Progress in Microbes and Molecular Biology

HH PUBLISHER

Original Research Article

Comparative Analysis of Deep Learning Algorithm for Cancer Classification using Multi-omics Feature Selection

Nur Sabrina Azmi[1], Azurah A Samah[1*], Vivekaanan Sirgunan[1], Zuraini Ali Shah[1], Hairudin Abdul Majid[1], Chan Weng Howe[1], Nies Hui Wen[1], Nuraina Syaza Azman[1]

Article History

Received: 16 May 2022;

Received in Revised Form: 27 September 2022;

Accepted: 4 October 2022;

Available Online: 6 October 2022

[1]Faculty of Computing, Universiti Teknologi Malaysia, 81310 Johor Bahru, Johor, Malaysia; nsabrina36@graduate.utm.my (NS); vivekaanan16@gmail.com (VS); aszuraini@utm.my (ZAS); hairudin@utm.my (HAM); cwenghowe@utm.my (CWH); huiwennies@utm.my (NHW); nsyaza7@graduate.utm.my (NSA)

*Corresponding author: Azurah A Samah; Faculty of Computing, Universiti Teknologi Malaysia, 81310 Johor Bahru, Johor, Malaysia; azurah@utm.my (AAS)

Abstract: Advancement of high-throughput technologies in omics studies had produced large amount of information that enables integrated analysis of complex diseases. Complex diseases such as cancer are often caused by a series of interactions that involve multiple biological mechanisms. Integration of multi-omics data allows more advanced analysis using features from various aspects of biology. However, analysing cancer multi-omics data on a large scale could be challenging due to the high dimensionality of the data. The recent development of advanced computational algorithms, especially deep learning, had sparked numerous efforts in applying these algorithms in multi-omics studies. This study aims to investigate how deep learning algorithms, namely stacked denoising autoencoder (SDAE) and variational autoencoder (VAE) can be used in cancer classification using multi-omics data. Moreover, this study also investigates the impact of feature selection in multi-omics analysis through the implementation of an embedded feature selection. The multi-omics data used in this study includes genomics, methylomics, transcriptomics and clinical data for a case study of lung squamous cell carcinoma. The classification performance has been compared and discussed in terms of the effectiveness of different models and the impact of feature selection. Results showed that VAE outperforms SDAE with 91.86% accuracy, 22.73% specificity and 0.21% Matthews Correlation Coefficient (MCC).

Keywords: Genomic; Cancer Classification; Multi-omics; Deep learning; Feature Selection; Recursive Feature Elimination; Stacked Denoising Autoencoder; Variational Autoencoder

1. Introduction

Cancer is a complicated and heterogeneous disease to human health as it has contributed to a tremendous number of deaths worldwide [1]. Cancer research has been carried out by past researchers for decades. The classification of cancer based on molecular level research attracted the attention of numerous researchers, as it offered a systematic, precise and objective diagnosis for different types of cancer [1]. Clinical practices these days prefer

classifying cancer based on its tissue or cell type origin as well as pathogens [2]. Types of genomic data that are regularly consolidated for cancer diagnosis include gene expression, DNA methylation, mRNA expression, microRNA (miRNA) expression and also protein expression [3]. Integration of these components is expected to improve our understanding of the clinical and biological importance [4]. Integrative analysis of multi-omics data is the key to connecting cancer genetics, clinical and epidemiological information to ensure patients receive an effective and proper diagnosis [3]. Accordingly, one of the significant goals of cancer multi-omics study is to discover possible cancer subtypes using molecule-level signatures, which would be handy in successful diagnosis and treatments [4]. Multi-omics analysis has become increasingly popular in biomedical research as these omics with different unique features of their own are being integrated using effective integration strategies to obtain valuable data for solving biological problems. However, analysing cancer multi-omics data on a large scale could be genuinely challenging as it requires an efficient and robust computational algorithm [4]. This is due to the heterogeneity of distinct omics data in terms of scale, dimension and quality, requiring subtle processing. Fortunately, deep learning has grown into a formidable force in handling big data, which now stands as one of the most powerful architectures in overcoming complex problems related to biological data. It has been continuously refreshing the state-of-the-art performance of many machines learning tasks and facilitating the development of numerous disciplines. Besides, deep learning allows computational models composed of multiple processing layers to understand data representations with multiple levels of abstraction [5]. Deep learning has also clearly demonstrated its power in promoting bioinformatics, including sequence analysis, structure prediction and reconstruction, biomolecular property and function prediction, biomedical image processing and diagnosis, biomolecule interaction prediction and systems biology [6].

Since computational methods are being widely exploited in the bioinformatics field lately, Stacked Denoising Autoencoder (SDAE) is considered a handy algorithm in terms of gene regulatory targets discovery, disease detection and drug discovery [7]. In addition, Variational Autoencoder (VAE) has shown promising character when applied for single omics-based research carried out for drug discovery and disease diagnosis as well. Since the model is unsupervised and entirely data driven, it does not rely on existing omics annotations [8]. That is why these two deep learning algorithms were chosen. This paper aims to compare the performance of two deep learning architectures namely VAE and SDAE using a multi-omics dataset derived from a case study. The performance of these models will be measured in terms of model accuracy and loss rate as integrated multi-omics data will be analysed thoroughly.

The rest of this paper is further organised as follows. Section II illustrates the literature review for detailed information in this study while in section III explains all the steps taken to complete this analysis. Then, section IV describes the result and discussion of the model performance and ends with a conclusion included in section V.

2. Materials and Methods

Experimental workflow has been done to understand further the processes and steps involved in implementing the algorithms and methodologies involved as a part of this research. Figure 1 demonstrates the experimental framework of this study.

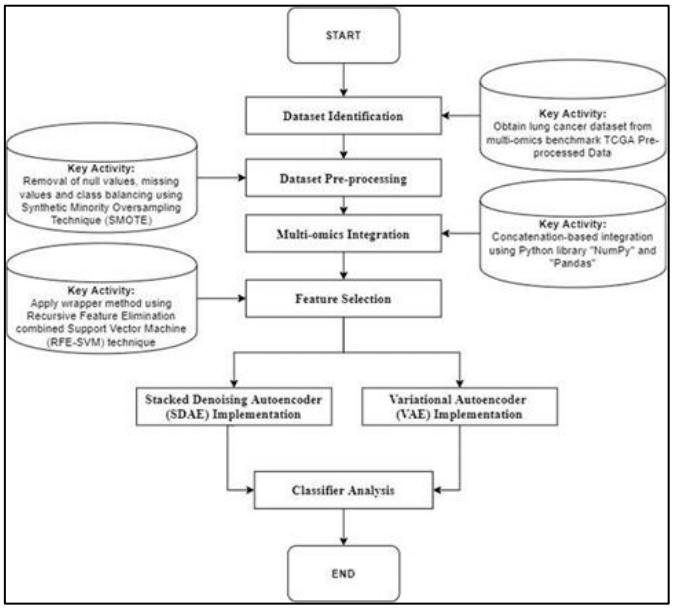

Figure 1. Experimental Framework

2.1. Omics Dataset

The cancer dataset used in this study was Lung Squamous Cell Carcinoma (LUSC) obtained from the Multi-Omics Cancer Benchmark TCGA Pre-processed Data publicly available at Multi-Omics Cancer Benchmark repository [9]. The URL for the dataset is http://acgt.cs.tau.ac.il/multi_omic_benchmark/download.html. Table 1 shows the summary of LUSC data.

Table 1. Summary of LUSC data.

Types of datasets	Omics field	Num. of patients	Num. of features
Gene Expression	Genomics	553	20531
DNA Methylation	Methylomics	388	1046
miRNA expression	Transcriptomics	413	5000
Survival Data (Clinical)	-	626	164

2.2. Data pre-processing

Data pre-processing is very important for integrative analyses towards multi-omics to reduce unnecessary biases and noises [10]. The retrieved omics dataset has been pre-processed previously. However, in this study, we recheck the presence of missing values in the dataset. The patient survival data is a representation of two different class labels labelled "0" and "1".

Patients who are occupied with "Primary Tumour" samples are marked with "1" while patients having "Solid Tissue Normal" samples are marked "0".

Class imbalance issues were also monitored throughout the data preparation process, whereby the initial class count showed that only 35 patients had "Solid Tissue Normal" (10.17 %) while the other 309 patients (89.83 %) had "Primary Tumour" (Refer to Table 2). To overcome this issue which may lead to unequal representation of classes and misclassifications, we proposed SMOTE as an oversampling technique, whereby minority class gets oversampled by creating synthetic examples instead of oversampling with unfamiliar replacement [11]. The presence of class labels in the form of survival data encourages supervised cancer classification and aids the process of predicting omics features that contributes the most to analyses of patients with LUSC.

Table 2. Details of classes in survival data (Clinical).

Data type	**Class**	
Survival Data (Clinical)	Primary Tumour	Solid Tissue Normal
Number of samples	309	35

The pre-processed omics data which were splitted into a ratio of 75% was used for training while the remainder of the 25% was used for testing. Additionally, data normalisation was also carried out in this phase using min-max normalisation method. The major goal of normalisation process was to scale every feature into the range of 0 to 1. It becomes imperative to carefully use these pre-processing steps as they have a huge influence on the integrative analysis [12].

2.3. Multi-omics integration

Analysis of multi-omics has become increasingly popular among professionals from the biomedical sector [13]. In conjunction with that, integrating multiple layers of information has been hugely challenging for researchers due to data heterogeneity [14]. To conduct the data integration of multi-omics, the integration strategy used in this study was early integration. This strategy was implemented after data has been pre-processed. Then, the pre-processed data were concatenated and the resulting data was formed into a single large matrix before having fed into model training. The early integration is common due to its simplicity, ease of implementation and mainly its ability to reveal the effects of interactions between the different layers by combining variables from each omics [15]. Table 3 summarises the number of samples and features after data integration.

Table 3. Summary of the number of samples and features after data integration.

Types of datasets	**Num. of sample**	**Num. of features**
Multi-omics	344	18663

Pandas, a Python library has been utilised to concatenate all the three omics relatively to each other. This was done to ensure that a complete multi-omics dataset comprising of patients with all three omics was produced. "Patient ID" has been used as the index or target data whereby only patients whose data was presented at all 4 data types, including patient survival (clinical) data, were eligible for further analysis.

2.4. Feature Selection

The multi-omics dataset consisted of a large volume of data that could hike the computational cost rate. Therefore, we decided to undergo feature selection which has successfully reduced the dimensionality and complexity of the multi-omics dataset. In general, feature selection is a technique majorly used for dimensionality reduction, which is necessary for several fields such as machine learning, pattern recognition, statistics and even data mining [16]. Feature selection performs subset selection based on feature relevance and consistency [16].

In this study, we choose the SVM-RFE method which is a recursive feature reduction technique that utilises SVM weights as a criterion for ranking [17]. SVM-RFE's key concept is to remove features that have the lowest squares of weight in each of the iteration. SVM-RFE technique searches for a subset of features by going around all the features available to be iterated and finally removing a certain amount of features, leaving the desired number of data available. For this study, a non-linear SVM-based RFE has been implemented to eliminate the least relevant features in predicting target variables.

Input: Training data $\{x_i, y_i\}\ _{i=1}^{N}$

Output: Ranked feature list R

Initialize: $S = \{1, 2, \dots, D\}_i$

$R = \emptyset$

While S is not empty, do:

1. Restrict the features of x_j to the remaining S
2. Get weight vectors by training SVM
3. Compute the ranking criteria $c_k = w_k^2, k = 1, \dots, |S|$
4. Find features with lowest value of c_k, called feature p
5. Add feature p into R ($R = \{p\}\ R$)
6. Remove feature p from S ($S = S \backslash p$)

Algorithm: Feature Selection based on SVM-RFE

The procedure that has been carried out in our case study was highly similar to the recursive process of the feature removal method in general:

1. Train classifier to find the weight vector (w).

2. Calculate criteria of ranking for all features.

3. Dispose features with the lowest rating criterion value.

Features used in the iteration had to be removed with backward feature elimination. The ranking score is given according to the components of SVM's weight vector w:

$$w = \sum_k a_k y_k x_k \quad (1)$$

In our case, multi-omics dataset was aggregated into two portions termed "features" and "targets". This step is taken to ensure SVM-RFE based feature selection is made without interfering with independent variables such as class labels during the biological data ranking process. Table 4 below shows the number of the multi-omics feature before and after the implementation of the feature selection technique.

Table 4. Implementation of SVM-RFE technique for Feature Selection.

	Before SVM-RFE	**After SVM-RFE**
Total features	18663	12000

2.5. Stacked Denoising Autoencoder (SDAE) and Variational Autoencoder (VAE) Implementation

The significance of this study is in the development of two deep learning approaches that are capable classify cancer using multi-omics data. The models used in this study were SDAE and VAE. The models were designed to analyse the contribution of integrated multi-omics data in identifying cancer patients, easing early diagnosis of LUSC. SDAE model was designed with an early stopping with a patience rate of 50 and the loss rate monitoring was set to ensure that the SDAE model development stops training when the monitored metric has stopped improving. The architecture of DAE was implemented with 50% Gaussian noise with a batch size of 4 and LASSO (L1) kernel regulariser. The input dimension of DAE was 12,000 features and the output dimension was 500, along with the Rectified Linear Unit (ReLU) activation function and adaptive moment estimation (ADAM) optimiser. Later, SDAE with 7 layers of dimensions of 12,000, 10,000, 8,000, 6,000, 4,000, 2,000 and 500, were built with similar parameters as DAE. The performance of this model was observed and analysed at the end of the experimental works.

Next, an input of 12,000 features has been fed into the input layer, with the learning rate set to 0.0005 for the VAE model. The VAE model was built with the compression of the input layer into a mean and log variance vector of the size latent dimension set into 100. Each

layer was initialised with glorot uniform weights and each step includes dense connections, batch normalisation and ReLU activation funnelled separately. Both encoder and decoder layers were implemented with ridge regression (L2) to add a squared magnitude of coefficient as a penalty term to the loss function. Each vector of latent dimension length was connected to the omics input tensor. The hidden layer took two Keras layers as input to the custom sampling function layer with a latent dimension output. Meanwhile, the decoding layer contained a single layer and sigmoid activation function. ADAM optimiser and Binary Cross entropy loss rates were observed as the model stores trained layer weights and reconstructed multi-omics features. Lastly, the performance of this model to classify cancer was also analysed and compared to SDAE.

2.6. Classifier Analysis

The model accuracy for SDAE and VAE models in cancer classification was calculated through training and testing/validation phases. The percentage of the correctly assigned multi-omics features into its proper classes, namely benign or malignant tumours was measured and the result obtained was well assessed. Both models' training and testing phases require different sets or portions of data to avoid overfitting. Integrated multi-omics data was split into training and testing data before being fed into the model. The efficiency of both deep learning architectures was evaluated by accuracy, sensitivity, precision, F1-score, model loss, specificity and MCC.

3. Results

To identify the best performing deep learning model in integrated multi-omics data, both existing SDAE and developed VAE models were compared and analysed in detail to obtain a firm justification, supporting our research objectives. Both models were accessed using a similarly structured model known as Artificial Neural Network (ANN). Table 5 shows the comparative analysis of SDAE and VAE models based on several evaluation metrices as described below.

Table 5. Performance measurement of both autoencoder models.

Deep Learning Architecture	Accuracy (%)	Sensitivity (%)	Precision (%)	F1-Score (%)	Model Loss (%)	Specificity (%)	MCC (%)
SDAE	43.97 %	59.48%	28.86 %	37.17%	0.69 %	29.31%	-0.05%
VAE	91.86 %	98.48 %	93.18 %	95.19 %	0.25 %	22.73%	0.21%

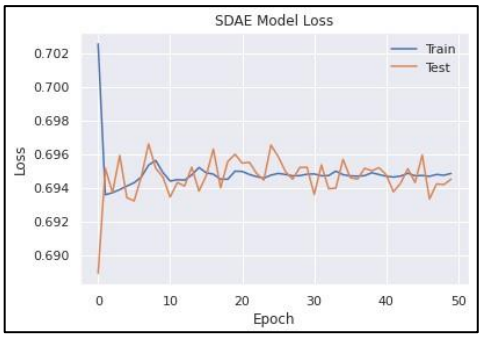

Figure 2. SDAE Model Loss

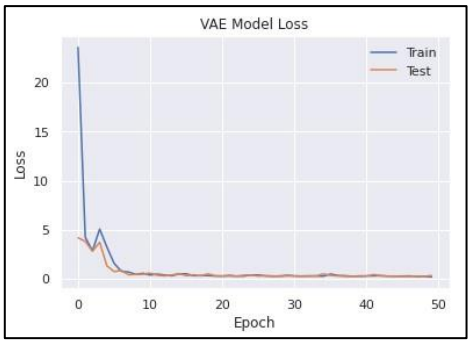

Figure 3. VAE Model Loss

4. Discussion

The above Table 5 is made up of analysis done on two models (SDAE and VAE) implementing similar models (ANN), data preparation and hyper-parameters including optimisers, loss functions, dropouts batch size as well as epochs. Integrated multi-omics data (12,000 features) were fed into both models accordingly to identify how these models interpret and reconstruct these features accurately, leading to an optimal reconstruction rate. VAE model has outperformed SDAE model in terms of an overall performance rate. VAE model has achieved 91.86% accuracy, while SDAE only managed to produce 43.97%. Meanwhile, SDAE produce 59.48% in sensitivity while VAE produce 98.48%. The sensitivity indicates the proportion of actual positive cases (cancerous patient) which is correctly predicted as positive. This means VAE performed well in classifying the majority class label the LUSC patients. Moreover, the SDAE precision was 28.86% while the VAE model precision was 93.18%. Precision defines the number of positive cases (cancerous patient) over the total number of misclassified and correctly classified positive cases (cancerous patient). In this case, the SDAE model has performed poorly in classifying cancerous patients. The F1-score which depended on the weighted average of sensitivity and precision scored 37.17% for the SDAE model while for the VAE model, the score was 95.19%.

This shows that VAE has a better classification and feature reconstruction rate compared to SDAE. Figure 2 and 3 shows the loss rate for both models in training and testing. The VAE model managed to produce a lesser loss rate (0.25%) compared to SDAE (0.69%). However, the specificity value for SDAE was higher (29.31) than VAE (22.73%). In addition, the MCC value for VAE was 0.21% because it was closer to 1 than the SDAE MCC value, which produces a negative value of -0.05%. In the MCC score, when the proposed model lies at +1, it shows a good model. This indicates that the VAE model extracts more information from multi-omics features in comparison to the SDAE model. Confusion matrix-based elements can be considered additional supporting factors for our claim showing that VAE was better than SDAE in handling complex multi-omics data.

5. Conclusions

In conclusion, the performance of SDAE and VAE models were compared to extract meaningful features from integrated multi-omics data that enable the classification of LUSC cells among patients. As a result, VAE has outperformed SDAE in producing excellent accuracy and minimal model loss rate. It was useful to use the weights of the VAE model to extract genes that were also useful for cancer prediction and possess the potential as biomarkers or therapeutic targets. One major drawback of these deep learning approaches is the requirement for large data sets, which are limited for certain cancer tissue. With that, the performance of both models will improve and reveal more meaningful patterns when more omics data becomes available. Together, deep learning models are highly scalable to large input data. Therefore, future studies should analyse different cancer types to identify cancer-specific biomarkers. Moreover, there is also a high potential to identify cross-cancer biomarkers through the analysis of aggregated heterogeneous cancer data.

Author Contributions: Conceptualization, methodology, AAS, NS, VS, HAM, NSA; literature search NSA, NS, AAS, VS, ZAS; experiment, result analysis and validation, VS, HAM, NS, NSA, NHW; writing—original draft preparation, VS, AAS, NSA, NS; writing—review and editing, AAS, NSA, NS, CWH; proofread, HAM, AAS, CWH, ZAS.

Funding: This work was funded by the Malaysian Ministry of Higher Education through the Fundamental Research Grant Scheme (Grant Number: FRGS/1/2018/ICT02/UTM/02/11).

Acknowledgments: The authors would like to express gratitude to the Malaysian Ministry of Higher Education for the financial sponsorship of this study through the Fundamental Research Grant Scheme (Grant Number: FRGS/1/2018/ICT02/UTM/02/11). The study is also supported by the Faculty of Computing, Universiti Teknologi Malaysia.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Duan R, Gao L, Gao Y, Hu Y, Xu H, Huang M et al. Evaluation and comparison of multi-omics data integration methods for cancer subtyping. PLOS Computational Biology. 2021;17(8):e1009224.

2. Sompairac N, Nazarov P, Czerwinska U, Cantini L, Biton A, Molkenov A et al. Independent Component Analysis for Unraveling the Complexity of Cancer Omics Datasets. International Journal of Molecular Sciences. 2019;20(18):4414.

3. Manzoni C, Kia D, Vandrovcova J, Hardy J, Wood N, Lewis P et al. Genome, transcriptome and proteome: the rise of omics data and their integration in biomedical sciences. Briefings in Bioinformatics. 2016;19(2):286-302.

4. Wu D, Wang D, Zhang M, Gu J. Fast dimension reduction and integrative clustering of multi-omics data using low-rank approximation: application to cancer molecular classification. BMC Genomics. 2015;16(1).

5. LeCun Y, Bengio Y, Hinton G. Deep learning. Nature. 2015;521(7553):436-444.

6. Li Y, Huang C, Ding L, Li Z, Pan Y, Gao X. Deep learning in bioinformatics: Introduction, application, and perspective in the big data era. Methods. 2019;166:4-21.

7. Xu A, Chen J, Peng H, Han G, Cai H. Simultaneous Interrogation of Cancer Omics to Identify Subtypes With Significant Clinical Differences. Frontiers in Genetics. 2019;10.

8. Pomraning K, Kim Y, Nicora C, Chu R, Bredeweg E, Purvine S et al. Multi-omics analysis reveals regulators of the response to nitrogen limitation in Yarrowia lipolytica. BMC Genomics. 2016;17(1).

9. Multi-Omic Cancer Benchmark [Internet]. Acgt.cs.tau.ac.il. 2022 [cited 26 September 2022]. Available from: http://acgt.cs.tau.ac.il/multi_omic_benchmark/download.html

10. Wang D, Gu J. Integrative clustering methods of multi-omics data for molecule-based cancer classifications. Quantitative Biology. 2016;4(1):58-67.

11. Chawla N, Bowyer K, Hall L, Kegelmeyer W. SMOTE: Synthetic Minority Over-sampling Technique. Journal of Artificial Intelligence Research. 2002;16:321-357.

12. Luo J, Wu M, Gopukumar D, Zhao Y. Big Data Application in Biomedical Research and Health Care: A Literature Review. Biomedical Informatics Insights. 2016;8:BII.S31559.

13. Vasaikar S, Straub P, Wang J, Zhang B. LinkedOmics: analyzing multi-omics data within and across 32 cancer types. Nucleic Acids Research. 2017;46(D1):D956-D963.

14. Zhang W, Li F, Nie L. Integrating multiple 'omics' analysis for microbial biology: application and methodologies. Microbiology. 2010;156(2):287-301.

15. Hasanin T, Khoshgoftaar T, Leevy J, Seliya N. Investigating Random Undersampling and Feature Selection on Bioinformatics Big Data. 2019 IEEE Fifth International Conference on Big Data Computing Service and Applications (BigDataService). 2019;346-356.