSION DATA

Jakelyne Machado Lima Silva[1], Joaquim dos Santos Costa[1], Edson Magalhaes da Costa [1], Maria Eliana da Silva Holanda [1], Lucas Henrique Martins Soares [1], Dhian Kelson Leite de Oliveira[1], Fabrício Almeida Araújo[2,3], Guilherme Damasceno Silva[4], Isadora Mendes dos Santos[1], Gilberto Nerino de Souza Junior[1] and Marcus de Barros Braga[1*]

[1]Universidade Federal Rural da Amazônia. Campus Paragominas.Paragominas, Pará. Brasil; [2]Universidade Federal Rural da Amazônia, Programa de Pós-Graduação em Biotecnologia Aplicada à Agropecuária, Belém, Pará, Brasil; [3]Universidade Federal do Pará. Campus Castanhal, Castanhal. Pará. Brasil; [4]Instituto Federal de Educação, Ciência e Tecnologia do Pará. Campus Ananindeua, Ananindeua. Pará, Brasil

ARTICLE INFO

Article History:
Received 06th August, 2021
Received in revised form
14th August, 2021
Accepted 06th September, 2021
Published online 30th September, 2021

Key Words:
Leukemia, Classification, Machine Learning, Ensemble Learning, Bioinformatics.

***Corresponding author:**
Marcus de Barros Braga

ABSTRACT

One of the great challenges of treating leukemia is targeting specific therapies for different categories. Classification models have been improved, making them decisive for improving the treatment of the disease. In this study, gene expression data was used and then different computational machine learning models were applied to establish the diagnosis of *Acute Lymphoblastic Leukemia* and *Acute Myeloid Leukemia* type leukemias. Three approaches, combined with data mining techniques, were used: one using a Support Vector Machine algorithm as core, the second one using an Artificial Neural Network and the third one using the Machine Learning Ensemble combination (Artificial Neural Network, Support Vector Machine, Random Forest, Gradient Boosting and k-NN). The Ensemble model achieved a consistent overall performance above 94% for five different learning algorithm evaluation metrics.

Citation: ***Jakelyne Machado Lima Silva, Joaquim dos Santos Costa, Edson Magalhaes da Costa, Maria Eliana da Silva Holanda et al.*** "Leukemia Diagnosis with Machine Learning Ensemble from Gene Expression Data*", International Journal of Development Research*, 11, (09), 50641-50646.

INTRODUCTION

There are two main classifications of leukemia, *Lymphocytic Leukemia* and *Myeloid Leukemia*, and each one can be subdivided as acute and chronic based on their aggressive growth rates (Taylor *et al*., 2017). Human acute leukemias are genetically very diverse. The consistent chromosomal changes identified in tumors point to the location of genes whose functions are critical in the growth potential of this particular type of cell. The identification of those genes located at breakpoints in dozens of translocations, many of which were previously unknown, provides unique insights into the function of these genes in normal cells, as well as their altered function in malignant cells (Rowley, 2000). One of the greatest challenges in cancer treatment has been to target specific therapies for pathogenetically distinct tumors, aimingthe maximum efficacy and minimizing toxicity.

The improvement in type of cancer classificationis decisive for improving the disease's treatment. For a long time, cancer classification was mainly based on the morphological appearance of the tumor, which has its limits. Tumors with similar histopathological appearance may follow significantly different clinical courses and show different responses to therapy. In some cases, such clinical heterogeneity is better understood by dividing morphologically similar tumors into subtypes with distinct pathogenesis, as in the case of acute leukemias (Golub *et al*., 1999). *Acute Lymphoblastic Leukemia* (ALL) is a cancer of the lymphoid line of blood cells, showing a rapid growth of immature lymphoblastic cells. *Acute Myeloid Leukemia* (AML) is a cancer of the myeloid line of blood cells. ALL is the most common form of leukemia in children, about 80% of cases, and appearing in only 20% of adults. AML occurs more frequently (80%) in adults over 60 years of age than in children. The five-year survival rate is 67% for ALL and about 40% for AML

(Siegel *et al*., 2016). However, this does not prevent them from appearing at any stage of life. In general, ALL has a better prognosis than AML. Early diagnosis is essential for the effective treatment of patients with this type of disease. The current standard diagnostic procedure relies on extensive blood count analysis, microscopic morphological investigations, bone marrow biopsy, and flow cytometry, which are all time-consuming and expensive (Masilamani *et al*., 2020). Because therapeutic strategies and prognosis vary considerably, ALL and AML must be differentiated in diagnosis (Pui *et al*., 2004; Randolph, 2004). This distinction can be achieved through the appropriate use of morphological, immunohistochemical and immunological methods (Löwenberg *et al*., 1999). Conventional clinical practice requires an experienced technical staff and no test is fully sufficient and reliable to establish the diagnosis (Mi *et al*., 2007). Distinguishing ALL from AML is critical to successful treatment. Chemotherapy regimens for ALL usually contain corticosteroids, vincristine, methotrexate, and L-asparaginase, while most AML regimens rely on a backbone of daunorubicin and cytarabine. Although remissions can be achieved using ALL therapy for AML (and vice versa), cure rates are markedly reduced and unwarranted toxicities are encountered (Pui and Evans, 1998; Bishop, 1999; Stone and Mayer, 1993).

In a pioneering study to find a more efficient diagnostic approach, Golub *et al*. [4] showed that ALL and AML can be differentiated based on gene expression profiles. Since then, the expression profile of messenger RNA (protein-coding gene) has been widely used in the classification of ALL and AML subtypes, as well as in predicting the prognosis/outcome of leukemias (Benjamin and Golub, 2004; Bullinger and Valk, 2005; Haferlach *et al*., 2007). However, the precise genes and pathways that exert critical control over the determination of lineage fate during the development of leukemia remain unclear (Mi *et al*., 2007). Microarray is a molecular biology technique where tens of thousands of probes representing a given DNA sequence are analyzed and quantified to provide a general gene expression profile of multiple biological samples. The resulting output from a Microarray experiment is a two-dimensional (2D) matrix with genes as rows and samples as columns (usually coming from different conditions). Each cell in the matrix is a real number that indicates how much a gene is expressed in a sample. These expression matrices usually have thousands of rows and tens or hundreds of columns (Feltes *et al*., 2019).

Due to high availability, Microarray data have become one of the largest sources of large-scale transcriptomic biological data, boosting bioinformatics studies and increasing knowledge of biological functions and diseases (Shi *et al*., 2017). However, despite the diversity of Microarray studies, the continuous improvement of sequencing technologies and the wide offer of transcriptomic analysis tools, the identification of expression patterns is still a great challenge (Walsh, C., Hu, P., Batt, J., *et al*., 2015), especially in diseases such as cancer. According to the World Health Organization (WHO), in 2019, cancer was the sixth leading cause of death worldwide (https://www.who.int/news-room/fact-sheets/detail/the-top -10-causes-of-death) and understanding the molecular pathways underlying the tumor is a challenge to be overcome, mainly due to the heterogeneity of its nature (Ho, J., *et al*., 2018). Machine Learning (ML) is a field of Computational Intelligence that is concerned with building computer programs that automatically improve themselves with experience. Over the past few decades, many successful Machine Learning applications have been developed: from data mining programs that learn to detect fraudulent credit card transactions, information filtering systems that learn users' reading preferences, to self-driven vehicles, that learn to drive on highways without the participation of a driver (Mitchel, T.M, 1997; Lecun *et al*., 2015). Among the various techniques available for analyzing DNA Microarray data, ML has been widely used for gene selection and expression dataset classification, as well as information discovery. In addition, cancer data have become a frequently used source to test new ML algorithms (Feltes *et al*., 2019). The popularization of industrialized Microarray chips dates back to 1995 (Schena, M., *et al*., 1995), but the application of ML for these purposes began in 1999 when Golub *et al*. (Siegel *et al*., 2016) designed a class discovery procedure for leukemia. Alon *et al*. (Alon, U., *et al*., 1999) used a clustering algorithm to analyze tumors and normal colon tissues. Microarray data were used to train classifier algorithms able to predict different conditions and help with diagnoses (Peterson, L.E, *et al*., 2005; Dıaz-Uriarte, R., and De Andres, S.A., 2006; Pirooznia, M. *et al*., 2008; Statnikov, A., Wang, L., and Aliferis, C.F., 2008). By grouping samples autonomously by the expression of their genes according to some similarity criteria, the grouping methods helped in the discovery of knowledge and in the biological inference about that set of genes or samples (Whitworth, G.B., 2010). The review by Thalamuthu (Thalamuthu *et al*., 2006) and the case study by Dash & Misra (Dash, R., and Misra, B.B., 2018) compare some of these Machine Learning methods in Microarray analysis. The work by Oyelade (Oyelade *et al*., 2016) also provides descriptions of grouping methods and insights on how best to choose and use them for Microarray data. The use of feature extraction and selection methods in gene expression data is also common for dimensionality reduction and data visualization, as a pre-processing step for other algorithms or to find a more relevant subset of genes. (Lazar *et al*., 2012; Ang *et al*., 2016) provide extensive reviews on the subject. This study used the gene expression data organized by Feltes *et al*. (Feltes *et al*., 2019) and applied different computational Machine Learning models to establish the diagnosis of the leukemia type (ALL or AML) from the gene expression data of the studied patients. Three intelligent approaches were used. Initially, a Support Vector Machine (SVM) (M. A. Hearst *et al*., 1998; Cortes, C. and Vapnik, V., 1995; Bernhard Schölkopf *et al*., 2000) and an Artificial Neural Network (ANN) (Wang SC, 2003; Kubat, M., 1994; Simon Haykin, 2000) were used to classify the data. Subsequently, a combination of computational intelligence algorithms, an approach known as Machine Learning Ensemble, was used to further improve the performance in classifying these two types of acute leukemia. Ensemble Learning refers to the procedures used to train various Machine Learning models and combine their results, treating them as a "committee" of decision makers. The principle is that the committee's decision, with the individual forecasts properly combined, should achieve better overall accuracy, on average, than any individual committee member. Numerous empirical and theoretical studies have shown that cluster models often achieve greater precision than individual models (Brown G., 2011). The results are presented and discussed in the next sections.

METHODOLOGY

The data used in this work was extracted from the Curated Microarray Database (CuMiDa) (Feltes *et al*., 2019), a repository containing 78 selected extensively cross-checked cancer Microarray datasets from 30,000 Gene Expression Omnibus (GEO) studies. CuMiDa offers a newer dataset, manually and carefully selected, with sample quality, unwanted probe extraction and background correction and normalization to create a more reliable data source. These data are available at: https://sbcb.inf.ufrgs.br/cumida and contains the training and test sets used in the work by Golub *et al*. (Golub *et al*., 1999). These datasets contain measurements corresponding to ALL and AML samples of bone marrow and peripheral blood. The intensity values have been resized so that the overall intensities of each chip are equivalent.

The dataset contains three files:

- ***actual.csv*** contains the identification of all 72 patients in the study and their labels (type of cancer, 47 ALL and 25 AML).
- ***data_set_ALL_AML_train.csv*** contains the subset with training data (38 bone marrow samples, 7129 genes).
- ***data_set_ALL_AML_independent.csv***contains the subset with the test data (34 peripheral blood samples, 7129 genes).

Figure 1. Original dataset

The training and test files contain the information for: Gene Description, Gene Accession and Raw Expression Values. Figure 1 shows a cutout of the three files in the dataset. Some pre-processing procedures were performed before applying ML algorithms for data classification. In the actual.csv file the "ALL" and "AML" labels were converted to numeric (0 and 1). Then, headers were removed and data (columns) that would not be used (such as "call") were eliminated. Subsequently, the files (rows and columns) were transposed and then the training data was normalized to the same scale as the test data. Finally, the PCA (Principal Component Analysis) technique (Rasmus B., and Smilde, A.K., 2014) was applied to identify the most important variables in the prediction from the measure of variance, thus reducing the dimensionality. The next step was the choice of ML algorithms to classify the database. As this is a classification problem, which characterizes the type of learning as supervised, part of the data, called labeled, is used to train the model and the other part (not labeled) is used for testing. The initial approach was to use the classifiers individually. The first ML algorithm chosen was a Support Vector Machine (SVM), implemented in the Python programming language (https://www.python.org/), using the Scikit-learn library (Pedregosa, F, *et al*., 2011) in the cloud programming environment Collab - Google Collaborative (https://colab.research.google.com/). Figure 2 shows the SVM parameters used.

```
SVC(C=1.0, break_ties=False, cache_size=200, class_weight=None, coef0=0.0,
    decision_function_shape='ovr', degree=3, gamma='scale', kernel='rbf',
    max_iter=-1, probability=False, random_state=None, shrinking=True,
    tol=0.001, verbose=False)
```

Figure 2. SVM parameters used to classify

Then, an Artificial Neural Network (ANN) of the Multilayer Perceptron (MLP) type was used to classify the gene expression data. The algorithm was implemented in the Python programming language, using the Keras library (https://keras.io/) in the Colab - Google Collaborative cloud programming environment. Table 1 shows the parameters of the ANN used.

Table 1. Used parameters of the ANN

Parameter	Values	
Input layer	*32 neurons*	*ReLU activation*
Hidden layer	*16 neurons*	*ReLU activation*
Output layer	*1 neuron*	*Sigmoide activation*
Loss function	*Binary Crossentropy*	
Optimizer	*Adam*	
Metrics	*Binary Accuracy*	
Batch size	*8*	
Epochs	*50*	

The Machine Learning Ensemble technique is used to combine more than two algorithms to produce the best learning model (Ahmad, W.D. and Bakar, A.A., 2020). This approach has two main goals: the first one is to increase the accuracy of the overall predictions compared to a single classifier, and the second one is to improve the generalization rate because of its specific measures.

As a result, the final classifier can resolve unsolved issues with a single predictive model. The performance of the model on examples not seen during training demonstrates the real capabilities of the model (Schapire, 2003). In order to obtain greater accuracy of classification, gene expression data from patients with leukemia were submitted to a Machine Learning Ensemble. A Machine Learning Ensemble model was then built using a 2-layer stack. The first layer is called Base-Learning and the algorithms used are all classifiers. The second layer is called Meta-Learning and is used to combine the results of all the first layer algorithms.

The steps involved in this process are described below:

Step 1. Dividing data into training sets and test sets;

Step2. Data pre-processing (training data augmentation);

Step 3. Submit training data to Base-Learning layer classifiers;

Step 4. Aggregate the result generated by each Base-Learning algorithm in the Meta-Learning layer;

Step 5. Submit test data to trained Ensemble to get final results.

The model was built with the Orange framework (https://orangedatamining.com/docs/), using the Stack method, which performs Ensemble Learning by stacking several algorithms to make the individual data classification. Then, the individual results from the classifier stack are used as input to a metamodel that aggregates these results. The classifier ensemble used for Base-Learning has Multilayer Perceptron, Support Vector Machine, Random Forest (Breiman, 2001), Gradient Boosting (Friedman, 2001) and k-NN (Farooq, M. *et al*., 2021). The Meta-Learning step used Logistic Regression. Before submitting the data to the Ensemble classifier model, a treatment step on the training data had to be performed. A training data augmentation was performed with the Create Instance Median/Mean method. This mechanism creates new instances for each class value in the training set, based on the mean and median of the attributes. This process generated new synthetic data from existing data, acting as a regularizer and improving the accuracy and precision of the model. With the data fitted to the model, the training set was submitted to Ensemble. Finally, the test data was submitted to the already trained Ensemble. Figure 3 shows the model diagram.

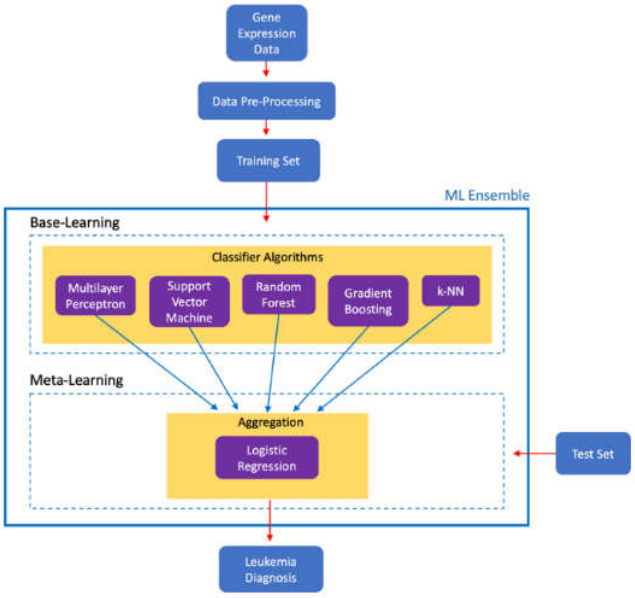

Figure 3. Proposed model diagram

The Classifier Algorithm Ensemble ran on a 2.2GHz Intel Core i7 – 6 Core processor, with 16GB of RAM. Table 2 shows the Ensemble parameters used in this study.

RESULTS AND DISCUSSION

The results of isolated application of classifiers for the diagnosis of leukemia are shown below. The SVM achieved an accuracy of 67% and we can see its confusion matrix in Figure 4a.

Table 2. Machine learn Ensamble parameters

Algorithm	Parameter	Values
Multilayer Perceptron	*Neurons in hidden layer*	*100*
	Activation function	*ReLU*
	Optimizer	*Adam*
Support Vector Machine	*Cost*	*1*
	Regression loss (ε)	*0,1*
	Kernel	*RBF*
Random Forest	*Number of trees*	*100*
	Attributes by break	*5*
Gradient Boosting	*Number of trees*	*100*
	Learning rate	*0,1*
k-NN	*Number of neighbors (k)*	*5*
	Metric	*Euclidean*
	Weight	*Uniform*

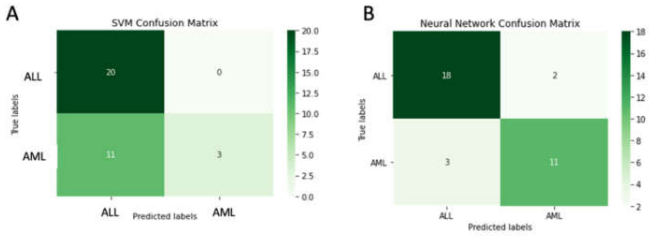

Figure 4. A) Confusion matrix for the SVM. B) Confusion matrix for the MLP

The MLP-type artificial neural network had a prediction hit rate of 85% of the cases and we can see the classification errors in Figure 4b. The results of applying the Ensemble stacked classifiers for the diagnosis of leukemia are shown below. Among the various existing metrics to assess the performance of ML algorithms in the classification task, we chose a few to measure the result, either individually or in the final aggregation of results. The metrics are AUC (Area Under ROC - Receiver Operating Characteristic), CA (Classification Accuracy), F1 (F-score), Precision and Recall.

Table 3 shows the performance measures (individual and aggregated) of the algorithms. When all results are analyzed, it is clear that the first SVM isolated implementationhad the worst result, with an accuracy of 67%. For this reason, the MLP neural network was implemented for the diagnosis of leukemia, which obtained a better result, reaching 85%, however, this could still be improved. We then proceeded to the Ensemble approach, where each classifier algorithm performs its task and, in the end, the results of all of them are aggregated. This stack of programs is trained with a set of data and later tested for other untrained data. Orange allows to evaluate the performance of each algorithm individually and then the overall performance of the Ensemble model, with different types of performance measurements. Some of these numbers are worth mentioning. Considering all five performance metrics, SVM had the worst individual result, similar to what was achieved in the previous isolated implementation. The MLP neural network, although having achieved an excellent AUC result, maintained the same previous level for the other 4 metrics observed. Random Forest had a perfect AUC result, classified without any error, however, the same performance was not observed in the other 4 performance measures. Gradient Boosting was the Ensemble algorithm that achieved the best individual result, above 90% for all metrics. The nearest neighbor's method (k-NN) had a very good AUC result, however, in the other 4 performance measures, this hit level decreased and was below 90. Finally, the overall result, already aggregated, of the Ensemble, as expected, achieved the best overall performance when compared to the individual results of the algorithms, remaining above 94% accuracy for all observed measures. Figure 5 presents the confusion matrices with the classification errors of the five classifier algorithms, as well as the Ensemble's aggregate result. Class 0 represents ALL and class 1 represents AML and correctly sorted instances are on the matrices's main diagonal. The gene expression data obtained from Microarray experiments are usually organized in matrices of n rows and m columns, called the gene expression profile. Rows represent genes and columns represent samples or their features. For illustration purposes, it is possible to carry out some simple studies on this data, for example, comparing the genes (n rows) or comparing the samples (m columns) of the matrix.

Table 3. Machine Learning Ensemble Performance Evaluation

Algorithm	AUC (%)	CA (%)	F1 (%)	Precision (%)	Recall (%)
MLP	*95,7*	*85,2*	*85,3*	*85,5*	*85,2*
SVM	*92,1*	*64,7*	*55,5*	*77,9*	*64,7*
Random Forest	*100*	*82,3*	*81,0*	*86,4*	*82,3*
Gradient Boosting	*91,4*	*91,1*	*91,2*	*91,4*	*91,1*
k-NN	*96,7*	*85,2*	*84,5*	*88,2*	*85,2*
Stack (Ensemble)	*96,7*	*94,1*	*94,0*	*94,6*	*94,1*

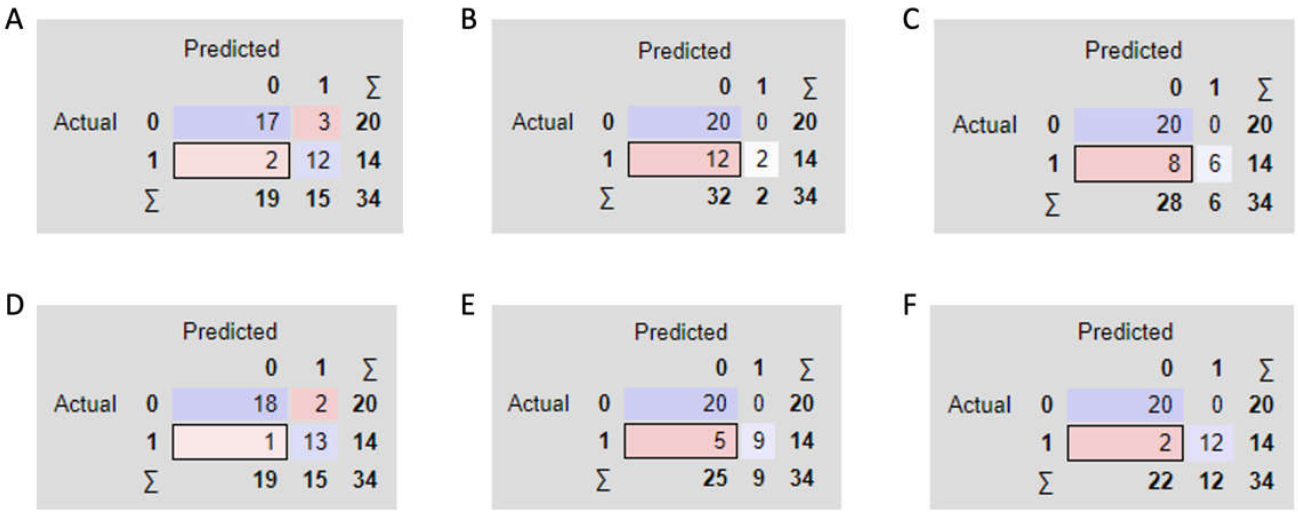

Figure 5. A) Confusion Matrix of the *MLP*. B) Confusion Matrix *SVM*. C) Confusion Matrix of the Random Forest. D) Confusion Matrix of the Gradient Boosting. E) Confusion Matrix of the k-NN. F) Confusion Matrix of the ML Ensemble

If two lines are found to be similar, it can be assumed that the two genes are co-regulated and possibly functionally related. These analyzes can facilitate the understanding of gene regulation; metabolic and signaling pathways; the genetic mechanisms of disease; and the response to drug treatments. However, considering the amount and complexity of the gene expression data, it is impossible for a human expert to calculate and compare the n x m gene expression matrix manually, as n is generally greater than 5,000 and m is greater than 10. In the case of our data, the matrix has 7,129 rows and 78 columns. Thus, Machine Learning techniques turn out to be very useful and effective, and have been widely applied to classify and characterize gene expression data. This is due to the nature of the Machine Learning approach, which manages to perform well in domains where there is a large amount of data and little theory or explanation about it, and this is exactly the case for the analysis of gene expression profiles. The Machine Learning Ensemble has been an active research topic in artificial intelligence, but it is still relatively new in bioinformatics applications. Constructing a discriminatory classifiercan be seen as finding (approaching) the true hypothesis from all possible hypothesis space. Each individual learning algorithm uses a different search strategy to identify the true hypothesis. If the training sample size is very small, which is the case when classifying Microarray data, the individual classifier can induce different hypotheses with similar performance from the search space. Thus, by averaging the different hypotheses, the combined classifier (Ensemble) can produce a good approximation of the true hypothesis. The computational motivation for this is to try to avoid the great locations of individual search strategies. The final classifier can provide a better approximation of the true hypothesis by performing different initial searches and combining the outputs. Finally, due to the limited amount of training data, an individual classifier may not be able to represent the true hypothesis. Thus, from the various base classifiers, it might be possible for the final classifier to approximate the true hypothesis. All data and codes are available at https://github.com/npca-ufra/leucemiaensemble.

CONCLUSION

The use of ML approaches to infer gene expression information from Microarray data has increased in recent years, especially in cancer-related work. The classification of cancer based on gene expression profile remains a challenging task, whether in identifying potential sources of therapeutic intervention, understanding the behavior of the tumor, or facilitating drug development. This study applied some Machine Learning techniques to classify two types of leukemia (ALL and AML), based on their gene expression data. First, two supervised learning algorithms, a support vector machine and an artificial neural network were used to make this diagnosis. Subsequently, to improve the overall classification performance, a ML Ensemble model was built, where several intelligent algorithms are combined, thus increasing the learning capacity and classification power. The following algorithms were combined: Artificial Neural Network, Support Vector Machine, Random Forest, Gradient Boosting and k-NN. The Ensemble model achieved a consistent overall performance (above 94% for five different learning algorithm evaluation metrics) in classifying the ALL and AML leukemia types from the gene expression data. The proposed model proved to be useful for Microarray data classification tasks and can be applied in the diagnosis of other types of cancer where sufficient gene expression data are available.

ACKNOWLEDGMENTS

The authors would like to thank to Universidade Federal Rural da Amazônia (UFRA), through Pró-Reitoria de Pesquisa e Desenvolvimento Tecnológico (PROPED), for the financial support to this study. This work is part of the Scientific Initiation Program PROGRIDI/PIBITI/UFRA (Edital PROPED 08/2020 and 09/2020). The authors also thank to PROCAD Amazônia 2018 (Edital 88887.200562/2018-00) for the financial support.

REFERENCES

Ahmad, W.D. & Bakar, A.A. 2020. *Ensemble Machine Learning Model for Higher Learning Scholarship Award Decisions*. International Journal of Advanced Computer Science and Applications, Vol. 11, No. 5, 2020

Alon, U., Barkai, N., Notterman, D.A., *et al*. 1999. *Broad patterns of gene expression revealed by clustering analysis of tumor and normal colon tissues probed by oligonucleotide arrays*. Proc. Natl. Acad. Sci. 96, 6745–6750.

Ang, J.C., Mirzal, A., Haron, H., *et al*. 2016. *Supervised, unsupervised, and semi-supervised feature selection: A review on gene selection*. IEEE/ACM Trans. Comput. Biol. Bioinformatics 13, 971–989.

Benjamin L. Ebert, Todd R. Golub. 2004. *Genomic approaches to hematologic malignancies*. Blood 2004; 104 4: 923–932. doi: https://doi.org/10.1182/blood-2004-01-0274.

Bernhard Schölkopf, Alex J. Smola, Robert C. Williamson, Peter L. Bartlett. 2000. *New Support Vector Algorithms*. *Neural Comput* 2000; 12 5: 1207–1245.

Bishop, JF. *Adult acute myeloid leukemia: update on treatment*. The Medical Journal of Australia. 1999 Jan;1701:39-43. DOI: 10.5694/j.1326-5377.1999.tb126866.x. PMID: 10026673.

Breiman, L. 2001. Random Forests. Machine Learning 45, 5–32. https://doi.org/10.1023/A:1010933404324.

Brown G. 2011. Ensemble Learning. In: Encyclopedia of Machine Learning. Sammut C., Webb G.I. eds. Springer, Boston, MA. https://doi.org/10.1007/978-0-387-30164-8_252

Bullinger L. & Valk, PJ. 2005. *Gene Expression Profiling in Acute Myeloid Leukemia.* J Clin Oncol 23:6296–6305.

Cortes, C., Vapnik, V. 1995. *Support-Vector Networks*. *Machine Learning* 20, 273–297 1995. https://doi.org/10.1023/A:1022627411411.

Dash, R., and Misra, B.B. 2018. *Performance analysis of clustering techniques over microarray data: A case study*. Phys. A Stat. Mech. Appl. 493:162–176.

Dıaz-Uriarte, R., and De Andres, S.A. 2006. *Gene selection and classification ofmicroarray data using random forest*.BMC Bioinformatics 7, 3.

doi: https://doi.org/10.1162/089976600300015565. 2000.

Farooq, M., Sarfraz, S., Chesneau, C., Ul Hassan, M., Raza, M.A., Sherwani, R.A.K., & Jamal, F. 2001. *Computing Expectiles Using k-Nearest Neighbours Approach*. Symmetry 2021, 13, 645. https://doi.org/10.3390/sym13040645.

Feltes, B.C., Chandelier, E.B., Grisci, B.I., Dorn, M. 2019.*CuMiDa: An Extensively Curated Microarray Database for Benchmarking and Testing of Machine Learning Approaches in Cancer Research*. Journal of Computational Biology, 2019.DOI: 10.1089/cmb.2018.0238.[Online] SBCB. Available at: http://sbcb.inf.ufrgs.br/cumida.

Friedman, J.H. 2001. *Greedy function approximation: a gradient boosting machine*. Annals of statistics, pages 1189–1232.

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. 1999. *Molecular Classification of Cancer: Class Discovery and Class Prediction by Gene Expression Monitoring*. Science 286 5439, 531-537. DOI: 10.1126/science.286.5439.531.

Haferlach, T., Kohlmann A., Bacher, U., Schnittger, S., Haferlach, C. & Kern, W. 2007. *Gene expression profiling for the diagnosis of acute leukaemia*. Br J Cancer 96:535–540.

Hearst, M. A. S. T. Dumais, E. Osuna, J. Platt and B. Scholkopf. 1998. *Support vector machines.* in *IEEE Intelligent Systems and their Applications*, vol. 13, no. 4, pp. 18-28, July-Aug. 1998, doi: 10.1109/5254.708428.

Ho, J., Li, X., Zhang, L., *et al*. 2018. *Translational genomics in pancreatic ductal adenocarcinoma: A review with reanalysis of tcga dataset*. Semin. Cancer Biol. DOI: 10.1016/j.semcancer. 2018.04.004.

Kubat, M. *Neural networks: A comprehensive foundation.* Simon Haykin, Macmillan, 1994, ISBN 0-02-352781-7. *The Knowledge*

Engineering Review, 134, 409-412. doi:10.1017/ S0269888998214044. 1994.

Lazar, C., Taminau, J., Meganck, S., *et al*. 2012. *A survey on filter techniques for feature selection in gene expression microarray analysis*. IEEE/ACM Trans. Comput. Biol. Bioinf. 9, 1106–1119.

Lecun, Y., Bengio, Y. & Hinton, G. 2015. *Deep Learning*, 436, NATURE, Vol. 521.

Löwenberg B, Downing JR. & Burnett A. 1999. *Acute myeloid leukemia.* N Engl J Med. 1999 Sep 30;34114:1051-62. Erratum in: N Engl J Med 1999 Nov 4;34119:1484. PMID: 10502596.

Masilamani, V., Devanesan, S., AlSalhi, MS., AlQahtany, FS. & Farhat, KH. 2020. *Fluorescence spectral detection of acute lymphoblastic leukemia ALL and acute myeloid leukemia AML: A novel photodiagnosis strategy, Photodiagnosis and Photodynamic Therapy*. Volume 29, 2020, 101634, ISSN 1572-1000, https://doi.org/10.1016/j.pdpdt.2019.101634.

Mi, S., Lu, J., Sun, M., Li, Z., Zhang, H., Neilly, M. B., Chen, J. 2007. *MicroRNA expression signatures accurately discriminate acute lymphoblastic leukemia from acute myeloid leukemia*. Proceedings of the National Academy of Sciences, 10450, 19971–19976. doi:10.1073/pnas.0709313104.

Mitchel, T.M. 1997. *Machine Learning*. McGraw-Hill International Editions.

Oyelade, J., Isewon, I., Oladipupo, F., *et al*. 2016. *Clustering algorithms: Their application to gene expression data*. Bioinf. Biol. Insights 10:237–253.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., & Thirion, B. 2011. *Scikit-learn: Machine Learning in Python*, Journal of Machine Learning Research 12, pp. 2825-2830.

Peterson, L.E., Ozen, M., Erdem, H., *et al*. 2005. *Artificial neural network analysis of dna microarray-based prostate cancer recurrence*. 1–8. In Proceedings of the 2005 IEEE Symposium on Computational Intelligence in Bioinformatics and Computational Biology, 2005, CIBCB'05. IEEE. Nov. 14–15, La Jolla, CA..

Pirooznia, M., Yang, J.Y., Yang, M.Q., *et al*. 2008. *A comparative study of different machine learning methods on microarray gene expression data*. BMC Genomics 9, S13.

Pui, CH, Evans, WE. *Acute lymphoblastic leukemia*. N Engl J Med. 1998 Aug 27;3399:605-15. doi: 10.1056/NEJM1998082 73390907. PMID: 9718381.

Pui, CH; Schrappe, M; Ribeiro, RC & Niemeyer. CM. 2004. *Childhood and adolescent lymphoid and myeloid leukemia*. Hematology Am Soc Hematol Educ Program. 2004:118-45. doi: 10.1182/asheducation-2004.1.118. PMID: 15561680.

Randolph, T.R. 2004. *Advances in Acute Lymphoblastic Leukemia.* American Society for Clinical Laboratory Science October 2004, 17 4 235-245.

Rasmus B., & Smilde, A.K. 2014. *Principal component analysis*. Anal. Methods,2014,6,2812.

Rowley, JD. 2000. *Molecular genetics in acute leukemia*. In: Leukemia. 14:513–517. Macmillan Publishers Ltd.

Schapire, R.E. 2003. *The boosting approach to machine learning: an overview*. Nonlinear Estim Classif [Internet]. 2003;171:149–71.

Schena, M., Shalon, D., Davis, R.W., *et al*. 1995. *Quantitative monitoring of gene expression patterns with a complementary DNA microarray*. Science, 270, 467–470.

Siegel, R.L., Miller, K.D., & Jemal, A. 2016. *Cancer statistics*. CA Cancer J. Clin. 66. 7–30.

Simon Haykin. *Neural Networks: A Guided Tour*. Soft Computing and Intelligent Systems: Theory and Applications. Elsevier. 2000.

Statnikov, A., Wang, L., and Aliferis, C.F. 2008. *A comprehensive comparison of random forests and support vectormachines for microarray-based cancer classification*. BMC Bioinformatics 9, 319.

Stone, RM, Mayer, RJ. *Treatment of the newly diagnosed adult with de novo acute myeloid leukemia*. Hematol Oncol Clin North Am. 1993 Feb;71:47-64. PMID: 8449864.

Taylor, J., Xiao, W., & Abdel-Wahab, O. 2017. *Diagnosis and classification of hematologic malignancies on the basis of genetics*. Blood, 130. 410–423.

Thalamuthu, A., Mukhopadhyay, I., Zheng, X., *et al*. 2006. *Evaluation and comparison of gene clustering methods in microarray analysis*. Bioinformatics 22, 2405–2412.

Walsh, C., Hu, P., Batt, J., *et al*. 2015. *Microarray meta-analysis and cross-platform normalization: Integrative genomics for robust biomarker discovery*. Microarrays Basel 4, 389–406.

Wang SC. *Artificial Neural Network*. In: Interdisciplinary Computing in Java Programming. The Springer International Series in Engineering and Computer Science, vol 743. Springer, Boston, MA. https://doi.org/10.1007/978-1-4615-0377-4_5. 2003.

Whitworth, G.B. 2010. *An introduction to microarray data analysis and visualization*. 19–50. In Methods in Enzymology, volume 470. Elsevier. San Francisco, CA.

Shi, A., Li, R., *et al*. 2017. *Microarray bioinformatics in cancer—A review*. J. BUON. 22, 838–843.

****