IJARSCT

ISSN (Online) 2581-9429

International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)

International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal

Impact Factor: 7.301

Volume 3, Issue 2, September 2023

Comparative Analysis of Machine Learning and Deep Learning Algorithms to Classify Cancer Based on Microarray Genes

Shefali Parihar[1] and Dr. Kalpana Sharma[2]

Research Scholar, Department of CSE, Bhagwant University, Ajmer, Rajasthan[1]

Assistant Professor, Department of CSE, Bhagwant University, Ajmer, Rajasthan[2]

Abstract*:* *Cancer classification is a topic of major interest in medicine because it allows accurate and efficient diagnosis and facilitates a successful outcome in medical treatment. previous studies using large-scale RNA profiling and supervised machine learning (ML) algorithms to construct molecular-based classifications of carcinoma cells from breast, bladder, adenocarcinoma, colorectal, gastro esophagus, kidney, liver, lung, ovarian Human tumors have been classified. , pancreas, and prostate tumors. These datasets are collectively known as the 11 Tumor Database, although this database has been used in many works in the ML field, no comparative study of different algorithms could be found in the literature. On the other hand, advances in both hardware and software technologies have led to substantial improvements in the accuracy of solutions using ML, such as Deep Learning (DL). In this study, we compare the most widely used algorithms in classical ML and DL to classify tumors described in the 11 tumor database.*

Keywords: Dataset, Database, Algorithm, ML, DL, Cancer etc

I. INTRODUCTION

Cancer describes a class of diseases in which genetic mutations cause malignant cells to form inside the human body. As these cells grow, they divide indiscriminately, spread throughout the organs and, in some cases, can result in loss of life. Cancer is the second leading cause of mortality globally after cardiovascular diseases [1]. Recently, gene expression analysis has emerged as an important tool for addressing fundamental challenges associated with cancer diagnosis and drug discovery [2,3]. Gene expression analysis also provides information about the contribution of different genes to cancer initiation and progression. Consequently, changes in gene expression can be used as markers for the early detection of cancer and to identify targets for drug development. Such approaches may open up the possibility of health care that is more personalized, preventive, and predictive [4].

Gene expression is the process by which information in DNA is converted into instructions for making a protein or other molecule. It involves transcription of DNA into messenger RNA (mRNA), followed by translation into proteins. Gene expression analysis is used to assess the sequence of genetic changes that occur under certain conditions in a tissue or a single cell [5]. This involves measuring the number of DNA transcripts present in sampled tissue or cells in order to obtain information about which genes are expressed and at what levels. One component of gene expression quantification is the comparison of sequenced reads corresponding to the number of base pairs sequenced from a DNA fragment to a recognized genomic or transcriptome source. The accuracy of quantification depends on sequenced reads containing sufficient specific information to allow bioinformatics algorithms to correlate reads with appropriate genes. Prevailing methods for assessing gene expression include DNA microarray and next-generation sequencing (NGS) methods. The DNA microarray method uses a two-dimensional array with microscopic spots, in which short sequences or genes bind to known DNA molecules through a hybridization process. NGS methods of massively parallel sequencing provide exceptionally high-throughput analysis, scalability, and speed, and they have been used to determine the nucleotide sequence of complete genomes, or single DNA or RNA segments [6,7]. Is. RNA-sequencing, also known as RNA-seq, is an NGS method that involves converting RNA molecules into complementary DNA (cDNA) and determining the sequence of nucleotides in the cDNA for gene expression analysis and quantification. Is. Compared to DNA microarrays, RNA-Seq [8,9] offers several advantages, including greater specificity and resolution, increased sensitivity to differential expression, and greater dynamic range. RNA-Seq can also be used to examine the transcriptome of any species to quantify RNA at a specific point in time.

II. LITERATURE REVIEW

Salem et al. [9] suggested genetic programming-based cancer classification with the help of information gain (IG) for feature selection. Lin et al. [10] introduced genetic algorithms with silhouette statistics for feature selection and classification on the SRBCT dataset. We have seen that the feature selection method is non-optimal as it generates thousands of features resulting in over-fitting of the model. Sharbaf et al. [11] proposed a hybrid approach for gene selection and classification of microarray datasets using cellular learning automata and ant colony optimization. They have investigated the effect of various rank-based feature selection methods and they use three classifiers for validation, namely Support Vector Machine (SVM), K-Nearest Neighbors (KNN), and Naive Bayes. Kumar et al. [12] built a feature selection and classification algorithm based on the MapReduce concept with a KNN classifier. Nguyen et al. [13] proposed a holistic gene selection for microarray data classification and applied their model on four standard datasets namely DLBCL, leukemia, prostate and colon datasets. To validate the method, five existing classifiers, namely, Linear Discriminant Analysis, KNN, Probabilistic Neural Network, SVM, and Multilayer Perceptron (MLP) were used and they claimed that the proposed method has stability across different classifiers, but They could not manifest the claimed stability beyond this. Five classifiers. Lofty and Keschwarz [14] introduced a hybrid of Principal Component Analysis (PCA) and Brain Emotion Learning for microarray cancer data classification. They validated the work on three datasets which are not sufficient to confirm about the generalizability of the method. Ravi et al. [26] have done a comprehensive review to reveal the potential of deep learning models in health informatics. He has illustrated various deep learning architectures such as deep feed-forward, convolutional networks and recurrent networks applied to problems in several problem areas. Kar et al. [15] proposed particle swarm optimization-based feature selection for classification of microarray cancer data. The acute lymphoblastic leukemia–acute myeloid leukemia (ALL-AML) and SRBCT datasets were used to validate the method. They perform an experiment ten times on each dataset and the average of these ten runs is reported as the final result.

Research Methodology

ML and DL techniques can learn features of a given problem from a certain amount of data. These data are usually randomly divided into two groups: training and validation. A training dataset is used to calibrate the parameters of the model, and a validation dataset is used to evaluate the performance of the model (Eraslan et al., 2019). We have also included Decision Tree-Methods (DT) such as Random Forest (RF). Unlike linear models, DT and RF are invariant to data scaling and work well with features at different scales. Finally, we applied deep neural networks (DNNs), such as fully connected neural networks, also known as multi-layer perceptron (MLP) and convolutional neural networks (CNN). MLPs are suitable for non-linear data, while CNNs automate the expensive work of engineering facilities; An indispensable task in the classical ML approach. The above algorithm is explained extensively in Michie, Spiegelhalter, and Taylor (1994) and Chollet (2007). The datasets used represent measurements of gene expression using cancer microarrays and normal biopsies (Statnikov et al., 2005; Bolon-Canedo et al., 2014), and are consolidated into "11 tumor databases", Which is available online for free.
https://github.com/simonorozcoarias/ML_DL_microArrays/blob/master/data11tumors2.csv). This database contains 174 samples with 12,533 gene expression microarrays for 11 different cancer types.

For the experiments, we divided the information into two groups; The first group corresponds to features (X) and the second group corresponds to classes (Y). The features form a matrix of size m × n and the classes are a vector of size n × 1, where m is the number of samples and n is the number of genes for each class (12,533). The dataset, which contains 174 samples, is randomly divided into two subgroups (80% training and 20% validation), consisting of 139 samples for training and 35 samples for validation. Initial calibration of the ML and DL algorithms (training) was done using the training set; Then, hyperparameter tuning was performed with the validation set and the accuracy of the algorithms was measured.

III. RESULT ANALYSIS

Table 1.1: Classification of cancers in 11 tumor databases.

Class	Cancer type	Number of patients
0	Ovary	29
1	Bladder/Ureter	10
2	Breast	28
3	Colorectal	25
4	Gastroesophagus	14
5	Kidney	13
6	Liver	9
7	Prostate	28
8	Pancreas	8
9	Adenocarcinoma	16
10	Lung squamous cell carcinoma	16

Four different datasets were created for training and validation of each ML or DL algorithm. For the first dataset, we did not apply any preprocessing operations; For the second, we performed a scaling process; For the third, we applied PCA with 96% of the variance retained to reduce the data dimensionality, achieving a dimensionality reduction from 12,655 to 85 features. Finally, for the final dataset, we applied both scaling and PCA, achieving dimensionality reduction from 12,655 to 115 features (principal components). We evaluated the performance of well-known ML classification algorithms including KNN, SVC, LR, LDA, NB, MLP, RF and DT. Next, we evaluated DL architectures, such as Fully Connected Neural Networks (FNN) and Convolutional Neural Networks (CNN). Two types of networks were used for DL; The first is a fully connected neural network and the second is a convolutional neural network. The FNN consists of three fully connected layers of 100 neurons each and a SoftSign activation function; then, a final layer of 11 neurons with sigmoid activation function are generated to generate cancer type probability.

Hierarchical analysis

Before evaluating the classification algorithms, we visualized internal clusters in the data and determined how these clusters are affected by various preprocessing methods applied to our data. Using the downloaded raw data, we built a hierarchical graph (unsupervised learning) using different methods and concluded that Ward's method produced the most balanced clusters. Then, using only Ward's method, we conducted additional analyzes using different datasets, including raw data, scaled data, data transformed by PCA, and data scaled and transformed by PCA. Finally, we created a dendrogram and a heatmap to find out if the data could be clustered into groups without any classes with the best results. Four well-separated groups, but the heatmap displayed other well-conserved groups, which may indicate that the four main groups can be divided into subgroups.

Based on a priori knowledge that the number of cancer types is eleven (11), we were interested in determining how the hierarchical clustering algorithm produced cluster assignments. Therefore, we applied the best parameters found previously (clustering method: Ward, and input: raw data and data reduced by PCA).

Table 1.2: Cluster structure and original number of individuals from each class of cancer.

Class	Original number	Clustering using raw data	Clustering using data processed by PCA
0	29	49	49
1	10	31	30
2	28	18	41
3	25	6	6
4	14	33	27
5	13	27	12

6	9	8	8
7	28	2	3
8	8	6	6
9	16	4	3
10	16	11	11

In this work, we show the application of unsupervised and supervised learning approaches of ML and DL for the classification of 11 cancer types based on microarray datasets. We observed that the best average results are obtained using the raw dataset and the LR algorithm using the training and validation data, leading to an accuracy value of 100% (validation set, using the hold-out split method). One can assume that overfitting has occurred because the confusion matrix has shown extremely well-behaved behavior; However, a comparison of training and validation accuracies between parameters using the entire dataset may indicate the true accuracy in both the training and validation datasets. Additional tests should be performed with independent data to rule out potential overfitting.

On the other hand, MLP and LDA showed a high accuracy value of 97.14% in the validation dataset. This improvement in accuracy was achieved by optimizing several parameters (number of neurons in the MLP) and preprocessing the dataset with PCA.

After tuning the four parameters, RF achieved high results with a maximum accuracy of 85.71%. In contrast, DT achieved 51.14% accuracy, indicating that despite tuning several parameters (in our case, three), DT does not work properly for the dataset used in this study.

Table 1.3: Best value of tuned hyperparameters in deep neural networks

Parameter	Best value	
	FNN	CNN
Batch size	20	10
Epochs	100	10
Training optimization algorithm	Adagrad	SGD
Learn rate	0.2	0.1
Momentum	0	0
Network weight initialization	Normal	Glorot_normal
Neuron activation function	Softsign	Linear
Weight constraint	3	1
Dropout regularization	0	0.4

IV. CONCLUSION

Cancer is predicted to become the deadliest disease for humans in the future (Degenais et al., 2019); Therefore, early diagnosis, recognition and treatment are needed to control the disease. ML and DL techniques are promising tools for classification of cancer types using complex datasets such as microarrays. In this study, we obtained predictions with 93.52% and 94.46% accuracy, which will allow patients with this type of pathology to quickly and accurately detect their disease and will also contribute to the discovery of new selective drugs. Treatment of this type of tumor. In our work, we propose a deep feed-forward neural network approach for classification of binary class microarray datasets. To validate the proposed method, eight standard microarray cancer datasets namely CNS, colon, prostate, leukemia, ovarian, lung-Harvard 2, lung-Michigan and breast cancer are used. To overcome the curse of dimensionality and other problems associated with the nature of data, PCA is used as a dimensionality reduction technique. Feature scaling is done using the min-max approach. To calculate the magnitude of error during training and testing, binary cross-entropy is applied as it is a standard loss function and recommended for binary classification problems. For optimization purposes, we have optimized ADAM. A comparative study of the proposed method is done with the state-of-the-art methods. Experimental results on these standard microarray datasets and comparative analysis with state-of-the-art methods show that the performance of the proposed method is highly acceptable. To measure the performance of the proposed method, we have contributed performance measures such as classification accuracy, precision, recall, F-measure, ROC curve, confusion matrix, and log-loss. The classification accuracy of the proposed method on four datasets, namely Leukemia, Lung-Michigan, Ovarian and Prostate, is 1.00, which shows an ideal classification performance. Furthermore, the proposed method scores an accuracy of 0.99 on Lung-Harvard2, 0.96 on CNS and Colon, and 0.95 on Breast Cancer.

REFERENCES

[1]. Miller, K.D.; Ortiz, A.P.; Pinheiro, P.S.; Bandi, P.; Minihan, A.; Fuchs, H.E.; Martinez Tyson, D.; Tortolero-Luna, G.; Fedewa, S.A.; Jemal, A.M.; et al. Cancer Statistics for the US Hispanic/Latino Population, 2021. *CA A Cancer J. Clin.* 2021, *71*, 466–487. [Google Scholar] [CrossRef]

[2]. 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]

[3]. Brewczyński, A.; Jabłońska, B.; Mazurek, A.M.; Mrochem-Kwarciak, J.; Mrowiec, S.; Śnietura, M.; Kentnowski, M.; Kołosza, Z.; Składowski, K.; Rutkowski, T. Comparison of Selected Immune and Hematological Parameters and Their Impact on Survival in Patients with HPV-Related and HPV-Unrelated Oropharyngeal Cancer. *Cancers* 2021, *13*, 3256. [Google Scholar] [CrossRef] [PubMed]

[4]. Ahmed, Z.; Mohamed, K.; Zeeshan, S.; Dong, X. Artificial Intelligence with Multi-Functional Machine Learning Platform Development for Better Healthcare and Precision Medicine. *Database* 2020, *2020*, baaa010. [Google Scholar] [CrossRef]

[5]. Anna, A.; Monika, G. Splicing Mutations in Human Genetic Disorders: Examples, Detection, and Confirmation. *J. Appl. Genet.* 2018, *59*, 253–268. [Google Scholar] [CrossRef] [PubMed][Green Version]

[6]. Slatko, B.E.; Gardner, A.F.; Ausubel, F.M. Overview of Next-Generation Sequencing Technologies. *Curr. Protoc. Mol. Biol.* 2018, *122*, cpmb.59. [Google Scholar] [CrossRef]

[7]. Briglia, N.; Petrozza, A.; Hoeberichts, F.A.; Verhoef, N.; Povero, G. Investigating the Impact of Biostimulants on the Row Crops Corn and Soybean Using High-Efficiency Phenotyping and Next Generation Sequencing. *Agronomy* 2019, *9*, 761. [Google Scholar] [CrossRef][Green Version]

[8]. Phan, T.; Fay, E.J.; Lee, Z.; Aron, S.; Hu, W.-S.; Langlois, R.A. Segment-Specific Kinetics of MRNA, CRNA, and VRNA Accumulation during Influenza Virus Infection. *J. Virol.* 2021, *95*, e02102-20. [Google Scholar] [CrossRef]

[9]. Salem H. Attiya G. El-Fishawy N. 'Classification of human cancer diseases by gene expression profiles', *Appl. Soft Comput.*, 2017, 50, pp. 124–134 (doi: 10.1016/j.asoc.2016.11.026)

[10]. Lin T.C. Liu R.S. Chen C.Y. :et al.: 'Pattern classification in DNA microarray data of multiple tumor types', *Pattern Recognit.*, 2006, 39, (12), pp. 2426–2438 (doi: 10.1016/j.patcog.2006.01.004)

[11]. 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, (6), pp. 231–238 (doi: 10.1016/j.ygeno.2016.05.001)

[12]. Kumar M. Rath N.K. Swain A. et al.: 'Feature selection and classification of microarray data using MapReduce based ANOVA and K-nearest neighbor', *Procedia Comput. Sci.*, 2015, 54, pp. 301–310 (doi: 10.1016/j.procs.2015.06.035)

[13]. Kar S. Sharma K.D. Maitra M.: 'Gene selection from microarray gene expression data for classification of cancer subgroups employing PSO and adaptive K-nearest neighborhood technique', *Expert Syst. Appl.*, 2015, 42, (1), pp. 612–627 (doi: 10.1016/j.eswa.2014.08.014)

[14]. Garcia V. Sanchez J.S.: 'Mapping microarray gene expression data into dissimilarity spaces for tumor classification', *Inf. Sci.*, 2015, 294, pp. 362–375 (doi: 10.1016/j.ins.2014.09.064)

[15]. Kar S. Sharma K.D. Maitra M.: 'Gene selection from microarray gene expression data for classification of cancer subgroups employing PSO and adaptive K-nearest neighborhood technique', *Expert Syst. Appl.*, 2015, 42, (1), pp. 612–627 (doi: 10.1016/j.eswa.2014.08.014)