International Journal of
INTELLIGENT SYSTEMS AND APPLICATIONS IN ENGINEERING
ISSN:2147-6799 www.ijisae.org Original Research Paper

Identifying Biomarkers from Medical Images Using Machine Learning Techniques

[1]Aditya Agnihotri, [2]Mahesh Manchanda, [3]Dr. Asif Ibrahim Tamboli

Submitted:20/03/2023 **Revised:**25/05/2023 **Accepted:**12/06/2023

Abstract: Genetic information is necessary for studying the biological processes that, when interrupted, cause certain cancers to form. Although advances in sequencing technology have made it possible to record the nuances of gene interaction in several data formats, using these methods to identify, diagnose, and treat cancer remains difficult. Machine learning has helped researchers in a number of areas, including supervised and unsupervised learning, as well as gene identification, but the results have been less than visually satisfying. Using RNA-SEQ data from The Cancer Genome Atlas, this research focuses on multi-class classification of cancer, extraction of key characteristics, and identification of relevant genes for 10 different types of cancer.Tests conducted with the restricted hardware resources at hand have shown that these limitations do not always exclude the possibility of positive results. Stacked de-noising auto encoders were employed for feature extraction and biomarker identification, while 1D convolutional neural networks were used for classification. Both the recovered features and the relevant genes were used in the classification process, with the former typically performing better (about 94% accurate) than the latter (95% accurate). By using stacked denoising auto-encoders to construct matrix weights and features, we were able to identify common cancer-related pathways and their associated genes.

Keywords*: convolutional neural networks, biomarker identification, auto encoders*

1. Introduction

Human physiology and biological processes rely heavily on genetic makeup to operate normally. The involvement of genes in this is crucial. However, the occurrence such molecular events is not completely predictable, which might lead to deviations in routine procedures. Such mechanistic shifts may lead to alterations or chromosomal rearrangements, either of which can have beneficial or deleterious effects but are closely linked to the onset of cancer (1). An important opportunity to identify cancer at an earlier stage or to limit its spread later on exists if genes or gene clusters that cause the development of malignant cells are identified (1).

The World Health Organization estimates that 8.2 million people worldwide die from cancer every year (2). Worldwide, clinicians and researchers continue to concentrate their efforts on improving cancer detection and therapy. Numerous advances in genomics have resulted from the advent of high volumes DNA sequencing technologies (1). The capacity to quickly recognize alteration profiles, RNA expressions, and micro-RNA profiles are among these findings.The relevance of this genetic data may be grasped by remembering that machine learning algorithms can be used to do statistical studies of cancer diagnosis, progression, and prognosis. It is also possible that precision medicine (1) (3) will ignore the sub-networks of genes and particular biomarkers that are responsible for cancer.

In several domains, including image processing, natural language processing, and voice recognition, experts have found success using machine learning and deep learning methods. However, in bioinformatics, the focus has often been on using clustering methods to find subgroups or biomarkers. Researchers have recently shifted their attention to developing supervised learning techniques for categorizing RNA-seq expression data. Very basic or simplistic methods have been employed to classify somatic mutations. While efforts have been made to identify cross-cancer biomarkers, multi-class categorization has received less attention.

By using machine learning techniques to the study of genetic data, we may more easily identify important genes and classify different types of cancer. Examples of methods used to reduce the number of dimensions include principal component analysis, k-means

1Lecturer, Department of Comp. Sc. & Info. Tech. Graphic Era Hill University, Dehradun Uttarakhand 248002, adityaagnihotri@gehu.ac.in
2Professor, Department of Comp. Sc. & Info. Tech. Graphic Era Hill University, Dehradun Uttarakhand 248002, mmanchanda@gehu.ac.in
3Assistant Prof. Department of Radiology, Krishna Institute of Medical Sciences, Krishna Vishwa Vidyapeeth "Deemed To Be University" Karad Malkapur, Karad (Dist. Satara), Maharashtra, India. PIN – 415539
Email ID:- drtamboliasif@gmail.com

clustering, and independent component analysis, while examples of methods used for classification include k-nearest neighbors, random forests, and support vector machines. Academics are increasingly looking to deep learning approaches to solve classification problems like object recognition and image categorization (1). Large datasets and powerful computers are largely to thank for this development. Deep learning's recent applications to genetic data for tasks like drug discovery, gene control, and protein classification have had an outsized effect on bioinformatics (6). This is now feasible due to the availability of vast amounts of data. As a result, deep learning architectures have been tried out in an attempt to improve classification accuracy and develop biomarkers for cancer detection based on gene expressions or mutation patterns.

2. Related Work

Bioinformatics experts have faced a difficult but crucial challenge in trying to identify cancer from genomic data. Now that DNA sequencing can be done at a lower cost, bigger datasets may be utilized for analysis in the service of diagnosis, therapy, and prognosis. This has led to the development of many different dimensionality reduction and classification techniques based on machine learning and feature extraction. Furthermore, attention has shifted from analyzing the impact of single genes to analyzing gene networks. Moreover, the same networks may also be responsible for a wide range of disorders.

As sequencing technology has improved, researchers have begun using more nuanced measures of gene expression, including as DNA sequencing and microarray expression (10). Recent years have seen a transition in cancer research based on gene expressions from microarrays to RNA-seq datasets. Cancer detection and related gene identification methods have, however, proved mostly consistent across data types.

Significant genes have been clustered together using cluster analysis to facilitate precise sample categorization. Both K-Nearest Neighbors and distance metrics with which it may be combined have been used to classify breast cancer (12) and to assess correlations between gene expressions in prostate cancer (11). In addition, wavelet transforms were used to identify driver genes in colon and leukemia samples, which were then used in a k-means clustering classification (13). Breast cancer subtypes(14) and low-dimensional cancer data(15) have both been classified using hierarchical clustering. SVMs have proved very successful in applications outside clustering, such as the categorization of gene expression patterns for various malignancies. Classifying leukemia subtypes was much facilitated by using multi-category SVMs (4). In addition, network algorithms have been used to locate the genetic circuitry that promotes the development of certain cancers (16).

The advent of multiple machine learning algorithms, however, has prompted investigation into their potential use in the fields of cancer detection and biomarker discovery. There has been a clear leaning toward using deep learning techniques for dimensionality reduction and classification since their introduction.

This design was utilized by Gupta et al. (1) in their research to learn an interpretable representation of gene expressions data from the yeast cell cycle. Clusters of genes scored from unlabeled input were compared to SDA's cluster analysis output. In addition to these clustering approaches, PCA was also examined using gene expression profiles. The data show that SDA is superior than PCA for capturing gene co-expressions.

In their study of breast cancer, Danaae et al. (5) utilized SDA to identify and extract highly linked genes from RNA-seq expression data. The efficiency of SDA was evaluated using PCA and KPCA as comparison methods. In addition to dimensionality reduction, they examined the weight matrix of SDA to locate causal genes. The genes in question are now known as DCGs, or Deeply Connected Genes. Panther pathways was utilized to examine how various genes and tumor suppressor genes do their respective tasks.

Researchers Bhat et al. (7) tried out Generative Adversarial Deep Convolution Networks for precise classification of breast cancer and prostate cancer gene expression datasets.

The effectiveness of DDBN in cancer classification was established by Karabulut et al. (8) in comparison to more conventional models such as SVM. Laryngeal, colorectal, and bladder cancers were studied separately in laboratory experiments. SVM, Random Forrest, and K-NN were used on all datasets for evaluation. According to the findings, DDBN fared better than the other categories considered.

Differentiating tumor samples from normal ones was the primary focus of the study by Liu et al. (9). They presented an SAE/SDA-inspired technique for expanding training samples. In this study, we suggest using 1DCNN to categorize tumors. Instead of the usual two-dimensional vectors used in picture classification, it accepts input in a single dimension. On every single dataset, 1DCNN outperformed SAE.

Teixeira et al. (17) used SDA to identify the most useful genes for ANN-based thyroid cancer categorization. They compared the deep learning feature extraction technique to more conventional approaches like principal component analysis and kernel principal component

analysis. Three sets of genes were identified as having similar roles when SDA output was evaluated using the Connected Weights Method.

Because of this, the usefulness of machine learning algorithms for obtaining features and identifying significant genes stands out in spite of the fact that the medical field as a whole is heading toward precision medicine. As a result, researchers have started to play with deep learning with the goals of multi-class classification and the finding of biomarkers. In recent years, the field of biology has joined the ranks of those other fields that have identified the advantages of utilizing deep learning to solve classification problems requiring large datasets and feature extraction.

3. Material and Method

Acquisition of data

As was previously indicated, gene expression datasets have seen the greatest utility in the context of anomaly categorization. With help from portals backed by The Cancer Genome Atlas (TCGA), we compiled the data used in this analysis.

RNA-seq Expressions

Read counts and normalized expressions of genes are available via the TCGA portal for 33 distinct cancer types. Each kind of data must use the same sequencing method and undergo the same preprocessing procedures to guarantee that they share the same genes necessary for multi-class categorization. Both raw and RSEM normalized datasets for RNA-seq expressions are available via the GDAC site at the Broad Institute. As can be observed in Table 1, this study makes use of a standardized dataset of Illumina Hiseq RSEM data.

Table 1: TCGA multi-class cancer dataset

Cancer Type	No. of Cancerous Samples
Breast invasive carcinoma(BRCA)	1100
Adrenocortial carcinoma(ACC)	70
Cervical and endocervical cancer(CESC)	304
Head and neck Squamous Carci-noma(HNSC)	520
Kidney renal papillary cell carcinoma(KIRP)	290
BrainLower Grade Glioma (LGG)	516
Lung adenocarcinoma (LUAD)	501
Pancreatic adenocarcinoma (PAAD)	178
Prostate adenocarcinoma (PRAD)	497
Stomach adenocarcinoma(STAD)	416
Uterine Carcinosarcoma (UCS)	57
Bladder urothelial carcinoma(BLCA)	408

Dataset Split

Each sample was labeled with the matching cancer type after the datasets for all 12 kinds were pooled into a single dataset. Every specimen was assigned a number from 0 to 11, each of which represents a distinct kind of cancer. About 4967 samples from 12 different cancer types were included in the dataset, which was then partitioned into training, validation, and test sets. The breakdown of each group was as follows: 70%, 15%, and 15%. The dataset was divided proportionally by class since there was an obvious class imbalance among the various kinds. To illustrate, this implies that each category was separated into three groups according to the percentages given above.

Preprocessing

After normalization, genes that consistently had 0 values across all samples were excluded from further analysis.

SDA

In the studies, 1DCNN was fed the results of SDA to determine the various cancers present. The SDA was trained using greedy-layer-wise training, in which the output of one layer was utilized as input to the next layer, and each layer was trained for a fixed number of iterations. Since it was found that fewer hidden units per layer allowed for greater feature incorporation, this parameter was steadily reduced. Five significant tests were conducted, resulting in five lists of highly rated genes and streamlined feature sets. There were two applications for SDA's results:

Using Reduced Features

The trimmed-down features of the dataset were produced by the last SDA layer. After the necessary number of cycles, these characteristics for each sample were saved as part of the training dataset. The final layer weights were also saved so that they could be used to prune features in the test and validation data sets.

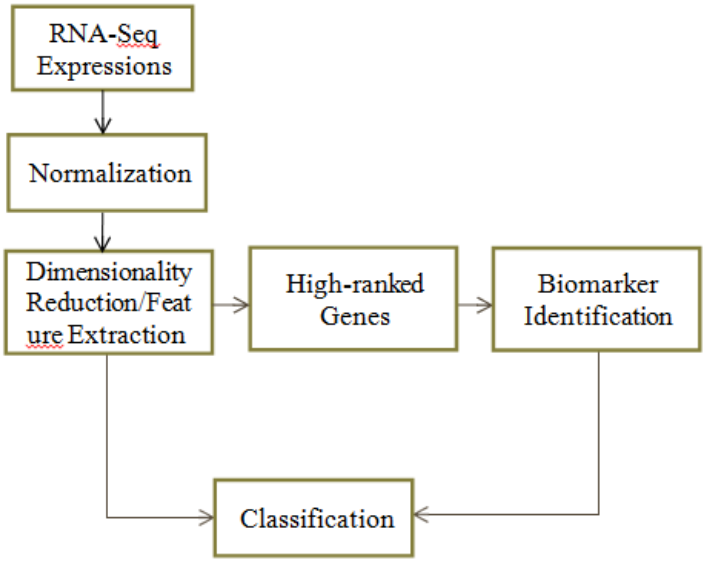

Figure 1: Workflow

The results, in the form of a weight matrix and a set of prioritized genes, were utilized in the preceding section. SDAs with varying numbers of layers were trained using a wide variety of hidden unit architectures. According to the research, SDA is able to better absorb the characteristics for reconstruction if the number of concealed units is steadily lowered. After filtering out the genes that had a value of zero in every sample, the initial count of 20531 dropped to 20313. The total number of secret units in the building was between 15,000 and 200, but the number of units in the top two storeys was always 15,000. For these trials, we only modified the third-to-last and final layers. Extensive testing was done with 3 and 4 layers because of the improved precision they provided.

The highest performance was shown with decreased features when a larger number of units were used in the reconstruction layer. One-class support neural networks (1DCNNs) were used to evaluate the characteristics. The accuracy peaked, however, at roughly 96.5%, after 4000 features.

In Figure. 1, we see a graph depicting the accuracy attained by 1dcnns with a range of layer counts and feature reductions. This graph features studies conducted using three and four layers. There were always 15000 units in the first layer, and 10000 units in the second.

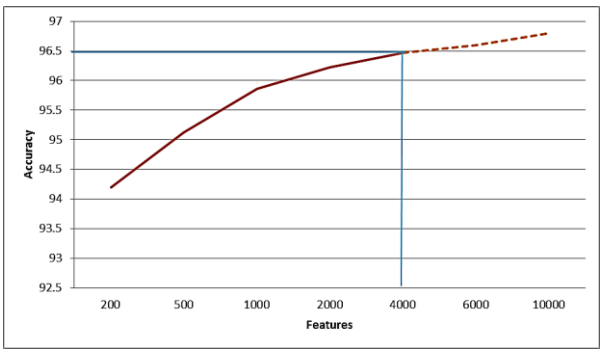

Fig 2: Accuracy with linear combination of reduced features

High-Ranked Genes

Genes were ranked using the sum of the weights from the weight matrix used in each SDA layer. Higher-scoring genes have been shown to be causal in cancer more often than other genes. According to (18), the genes with very negative or extremely positive weights in the SDA weight matrix are important genes. That's why the genes with the largest standard deviations from the mean weights are considered to be the best. Therefore, we isolated the important genes by looking at their divergence from the mean. The trials were limited in scope due to lack of funding, yet they nevertheless perform very well in terms of identifying genes of interest.

It was found that the genes that deviated the most from the mean were the most important. Genes that were shared by many SDA structures were also determined to be cancer-relevant in other diseases. The goal of this study has always been to maximize performance while minimizing gene count, and results show that structures with 200-1000 features improve performance by 4-5 standard deviations.

The literature was combed for evidence that four genes are implicated in various kinds after similarities were discovered across all routes, all sets, and all standard deviations. As can be seen in Table 2, the research demonstrates the potential and importance of realized genes.

Genes	Cancer Types
WNT10A	BRCA [19] LUAD [20] BLCA [21] PRAD [22] PAAD [23]
PIK3C2G	BRCA [24]

4. Conclusion

The study's results show that the proposed approach might be used to classify 10 distinct types of cancer and pinpoint important genes. For low-dimensional feature sets, the average accuracy of a combination of SDA and 1DCNN was 94%, while for highly ranked genes, it was 95%. This highlights the potential use of relevant gene sets in cancer classification as well as cross-cancer gene and pathway discovery. Panther Database aided in the identification of pathways and genes involved in cancer among the datasets generated by several studies. The research lends credence to the idea that these genes are linked to a wide variety of cancers. This highlights the adaptability of our method for recognizing biomarkers and classifying cancers into many classes. This bolsters efforts to identify genes whose roles are less well understood.

References

[1] Gupta A, Wang H, Ganapathiraju M. Learning structure in gene expression data using deep architectures, with an application to gene clustering. In: Bioinformatics and Biomedicine (BIBM), 2015 IEEE International Conference on. 2015. p. 1328–35.

[2] Yuan Y, Shi Y, Li C, Kim J, Cai W, Han Z, et al. DeepGene : an advanced cancer type classifier based on deep learning and somatic point mutations. BMC Bioinformatics [Internet]. 2016;17(Suppl 17). Available from: http://dx.doi.org/10.1186/s12859-016-1334-9

[3] Fawzy H, Kamel M, Al-amodi HSAB. Exploitation of Gene Expression and Cancer Biomarkers in Paving the Path to Era of Personalized Medicine. Genomics Proteomics Bioinformatics [Internet]. 2017;15(4):220–35. Available from: http://dx.doi.org/10.1016/j.gpb.2016.11.005

[4] Lee Y, Lee C-K. Classification of multiple cancer types by multicategory support vector machines using gene expression data. Bioinformatics. 2003;19(9):1132–9.

[5] Danaee P, Ghaeini R, Hendrix DA. A deep learning approach for cancer detection and relevant gene identification. In: PACIFIC SYMPOSIUM ON BIOCOMPUTING 2017. 2017. p. 219–29.

[6] Min S, Lee B, Yoon S. Deep learning in bioinformatics. Brief Bioinform. 2017;18(5):851–69.

[7] Bhat RR, Viswanath V, Li X. DeepCancer : Detecting Cancer through Gene Expressions via DeepGenerative Learning. (Ml).

[8] Karabulut EM. Discriminative deep belief networks for microarray based cancer classification .2017;28(3):1016–24.

[9] Liu J, Wang X, Cheng Y, Zhang L. Tumor gene expression data classification via sample expansion- based deep learning. Oncotarget. 2017;8(65):109646.

[10] Wang Z, Gerstein M, Snyder M. RNA-Seq: a revolutionary tool for transcriptomics. Nat Rev Genet.2009;10(1):57.

[11] Singh D, Febbo PG, Ross K, Jackson DG, Manola J, Ladd C, et al. Gene expression correlates of clinical prostate cancer behavior. 2002;1(March):203–9.

[12] Rules C, Medjahed SA. Breast Cancer Diagnosis by using k-Nearest Neighbor with Different Breast Cancer Diagnosis by using k-Nearest Neighbor with Different Distances and Classification Rules. 2013;(January).

[13] Mishra P, Bhoi N, Meher J. Effective clustering of microarray gene expression data using signal processing and soft computing methods. In: Electrical, Electronics, Signals, Communication and Optimization (EESCO), 2015 International Conference on. 2015. p. 1–4.

[14] Woodward WA, Krishnamurthy S, Yamauchi H, El-Zein R, Ogura D, Kitadai E, et al. Genomic and expression analysis of microdissected inflammatory breast cancer. Breast Cancer Res Treat. 2013;138(3):761–72.

[15] Kumar, C. ., & Muthumanickam, T. . (2023). Analysis of Unmanned Four-Wheeled Bot with AI Evaluation Feedback Linearization Method. International Journal on Recent and Innovation Trends in Computing and Communication, 11(2), 138–142. https://doi.org/10.17762/ijritcc.v11i2.6138

[16] Khan J, Wei JS, Ringner M, Saal LH, Ladanyi M, Westermann F, et al. Classification and diagnostic prediction of cancers using gene expression profiling and artificial neural networks. Nat Med. 2001;7(6):673.

[17] Martinez-Ledesma E, Verhaak RGW, Treviño V. Identification of a multi-cancer gene expression biomarker for cancer clinical outcomes using a network-based algorithm. Sci Rep. 2015;5:11966.

[18] Teixeira V, Camacho R, Ferreira PG. Learning influential genes on cancer gene expression data with stacked denoising autoencoders. In: Bioinformatics and Biomedicine (BIBM), 2017 IEEE International Conference on. 2017. p. 1201–5.

[19] Microbe-host AI. ADAGE-Based Integration of Publicly Available Pseudomonas aeruginosa Gene Interactions. 1(1):1–17.

[20] Braune E-B, Seshire A, Lendahl U. Notch and Wnt Dysregulation and Its Relevance for Breast Cancer and Tumor Initiation. *Biomedicines*. 2018;6(4):101. doi:10.3390/biomedicines6040101

[21] Tammela T, Sanchez-Rivera FJ, Cetinbas NM, et al. A Wnt-producing niche drives proliferative potential and progression in lung adenocarcinoma. *Nature*. 2017;545(7654):355-359. doi:10.1038/nature22334

[22] bin Saion, M. P. . (2021). Simulating Leakage Impact on Steel Industrial System Functionality. International Journal of New Practices in Management and Engineering, 10(03), 12–15. https://doi.org/10.17762/ijnpme.v10i03.129

[23] Zhang M, Li H, Zou D, Gao J. Ruguo key genes and tumor driving factors identification of bladder cancer based on the RNA-seq profile. *Onco Targets Ther*. 2016;9:2717-2723. doi:10.2147/OTT.S92529

[24] Ahmad I, Sansom OJ. Role of Wnt signalling in advanced prostate cancer. *J Pathol*. 2018;245(1):3-5.doi:10.1002/path.5029

[25] Fakhar M, Najumuddin, Gul M, Rashid S. Antagonistic role of Klotho-derived peptides dynamics in the pancreatic cancer treatment through obstructing WNT-1 and Frizzled binding. *Biophys Chem*. 2018;240(June):107-117. doi:10.1016/j.bpc.2018.07.002

[26] Fidalgo F, Rodrigues TC, Pinilla M, et al. Lymphovascular invasion and histologic grade are associated with specific genomic profiles in invasive carcinomas of the breast. *Tumor Biol*. 2015;36(3):1835-1848. doi:10.1007/s13277-014-2786-z