IJNRD.ORG ISSN : 2456-4184
INTERNATIONAL JOURNAL OF NOVEL RESEARCH AND DEVELOPMENT (IJNRD) | IJNRD.ORG
An International Open Access, Peer-reviewed, Refereed Journal

Threshold Methodology to Predict Brain with Gene Expression Pattern by using Machine Learning Algorithm

[1] Y.Tezaaw ,[2] Dr.K.Vijaya Lakshmi
[1] Research Scholar, [2]Associate Professor
[1,2] Dept.of Computer Science, S.V.University,Tirupati, A.P,India

Abstract: The Tumor that has been developed in an organ of brain or spinal cord is said to be brain cancer which is further classified into several types based on its location as well as origin namely glioma, pituitary adenomas, meningioma, schwannoma and medulloblastoma. One of the most frequent from the above brain cancer is Glioma which is represented as glial cells that plays the major type in brain whereas it is further classified as pilocytic astrocytoma, medulloblastoma, glioblastoma and ependymoma based on morphological appearance. The best way of detecting cancer during earlier stage is Gene Expression (GE) which reflects biochemical processes in cells, tissues and an organism's genetic aspects. The information about gene expression helps to measure the levels of gene expression as well as generate valuable data in computational analysis through sequencing methods of Deoxyribonucleic Acid (DNA) and Ribonucleic Acid (RNA) microarrays. At present, several researchers focuses Machine Learning (ML) technique in predicting diseases using GE data. The discovery of genomes study assist in interaction among genes as well as the disease and its interaction leads for specific phenotype that developed exponentially.

Keywords: Gene Expression (GE), Principal Component Analysis (PCA), hyperparameter tuning, brain cancer, Machine Learning (ML)

1. Introduction

Cancer refers to a group of disorders in which the human body develops malignant cells as a result of genetic mutation. As they mature, these cells are divided randomly and spread over the organs and, in many instances, can cause death. The second biggest cause of death worldwide behind cardiovascular illnesses is cancer [1]. Recent developments in gene expression analysis have made it a crucial tool for overcoming the fundamental difficulties in cancer diagnosis and medication discovery [2,3]. The significant function that many genes play in the development and progression of cancer is also revealed through gene expression studies. As a result, modifications in gene expression can serve as indicators for early identification of cancer and as a guide for choosing potential therapeutic targets. These methods may pave the way for more individualized, preventative, and predictive healthcare [4]. A gene is a sequence of instructions that instructs a cell to make proteins or other molecules based on the information in DNA. Translation involves converting DNA into messenger RNA (mRNA) and then converting that into proteins. The order of genetic changes occurring in a tissue or single cell under specific circumstances is evaluated using gene

expression analysis [5]. In this method, DNA transcript levels are measured in a sample tissue or cell in order to find out which genes are present and how much they are expressed. Comparing the number of base pairs from a DNA fragment sequenced to a known genomic or transcriptome source is one step in the quantification of gene expression. The sequenced reads must have enough distinguishing information to enable the application of bioinformatics techniques to connect the reads to the proper genes if the quantification to be accurate.

The primary cause of death for both women and men under 40 is brain cancer [6, 7]. Malignant brain tumors are also becoming more common [3], which has a negative influence on society and people's health [8, 9]. Primary brain tumors develop from the brain's own cells, whereas secondary brain tumors develop from cancerous cells that have travelled to the brain from elsewhere [10]. Studies have revealed that brain tumors are incredibly diverse, making categorization, segmentation, prediction and diagnosis are extremely difficult [11]. Cancer diagnosis, prognosis, and treatment have all benefited from microarray-based GE analysis. DNA microarray technology has also substantially altered about the understanding of what causes cancer in recent years. Small sample sizes and a wide range of gene expression levels in cancer microarray data can result in the "curse of dimensionality," which makes it challenging to categorize the data. ML methods are used by the bioinformatics community to categorize microarray data in various ways. The majority of the studies using microarray data sets for cancer classification that have been examined focus on accuracy in cancer classification without disclosing relevant biological information about the cancer classification approach. Microarray data sets can be interpreted biologically as well as classified accurately by models. There have been very few studies that address this issue.

2. Literature Review

The Tumor in brain can be investigated and diagnosed based on the GE is made automated through ML method is investigated through the evaluation is available literature. In this case, the literature review includes various classification techniques to determine accurate prediction in cancer through GE and its related functions.

According to Done et al., the traditional classification system of glioma cells is based on the type, including oligodendroglioma or astrocytoma, along with the mutation status of genes such as ATRX, Isocitrate Dehydrogenase gene (IDH) 1 and IDH2, and 1p/19q deletions as well as TP53 [13]. The Alzheimer's Disease Neuroimaging Initiative (ADNI), AddNeuroMed1 (ANM1), and ANM2 are three different datasets that are used in [14] to separate AD from CN. Highly significant genes are chosen using a variety of Gene Selection (GS) techniques, including Convergent Functional Genomics (CFG), transcription factors, variational autoencoders, hub genes and transcription factors. SVM, RF, L1-regularized LR (L1-LR), Logistic Regression (LR), and Deep Neural Network (DNN) are the five models used for classification for ANM1, ANM2, and ADNI, respectively and the resulting AUC values such as 87.4%, 80.4%, and 65.7%. The blood and brain bio-signature are compared by the authors, who also examine the biological processes of the AD-related blood genes. There are 140 genes are given in common between the blood and the brain, they use 1291 brain genes that were derived out of a gene expression dataset together with 2021 blood genes that have been gathered from the remaining three datasets. Li et al [15] provides an analysis of GE data from blood samples also correlates with blood and brain genes from AD patients. They discover 789 genes that express differently in the brain and the blood. The GS technique using the Least Absolute Shrinkage and Selection Operator (LASSO) regression algorithm. Models for classification include SVM, RF, and Logistic Ridge Regression (RR). The AD cases are distinguished from the control cases with an accuracy of 78.1%. Multiple brain areas are examined in [16] to find potential AD diagnostic biomarkers. The identification of AD biomarkers uses GE data collected from six different brain areas. The most useful genes are chosen using a t-test. AD biomarkers are checked using significance tests to determine their suitability for clinical diagnosis. In order to predict AD using a DNN, the researchers of [17] combine datasets on DNA methylation and GE. The most informative characteristics are chosen using t-stochastic closest neighbor approaches and PCA.

The metastatic melanoma is distinguished from the primary melanoma by a method which is created by Metri et al [18] at the same time that the SVM-based techniques were being created, but they employed Adaboost around RF. The researchers used two independent testing datasets as well as a training dataset. According to the study's findings, six genes namely, TMEM45B, SPRR3, KRT16, ALDH1A1, HSP90AB1 and KIT are involved in the metabolism of phospholipids, the construction of protein-lipid complexes, the control of inflammatory response, the inhibition of protein kinase function, and the regulation of innate immune response in metastatic melanoma. Metri et al [18] utilizing Adaboost around Random Forest (RF) to demonstrate promising results while taking into account more biological information, however SVM outperformed other feature selection techniques. This result illustrates the ability of SVM approaches reflects input parameter choices with less sensitive and differentiate the classes in complex datasets.

3. Research Methodology

The technique can be applied more easily, similar to how standard clustering methods work, if there are known to be a certain number of independent parts. The optimal number of PCA vectors has accomplished with 85 to 90% of information from the dataset. This may help the ML classifier model through collection of various classifiers in a single supervised Lazy predict library and the architecture of Balanced Hyper parameter tuned prediction of Brain Tumor classification is shown in Figure 1.

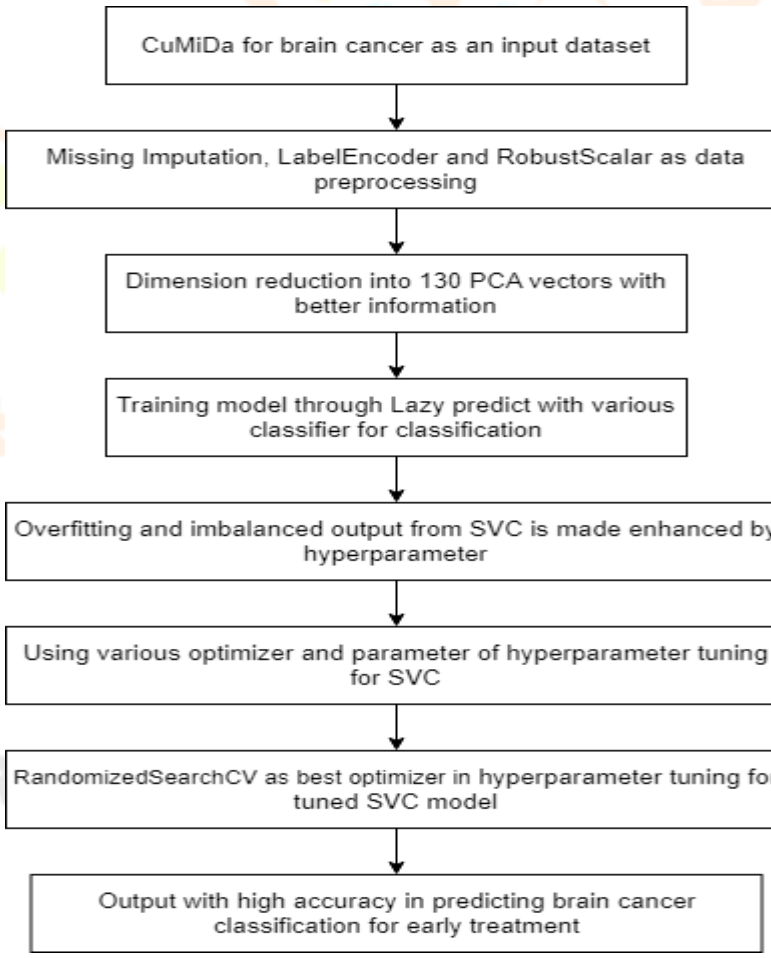

Figure:1 Proposed Architecture for Predicting Brain Tumor through GE Omnibus

The proposed architecture is segregated into several phases namely data collection, data preprocessing, feature extraction and data modeling for accomplishing the prediction in classifying the brain cancer. Moreover, the proposed architecture is introduced to generate the balanced accuracy to avoid overfitting of data from the ML

model. Hence, the hyperparameter tuning is done through random search optimizer as well as parameter tuning like learning rate, loss functions, maximum depth, etc. Thus, the Hyper parameter tuned model generates balance accuracy that improves the accuracy of the ML classifier model.

3.1 Dataset

The GE brain cancer dataset is collected from CuMiDa in which the respiratory contains dataset of cancer micro array which involves GE omnibus and CSV file containing the GE levels of 54676 columns of genes from 130 rows of samples. The new dataset is provided by CuMiDa and it distinguishes itself from other databases that have been meticulously and thoroughly edited to remove extraneous probes, improve sample quality and compensate for background noise, which makes the data source are more acceptable for computational research. There are 4 distinct forms of brain cancer such as pilocytic astrocytoma, medulloblastoma, glioblastoma and ependymoma as well as normal patient cases. The main aim of CuMiDa's is to provide these datasets with homogeneous, cutting-edge biological preprocessing together with multiple threefold cross-valid benchmark outcomes to support ML studies concentrated on cancer research.

3.2 Feature extraction and data modeling

The data get collected as well as check for missing value which is implemented with imputing the missing value. Once the missing imputation is done, the data has been preprocessed using label encoder and Robust scaler for scaling the all-variable unit as unique is handled. The dataset was first subjected to PCA with preprocessing stage. Using the top most best performed ML model is considered with untuned ML model and the top model can be improved its accuracy, the parameter of classification can be improved by proposed random search CV hyperparameter tuning.

3.3 Hyper parameter tuning working principle

Let us consider the feature variable as X and target variable as Y with an unidentified joint distribution as D(X,Y) in which the sample dataset is segregated as S with m observations. The ML model can acquire data using functional relationship among X and Y by generating a prediction model as $\hat{F}(X, \theta)$ that can be controlled through n-dimensional hyperparameter configuration $\theta = (\theta_1, \theta_2, \dots, \theta_n)$ from the hyperparamater search spacing $\Theta = (\Theta_1, \Theta_2, \dots, \Theta_n)$. The prediction performance has been calculated through point-wise relationship among prediction function $\hat{F}(X, \theta)$ and true label Y.

3.3.1 In this work, Google Colab is used with Jupiter IDE which assist to share and create document that can be narrated with text, livecode and visualizations. The dataset is collected and split it into 60% train dataset and 40% test dataset.

3.3.2 The tunability hyperparameter used other tools like Scipy, Seaborn and Pandas. From the top most best performed ML model is considered with untuned ML model. Improvement of model accuracy is based on the parameter of regression is enhanced through proposed Randomized Search CV hyperparameter tuning. Lazy predict is best python libraries which assist for semi-automate the ML based tasks.

3.3.3 Lazy predict help in building several and different basic ML models with certain code and assist in understanding which models may execute better accuracy without any parameter tuning. In this research lazy classifier is used for solving regression based dataset in predicting the brain tumor classification accurately to provide earlier treatment. From the lazy classifier analysis, SVC is considered to be top most model which is evaluated through confusion matrix metrics.

5. Conclusion

This research has presented architecture for the prediction of the class or severity of brain cancer which is not concentrated enough attention in the literature using GE data. the GE based prediction of brain cancer can be done through ML technique in which the numerous features information is grouped in term of cluster by feature extraction. The PCA technique can be used to extract the features by mapping the data with high dimensions into a lower-dimensional space while reducing the total squared error. This technique preserves the most variation from the original data. As a result, there are also less complicated issues related to time and space. The approach primarily helps with signal differentiation from diverse sources.

Reference

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.
2. Munkácsy, G.; Santarpia, L.; Gy″orffy, B. Gene Expression Profiling in Early Breast Cancer—Patient Stratification Based on Molecular and Microenvironment Features. Biomedicines 2022, 10, 248.
3. Brewczy ´nski, A.; Jabło ´nska, B.; Mazurek, A.M.; Mrochem-Kwarciak, J.; Mrowiec, S.; Snietura, 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.
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.
5. Anna, A.; Monika, G. Splicing Mutations in Human Genetic Disorders: Examples, Detection, and Confirmation. J. Appl. Genet. 2018, 59, 253–268.
6. Siegel, R.L.; Miller, K.D.; Jemal, A. Cancer statistics, 2021. CA A Cancer J. Clin. 2021, 66, 7–30.
7. Rehman, M.U.; Cho, S.; Kim, J.; Chong, K.T. BrainSeg-Net: Brain MR Image Segmentation via Enhanced Encoder–Decoder Network. Diagnostics 2021, 11, 169.
8. Havaei, M.; Davy, A.; Warde-Farley, D.; Biard, A.; Courville, A.; Bengio, Y.; Pal, C.; Jodoin, P.-M.; Larochelle, H. Brain segmentation with Deep Neural Networks. Med. Image Anal. 2017, 35, 18–31.
9. Isensee, F.; Jäger, P.F.; Full, P.M.; Vollmuth, P.; Maier-Hein, K.H. nnU-net for brain segmentation. In International MICCAI Brainlesion Workshop; Springer: Cham, Switzerland, 2020; pp. 118–132.
10. Zeineldin, R.A.; Karar, M.E.; Coburger, J.; Wirtz, C.R.; Burgert, O. DeepSeg: Deep neural network framework for automatic brain segmentation using magnetic resonance FLAIR images. Int. J. Comput. Assist. Radiol. Surg. 2020, 15, 909–920.
11. Perrin, S.L.; Samuel, M.S.; Koszyca, B.; Brown, M.P.; Ebert, L.M.; Oksdath, M.; Gomez, G.A. Glioblastoma heterogeneity and the tumour microenvironment: Implications for preclinical research and development of new treatments. Biochem. Soc. Trans. 2019, 47, 625–638.
12. Swarna Priya, R.M.; Maddikunta, P.K.R.; Panimala, M.; Koppu, S.; Gadekallu, T.R.; Chowdhary, C.L.; Alazab, M. An Effective Feature Engineering for DNN Using Hybrid PCA-GWO for Intrusion Detection in IoMT Architecture. Comput. Commun. 2020, 160, 139–149.
13. Dono A, Ballester LY, Primdahl D, Esquenazi Y, Bhatia A. 2021. IDH-mutant low-grade glioma: advances in molecular diagnosis, management, and future directions. Current Oncology Reports 23(2):20.
14. Lee, T.; Lee, H. Prediction of Alzheimer's disease using blood gene expression data. Sci. Rep. 2020, 10, 3485.

15. Li, X.; Wang, H.; Long, J.; Pan, G.; He, T.; Anichtchik, O.; Belshaw, R.; Albani, D.; Edison, P.; Green, E.K.; et al. Systematic analysis and biomarker study for Alzheimer's disease. Sci. Rep. 2018, 8, 17394.
16. Wang, L.; Liu, Z.P. Detecting diagnostic biomarkers of Alzheimer's disease by integrating gene expression data in six brain regions. Front. Genet. 2019, 10, 157.
17. Park, C.; Ha, J.; Park, S. Prediction of Alzheimer's disease based on deep neural network by integrating gene expression and DNA methylation dataset. Expert Syst. Appl. 2020, 140, 112873.
18. Metri R, Mohan A, Nsengimana J, et al. Identification of a gene signature for discriminating metastatic from primary melanoma using a molecular interaction network approach. Sci Rep 2017; 7:17314