UTM UNIVERSITI TEKNOLOGI MALAYSIA
INTERNATIONAL JOURNAL OF INNOVATIVE COMPUTING
ISSN 2180-4370
Journal Homepage : https://ijic.utm.my/

Classifying Sarcoma Cancer Using Deep Neural Networks Based on Multi-Omics Data

Nur Sabrina Azmi, Azurah A Samah*, Hairudin Abdul Majid, Zuraini Ali Shah, Haslina Hashim & Nuraina Syaza Azman
School of Computing, Artificial Intelligence and Bioinformatics Group (AIBIG),
Faculty of Engineering,
Universiti Teknologi Malaysia,
81310 UTM Johor Bahru, Johor, Malaysia

Ezzeddin Kamil Mohamed Hashim
Biomedicine Programme, School of Health Sciences,
Universiti Sains Malaysia (USM),
16150 Kubang Kerian, Kelantan, Malaysia

Submitted: 17/1/2022. Revised edition: 9/3/2022. Accepted: 27/3/2022. Published online: 16/5/2022
DOI: https://doi.org/10.11113/ijic.v12n1.360

***Abstract*—The challenge in classifying cancer may lead to inaccurate classification of cancers, especially sarcoma cancer since it consists of rare types of cancer. It is hard for the clinician to confirm the patient's condition because the specialist pathology can only make an accurate diagnosis. Therefore, instead of a single omics being used to identify the disease marker, integrating these omics to represent multi-omics brings more advantages in detecting and presenting the phenotype of the cancers. Nowadays, the advancement of computational models, especially deep learning, offered promising approaches in solving high-level omics of data with faster processing speed. Hence, the purpose of this study is to classify cancer and non-cancerous patients using Stacked Denoising Autoencoder (SDAE) and One-dimensional Convolutional Neural Network (1D CNN) to evaluate which algorithm classifies better using high correlated multi-omics data. The study employed both computational models to fit the multi-omics dataset. Sarcoma omics datasets used in this study was obtained from the Multi-Omics Cancer Benchmark TCGA Pre-processed Data of ACGT Ron Shamir Lab repository. The results obtained for the SDAE was 50.93% and 52.78% for the 1D CNN. The results show 1D CNN model outperformed SDAE in classifying sarcoma cancer.**

***Keywords*—Multi-omics analysis, Cancer classification, Neural Network, Stacked Denoising Autoencoder (SDAE), One-dimensional Convolutional Neural Network (1D CNN)**

I.INTRODUCTION

From the early days of computers being built until now, these intelligent machines gained more and more attention, especially when we are in the fourth industrial revolution since the data itself has become an integral part of our lives. As the volume of data is growing every day, it becomes a barrier for academia and researchers to understand the meaning of the data, especially biomedical data. However, the capability of big data has created a bigger opportunity for health care research in drug invention, advanced treatment, customized medicine that can improve patient diagnosis and cost-effectively (Adibuzzaman *et al.*, 2018). Every data is treated and handled differently based on the field of study and the degree of the perspective of the study.

The rise of technology and evolution of human studies with the initiation from Genome-Wide Association Study (GWAS) and Next Generation Sequencing (NGS) produce a new field of study, which is bioinformatics. Bioinformatics is an interdisciplinary field mainly involving molecular biology and genetics, computer science, mathematics, and statistics (Can,2013), where they applied computational methods to analyze and interpret biological data. It provides a broad understanding based on the correlated mechanisms such as transcriptome, genome, epigenome, and proteome profiling (Hu *et al.*, 2018). In the medical era, health and disease can be distinguished using multi-omics.

The advanced growth of data in medical applications in clinical settings and healthcare affect the bioinformatics field to analyze those data. As the gap of knowledge increases, many researchers are trying to transform the data into more understandable so that it can provide more useful information. One of the methods is profiling multi-omics able to administer advantageous understanding and guidance for curative targets and the development of biomarkers by integrating analysis.

Comprehensive characterization of patients using multi-omics methods was the advanced revolution assembled by the Human Genome Project (HGP) and powered by academia and biotechnology companies (Goodwin *et al*., 2016).

The main component of multi-omics is the single omics itself. The omics indicate the molecular properties of an organism based on a biological study such as metabolomics, proteomics, transcriptomics, genomics, and epigenomics (Chung *et al*., 2018). The multi-omics approach represents the relationship between omics and diseases based on the interrelation of omics. Integrated omics deliver more insight, such as biological pathways or different processes between the disease and control groups (Hasin *et al*., 2017, Manzoni *et al*., 2018). The study of omics debriefs millions of markers with similar biochemical properties where single omics contribute limited observation of disease while multi-omics improved understanding of disease etiology and molecular function from different omics levels (Sun *et al*., 2017).

According to the National Cancer Institute, in 2018, an estimated 609,640 out of 1,735,350 people will die because of cancer in the United States. The report concluded that deaths majorly came from cancers. As a precaution, an efficient diagnosis needs to be accurate to detect these cancerous diseases so that proper treatment can be given to these patients. With the advancement of technologies, deep learning models are favored by others because of the capability to represent the internal features in the form of a high-level model problem (Bacciu *et al*., 2018). The large size of multi-omics data results in high bias and variance to the results, especially in classification (van Karnebeek *et al*., 2018). Therefore, deep learning approaches corporate well with multi-omics data to classify cancers.

II. Previous Works

The rapid evolution of machine learning on multi-omics analysis offers a comprehensive biological complexity and intrinsic correlation between omics layers significantly in cancers. Despite other fields having almost entirely been digitized, pathologists still heavily rely on analogue technologies, such as microscopes, glass slides, and written reports to diagnose sarcoma. Hence, the researchers developed several computational models that implement artificial intelligence in pathological sarcoma management, focusing on omics-based data features. There are three commonly deep learning models used such as deep neural network (DNN), recurrent neural networks (RNN) and CNN.

In short, DNN is a fully connected neural network that is categorized into three main techniques such as multilayer perceptron (MLP), autoencoder (AE), and deep belief network (DBN). Unfortunately, MLP, AE and DBN application purely in sarcoma studies is limited due to the highly heterogeneous group with various subtypes. However, the performance of AE compared to MLP and DBN are prominent in general. AE ignores the data noise to reduce data dimension and extract vital biological information. Besides, AE has been acknowledged to analyze high-dimensional gene expression data (Chen *et al*., 2016, Khalili *et al*., 2016) and integrate heterogeneous data (Miotto *et al*., 2016, Chen *et al*., 2016). Apart from denoising autoencoder (DAE), SDAE promises improvement in the model performance (Vincent *et al*., 2010). SDAE is efficient to abstract high layer features such as multi-omics.

In medical imaging applications, the performance of CNNs is not debatable. Since the medical images often consist of high multi-dimensional, CNN reduces the dimension of data representing bind input images to classification. In conclusion, several previous works of deep learning architecture in previous studies are in Table I. Thus, we propose to implement and compare the performance of two different algorithms: SDAE and CNN, with multi-omics sarcoma cancer data.

TABLE I. SUMMARY OF PREVIOUS WORKS BY PAST RESEARCHER

Architecture	**Previous Works**
DNN	● Protein structure prediction (Lyons *et al*., 2014; Heffernan *et al*., 2015) ● Gene expression regulation (Leung *et al*., 2014; Zhang *et al*., 2018) ● Anomaly classification (Rasool *et al*., 2018)
MLP	● Predict the effects of the genetic variants using MLP and SAE (Xie *et al*., 2018)
AE	● SAE - Protein amino acid sequences (Heffernan *et al*., 2015) ● Stacked sparse autoencoder (SSAE) - The classification of births observations (Fergus *et al*., 2015)
DBN	● Amino acid sequences (Kesh and Raghupathi, 2004) ● Protein secondary structure prediction (Lyons *et al*., 2014) ● Breast Histopathology Images (Beevi *et al*., 2017)
RNN	● Identify non-coding RNAs (ncRNAs) (Hill *et al*., 2004)
CNN	● Biosensors (Page *et al*.,2014) ● Gene expression regulation (Leung *et al*., 2014; Lee and Yoon, 2015) ● Transcriptomic (Jurman *et al*., 2017)

III. Materials and Method

The methodology design starts with data pre-processing, where redundant, outliers and missing data are handled before integrating multi-omics to enrich raw data into desired output for analysis (refer Fig. 1). The next step is selecting highly correlated features with the targeted class, then reducing dimension of the dataset and training SDAE and 1D CNN with the dataset. Lastly, the testing and evaluation phase is where the accuracy performance of the model in classification technique compares both methods in classifying sarcoma cancer. All methodology steps below are done using Python programming language in the Google Colaboratory environment.

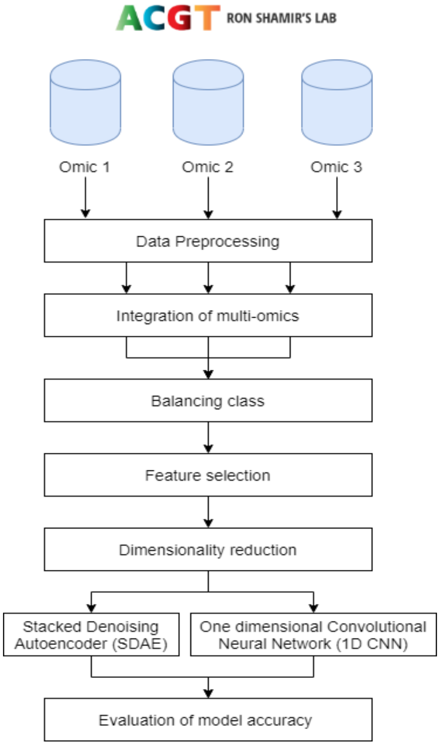

ACGT RON SHAMIR'S LAB
Omic 1
Omic 2
Omic 3
Data Preprocessing
Integration of multi-omics
Balancing class
Feature selection
Dimensionality reduction
Stacked Denoising Autoencoder (SDAE)
One dimensional Convolutional Neural Network (1D CNN)
Evaluation of model accuracy

Fig. 1. The methodology of the study

A. *Omics Datasets*

In this study, the omics datasets are obtained from the ACGT Ron Shamir Lab website and downloaded from the Multi-Omics Cancer Benchmark TCGA Pre-processed Data (Ron Shamir's lab – Tel Aviv University, 2022). Datasets in the Multi-Omics Cancer Benchmark TCGA Pre-processed Data repository contains 11 different types of cancers, including the clinical data of The Cancer Genome Atlas (TCGA) datasets of patients. Briefly, the dataset for every kind of cancer consists of ribonucleic acid (RNA) expression, deoxyribonucleic acid (DNA) methylation, microRNA and survival patient data of cancer.

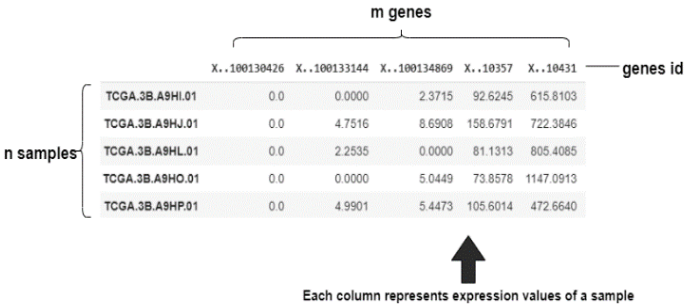
TABLE II. THE TOTAL NUMBER OF FEATURES BASED ON EVERY TYPE OF OMICS

Omics type	**Dataset**	**Number of features**	**Number of samples**
Transcriptomics	Gene expression	20531	265
Genomics	MicroRNA	1046	263
Epigenomics	DNA methylation	5000	269

Throughout this study, three data types are used: gene expression, microRNA, and DNA methylation, representing different types of omics in the sarcoma cancer dataset. The datasets contain various features expression levels based on the dataset type of other numbers of classes. There are four types of classes in the sarcoma cancer dataset: primary tumor, recurrent tumor, solid tissue normal, and metastatic. Table II shows the summary of datasets used in the study.

B. *Data Pre-Processing*

The first crucial step in the methodology design is data pre-processing. This step aims to come up with "clean" and "tidy" datasets that are fitted for successful statistical analysis while avoiding outliers in the data (Salgado *et al*., 2016). Data pre-processing is an iterative process where it needs to be done until the data meet the purpose of the analysis (Salgado *et al*., 2016). A snippet of the gene expression dataset containing gene expression values based on the sample is shown in Fig. 2 below. The microRNA and DNA methylation datasets are similar to gene expression, including the expression values respective features in the omics.


m genes
genes id
n samples
Each column represents expression values of a sample

	X..100130426	X..100133144	X..100134869	X..10357	X..10431
TCGA.3B.A9HI.01	0.0	0.0000	2.3715	92.6245	615.8103
TCGA.3B.A9HJ.01	0.0	4.7516	8.6908	158.6791	722.3846
TCGA.3B.A9HL.01	0.0	2.2535	0.0000	81.1313	805.4085
TCGA.3B.A9HO.01	0.0	0.0000	5.0449	73.8578	1147.0913
TCGA.3B.A9HP.01	0.0	4.9901	5.4473	105.6014	472.6640

Fig. 2. Raw gene expression dataset

Several distinct steps are involved in pre-processing data. The general steps taken to pre-process data are data cleaning, integration, transformation, and reduction (Son *et al*., 2006). Since the datasets obtained from the repository are already pre-processed, the checking needs to be done to ensure no missing, noisy, and redundant data. As a result, there are no missing values in the dataset. Next, transform data into values of a format, scale or unit that is more suitable for analysis (Son *et al*., 2006). The approach used in data transformation is normalization. This method scaled numerical variables in a range of 0 to 1.

C. *Integration of Multi-Omics*

The objective of this phase is to produce a multi-omics dataset from three single omics through the integration process. The integration of multi-omics methods highlights the interrelationships of the functions and the biomolecules involved between omics as different layers are encouraged to explain the biology complexes systematically and holistically (P. Chalise and B. L. Fridley,2017). The integration of multi-omics used in this study is a concatenation method combining the omics based on the patient id by using "merge" python function. As a result, a multi-omics dataset that contains 271 samples with 26577 features are produced, as shown in Table III below.

TABLE III. THE TOTAL NUMBER OF FEATURES AFTER INTEGRATION

Omics type	Number of features	Number of samples
Multi-omics	26577	271

D. Balancing Class

The focus of this study is to create a multi-omics that contain transcriptomics, genomics and epigenomics in the samples. Hence, the samples that did not fulfil the conditions are removed (10 samples). Earlier, the datasets contained four classes, but after data pre-processing is carried out, there is no sample in solid tissue normal class while metastatic class contain only one sample in the multi-omics dataset. The low number of samples in the metastatic class needs to remove due to the huge gap between other classes. Hence, binary classification is done to classify primary tumor and recurrent tumor.

Fig. 3 shows the distribution of class in the multi-omics dataset. The big gap between the samples in both classes is called imbalance class. Imbalance class in multi-omics often occurred in diseases classification (Haas *et al*., 2017). For instance, the most class for hypertension is primary hypertension with 95%, while 5% of endocrine hypertension class shows as minority class (Rimoldi *et al*., 2014). As a result, machine learning with imbalanced data resulting overfitting. Therefore, this study considered over-sampling minority class as the under-represented class compare to other techniques (Reel *et al*., 2021).

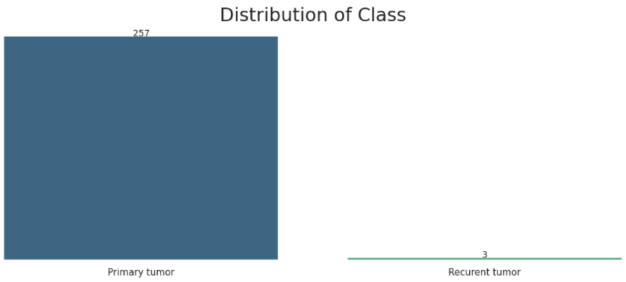


Fig. 3. Graph of distribution of class

The samples in recurrent tumor class are up sampled. The method is called random oversampling, where a common technique to oversample the minority classes to increase the number of minority observations until a balanced dataset has been reached. Thus, both classes contain the same amount of sample (180 samples), as shown in Fig. 4 below.

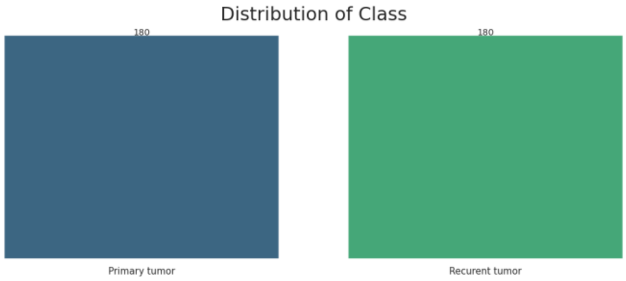


Fig. 4. Graph of distribution of class after random oversampling

E. Feature Selection

Using the support vector machines recursive feature elimination (SVM-RFE) method by manually selecting high correlated features with respective class, 21577 out of 26577 features are selected to rank the variables to be chosen. Around 5000 lowly correlated features in the dataset are removed. The aim of the method is by quantifying the changes in the cost function with presume the estimated value constant. Hence, the retraining variable can be removed to avoid by the classifier. In simple words, when variable with *p* ranked and 1 ranked variable are compared, *p* is considered the least relevant variable. After the method ranked the features, 21577 features in the top rank that has been selected. Table IV below shows the summary of multi-omics dataset after using SVM-RFE.

TABLE IV. THE TOTAL NUMBER OF FEATURES AFTER USING SVM-RFE

Omics type	Number of features	Number of samples
Multi-omics	21577	360

F. Dimensionality Reduction

The most extensive dimensionality reduction method is Principal Component Analysis (PCA) (Ringnér,2008). In machine learning classification problems, high number of features challenges a model to train the data because of the high dimensional space. PCA reduces the number of random variables by obtaining a set of principal variables with minimal loss information. PCA finds the directions of the most significant variance in the data set and represents each data point by its coordinates along each of these directions. PCA produce several principal components (PC) which the sequence of PC contains the highest variance. As the number of PC increases, the variance of the dataset represented decreases.

Higher variability of components is captured in the first principal component (PC1). The second principal component (PC2) captures more information than the third, and so on. In this study, PCA identified only PC1 and PC2, representing the multi-omics dataset's highest variance. Hence, PC1 and PC2 are further used in this study as features. The summary of multi-omics data after using PCA is shown in Table V below. However, the total amount of information in PCs is 33.71% which do not represent 90% of the dataset.

TABLE V. PCA DIMENSIONALITY REDUCTION RESULT

Omics type	Number of features	Number of samples
Multi-omics	2	360

Figure 5 shows the visualization of PCA where the red dot cluster represents the primary tumor while green dots represent the recurrent tumor. The cluster of primary tumor datasets is gathered, while recurrent tumor datasets are dispersed instead of clustering together, affecting classification performance since lower information is captured.

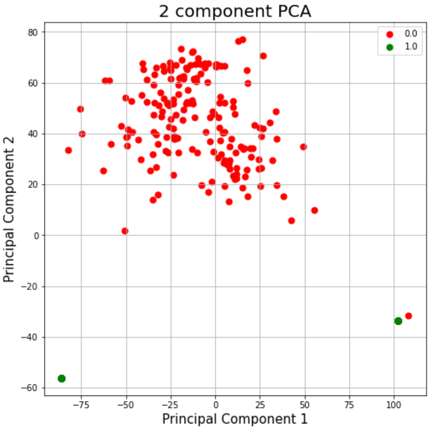


Fig. 5. The graph contains PC1 and PC2

G. *Stacked Denoising Autoencoder*

Using the hyperparameters as shown in Table VI, the pre-train part of SDAE is built. The input data refers to reduced dimensionality of multi-omics dataset. With 0.5 gaussian noise are added to the multi-omics to produce a corrupted input. The special about DAE is that it trained the model to reconstruct the corrupted version of input data without noise applied to it earlier in the process. Hence, DAE is considered as a feature extraction layer in a pre-training layer where the result of DAE is the output for fully connected layer and classifier layer.

TABLE VI. HYPERPARAMETER OF PRE-TRAIN

Hyperparameter setting	
Gaussian noise	0.5
Layers	[2, 1]
Epoch	400
Optimizer	Adamax
Activation	ReLU
Loss	MSE

The activation function used in the autoencoder (encoder and decoder) is Rectified Linear Unit (ReLU) Activation Function. ReLU is a famous activation practiced the most in neural networks, including CNN (Chandra and Sharma,2016). The combination of sigmoid-based or tanh-based with ReLU in deep networks allows the optimization and learning to be easily done if the number of layers and number of nodes are huge (Upadhyayula and Venkataramanan,2020). After a model of SDAE developed in the Google Colaboratory environment, the model is trained using a testing dataset. With 400 epochs, the multi-omics dataset is trained in the model. The loss function is generally used in regression using mean squared error (MSE).

SDAE consists of two phases which are pre-train and fine-tuning. After pre-train the model, the fine-tuning of the model is conducted. Fine-tuning is the most important phase to evaluate the classification of multi-omics cancer data. Fine-tuning helps the model be fine-tuned to minimize the error in predicting the class trained by the training dataset. After encoders stack training, top of the stack is added with an output layer. Hyperparameters of fine-tuning such as activation, optimizer and loss are assigned with parameters as describe in Table VII.

As for the sigmoid activation function chosen in the study is because the output produced within range of 0 until 1 (Upadhyayula and Venkataramanan,2020). Binary classification used in the study are 0 that represent “Primary tumor” and 1 that represent “Recurrent tumor”. Hence the sigmoid function is the most suitable activation function to predict the classes in the dataset as the output. In short, sigmoid activation function is used as the classifier layer to predict 0 to 1 range of the target class. The purpose of the implementation of the “Adam” optimizer is to handle sparse gradients on highly noised data such as multi-omics, while “binary_crossentropy” loss suits binary classification where the target values are in the set between 0 and 1.

TABLE VII. TABLE HYPERPARAMETER OF FINE-TUNING

Hyperparameter setting	
Epoch	100
Activation	Sigmoid
Batch	16
Optimizer	Adam
Loss	Binary_crossentropy

Based on the epoch, we can conclude that the performance measurement of the accuracy of SDAE is 50.93%, with 0.0493 loss. The loss value based on the epoch decreases as the epoch increases while model accuracy increases. Since the model learns from the loss function of the algorithm, it is evaluated on the performance of the method based on given data. If predictions deviate too much from actual results, loss function will cough up a large number. Gradually, with the help of some optimization function, the loss function learns to reduce the error in prediction.

The model testing concludes the performance of the SDAE model by obtaining 50.93% of accuracy, 50.93% of sensitivity and 67.48% of F1-Score. The summary of SDAE model accuracy and loss is shown in Table VIII below.

TABLE VIII. TABLE FINAL RESULTS OF SDAE MODEL LOSS AND ACCURACY

Model	Model accuracy	Model loss
SDAE	50.93%	70.94%

H. *One Dimensional Convolutional Neural Network*

The second model of the proposed method is 1D CNN. It used one dimensional for the input because the multi-omics dataset only consists of the same type of expression (numerical) values in the data structure. The hyperparameters implemented in 1D CNN is shown in Table IX.

TABLE IX. TABLE HYPERPARAMETER OF 1D CNN MODEL

Hyperparameter setting	
Filter	64
Kernel size	4
Activation	Sigmoid
Loss	mse
Optimizer	adam

Fig. 6 shows the flow of the architecture of the 1D CNN model from the input layer until the classifier layer.

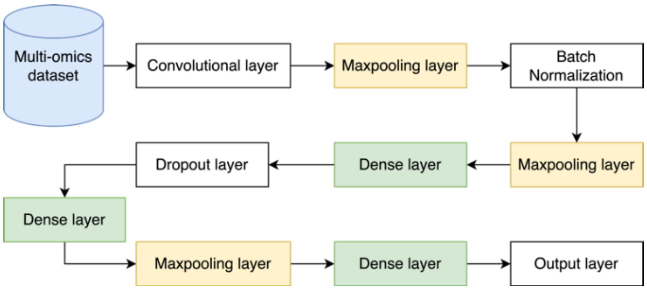


Fig. 6. The architecture of the proposed 1D CNN model

The details of each layer are discussed below:

a) Input data: Input data is obtained from the result of PCA. Next, the matrix must be initialized for the *n x m* value. The n refers to the total number of features which is 26577, while m refers to the dimension 1.

b) Convolutional layer: Apply filters to the input to generate the feature maps or the activation maps using the sigmoid activation function. According to Fig. 6, the function filters refer to feature maps which are 64. Feature maps refer to the number of times the inputs are integrated whole kernel size is 4. Kernel size is the number of input time steps considered as the input sequence is read onto the feature maps.

c) Max pooling layer: To reduce the complexity of output and prevent overfitting upon feature map by creating a new set of the same number of pooled feature maps separately by taking the maximum value of the feature map. The input of this layer is the convolutionized multi-omics dataset. Pooling involves selecting a pooling operation, like a filter applied to feature maps. The pooling operation or filter size is smaller than the size of the feature map.

d) Batch Normalization: This layer reduces the amount by what the hidden unit values shift around (covariance shift). It also allows each layer of a network to learn a little bit more independently of other layers.

e) Dense layer: A dense layer is just a regular layer of neurons in a NN where each neuron receives input from all the neurons in the previous layer, thus densely connected.

f) Dropout layer: Weights in the dropout layer are randomly assigned. Since we chose a rate of 0.4, 40% of the neurons will receive a zero weight. The network becomes less sensitive to react to smaller variations in the data. Therefore, it should further increase our accuracy on unseen data.

g) Dense layer with sigmoid activation: Reduce the vector of height j to a vector of two since we have two classes that we want to predict (“Primary tumor”, “Recurrent tumor”). The activation function used in this study is sigmoid because it forces all two outputs of the neural network to sum up to one. Therefore, the output value will represent the probability for each of the two classes.

As the result of the designed architecture of the proposed 1D CNN method, the model is compiled using the loss function of *mse* and *adam* optimizer. The model testing concludes the accuracy of the 1D CNN by obtaining 52.78% of accuracy, 52.78% of sensitivity and 69.09% of F1-Score. The summary of 1D CNN model accuracy and loss is shown in Table X below.

TABLE X. TABLE RESULTS OF 1D CNN MODEL LOSS AND ACCURACY

Model	Model accuracy	Model loss
1D CNN	52.78%	52.78%

IV. Result And Discussion

Before training the model, the multi-omics dataset is separated into training and testing datasets. Separations of the dataset are divided into 70% training and 30% for testing with actual 252 and 108 samples, respectively. Data partitioning is done to train the model with several training data and then tested with testing data to predict the class based on the training dataset. Table XI below shows the data partitioning of multi-omics data for both models.

TABLE XI. TABLE DATA PARTITIONING

Dataset	Data Division		Total
	Training (70%)	Testing (30%)	
Multi-omics	252	108	360

A graph comparing the accuracy and loss of both models is shown in Fig. 7 below. Based on the result, 1D CNN outperforms SDAE with 52.78% accuracy with 1.85% slight difference in the accuracy result. From the observation, both models do not produce over 90% accuracy, as the researchers and literature claim. Several possible answers lead to lower accuracy performance. Firstly, the omics dataset. According to the PCA result, the representation of the dataset on PCA is only 33.71%. The good representation of data needs to acquire more than 90% to represent the datapoint in PCA. This shows that the variation of information is lower.

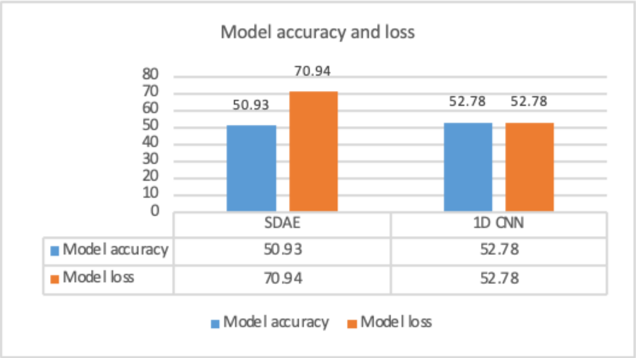


Fig. 7. Graph of SDAE and 1D CNN accuracy

We conclude that the accuracy of the models maintains between 50% during evaluating the testing dataset while model loss based on the epoch of the model between training and testing dataset overlap with each other. However, a good representation of data classified in the model must produce smaller loss values, but in this SDAE model, the loss function consists of a bit high of loss values and not decreasing along with the epoch.

V. Conclusion

The study is an exercise to explore the model performance of SDAE and 1D CNN using multi-omics data. The study concludes that 1D CNN outperforms SDAE. To the best of our knowledge, limited works of sarcoma data (Ron Shamir's lab – Tel Aviv University, 2022) restrict our analysis. We believe that the main component of this study is data pre-processing, which impacts the performance of both models. The implementation of a simple concatenate-based method used in this study to integrate multi-omics data does not consider the regulatory relationship between omics. Hence, the integrated omics result in more complex, noisy and no interrelation omics. Moreover, two approaches of dimension reduction are carried out: feature selection using SVM-RFE and feature extraction using PCA. There is a possibility of removing essential features in the approaches from the multi-omics dataset. Thus, further study is proposed to produce higher accuracy where several limitations are discussed to be improved.

Acknowledgment

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

References

[1] Adibuzzaman, M., Delaurentis, P., Hill, J. and Benneyworth, D. (2018). Big Data in Healthcare – The Promises, Challenges and Opportunities from a Research Perspective: A Case Study with a Model Database. 384-392.

[2] Bacciu, D. (2018). Bioinformatics and Medicine in the Era of Deep Learning, 25-27.

[3] Beevi, K., Nair, M., and Bindu, G. (2017). A Multi-Classifier System for Automatic Mitosis Detection in Breast Histopathology Images Using Deep Belief Networks. *IEEE Journal of Translational Engineering in Health and Medicine*, 5, 1-11. Doi: 10.1109/jtehm.2017.2694004.

[4] Can, T. (2013). miRNomics: MicroRNA Biology and Computational Analysis.

[5] Chandra, B. and Sharma, R. K. (2016). *Fast Learning in Deep Neural Networks, Neurocomputing.* Elsevier, 171, 1205-1215.

[6] Chen, Q., Song, X., Yamada, H., and Shibasaki, R. (2016). Learning Deep Representation from Big and Heterogeneous Data for Traffic Accident Inference. *Proceedings of the AAAI Conference on Artificial Intelligence*, 30(1).

[7] Chen, L., Cai, C., Chen, V., and Lu, X. (2016). Learning a Hierarchical Representation of the Yeast Transcriptomic Machinery using an Autoencoder Model. *BMC Bioinformatics*, 17(S1). Doi: 10.1186/s12859-015-0852-1.

[8] Chung, R.-H. and Kang, C.-Y. (2018). A Multi-omics Data Simulator for Complex Disease Studies and Its Application to Evaluate Multi-omics Data Analysis Methods For Disease Classification. Oxford University Press, 35, 1-24.

[9] Goodwin, S., Mcpherson, J. D. and Mccombie, W. R. (2016). Coming of Age: Ten Years of Next-generation Sequencing Technologies, Nature Publishing Group. Nature Publishing Group, 17(6), 333-351.

[10] Haas, R., Zelezniak, A., Iacovacci, J., Kamrad, S., Townsend, S., and Ralser, M. (2017). Designing and Interpreting 'Multi-omic' Experiments that May Change Our Understanding of Biology. *Current Opinion in Systems Biology*, 6, 37-45. Doi: 10.1016/j.coisb.2017.08.009.

[11] Hasin, Y., Seldin, M. and Lusis, A. (2017). Multi-omics Approaches to Disease, Genome Biology. Genome Biology, 18(1), 1-15.

[12] Heffernan, R., Paliwal, K., Lyons, J., Dehzangi, A., Sharma, A., and Wang, J. *et al.* (2015). Improving Prediction of Secondary Structure, Local Backbone Angles and Solvent Accessible Surface Area of Proteins by Iterative Deep Learning. *Scientific Reports*, 5(1). Doi: 10.1038/srep11476.

[13] Hu, Y., An, Q., Sheu, K., Trejo, B., Fan, S. and Guo, Y. (2018). Single Cell Multi-Omics Technology: Methodology and Application, *Frontiers in Cell and Developmental Biology*.

[14] Hill, S., Kuintzle, R., Teegarden, A., Merrill, E., Danaee, P., and Hendrix, D. (2018). A Deep Recurrent Neural Network Discovers Complex Biological Rules to Decipher RNA Protein-Coding Potential. *Nucleic Acids Research*, 46(16), 8105-8113. Doi: 10.1093/nar/gky567.

[15] Jurman, G., Maggio, V., Fioravanti, D., Giarratano, Y. and Landi, I. (2017). Convolutional Neural Networks for Structured Omics: OmicsCNN and the OmicsConv Layer, 1-7.

[16] Khalili, M., Majd, H.A., Khodakarim, S., Ahadi, B., and Hamidpour, M. (2016). Prediction of the Thromboembolic Syndrome: An Application of Artificial Neural Networks in Gene Expression Data Analysis. *Journal of Paramedical Sciences*, 7, 15-22.

[17] Lee, T. and Yoon, S. (2015). Boosted Categorical Restricted Boltzmann Machine for Computational Prediction of Splice Junctions. *32nd International Conference on Machine Learning, ICML 2015*, 3, 2473-2482.

[18] Leung, M., Xiong, H., Lee, L., and Frey, B. (2014). Deep Learning of the Tissue-regulated Splicing Code. *Bioinformatics*, 30(12), 121-129. Doi: 10.1093/bioinformatics/btu277.

[19] Lyons, J., Dehzangi, A., Heffernan, R., Sharma, A., Paliwal, K., and Sattar, A. *et al.* (2014). Predicting Backbone Cα Angles and Dihedrals from Protein Sequences by Stacked Sparse Auto-encoder Deep Neural Network. *Journal of Computational Chemistry*, 35(28), 2040-2046. Doi: 10.1002/jcc.23718.

[20] Manzoni, C., Kia, D. A., Vandrovcova, J., Hardy, J., Wood, N. W., Lewis, P. A. and Ferrari, R. (2018). Genome, Transcriptome and Proteome: The Rise of Omics Data and Their Integration in Biomedical Sciences. *Briefings in Bioinformatics*, 19(2), 286-302.

[21] Miotto, R., Li, L., Kidd, B., and Dudley, J. (2016). Deep Patient: An Unsupervised Representation to Predict the Future of Patients from the Electronic Health Records. *Scientific Reports*, 6(1). Doi: 10.1038/srep26094.

[22] P. Chalise and B. L. Fridley. (2017). Integrative Clustering of Multi-Level 'Omic Data Based on Non-negative Matrix Factorization Algorithm. *PLoS One*, 12(5), 1-18.

[23] Page, A., Turner, J. T., Mohsenin, T. and Oates, T. (2014). Comparing Raw Data and Feature Extraction for Seizure Detection with Deep Learning Methods. *Proceedings of the 27th International Florida Artificial Intelligence Research Society Conference, FLAIRS 2014,* 284-287.

[24] Kesh, S., and Raghupathi, W. (2004). Critical Issues in Bioinformatics and Computing. *Perspectives in Health Information Management*, 1, 9.

[25] Rasool, F., Faisal, L., Azade, N. and Manfred, H. (2018). Using Deep Learning to Enhance Head and Neck Cancer Diagnosis and Classification, *2018 IEEE International Conference on System, Computation, Automation and Networking, ICSCA 2018*, June 2013.

[26] Reel, P. S., Reel, S., Pearson, E., Trucco, E., and Jefferson, E. (2021). Using Machine Learning Approaches for Multi-omics Data Analysis: A Review. *Biotechnology Advances,* 49 (December 2020), 107739. https://doi.org/10.1016/j.biotechadv.2021.107739.

[27] Rimoldi, S., Scherrer, U., and Messerli, F. (2013). Secondary Arterial Hypertension: When, Who, and How to Screen? *European Heart Journal*, 35(19), 1245-1254. Doi: 10.1093/eurheartj/eht534.

[28] Ringnér, M. (2008). What is Principal Component Analysis? *Nature Biotechnology*, 26(3), 303-304. Doi: 10.1038/nbt0308-303.

[29] Ron Shamir's lab – Tel Aviv University. (2022). Retrieved 19 January 2022, from http://acgt.cs.tau.ac.il/.

[30] Salgado, C. M., Azevedo, C., Proença, H. and Vieira, S. M. (2016) Secondary Analysis of Electronic Health Records, Springer Nature.

[31] Son, N. H. (2006). Data Mining Course—Data Cleaning and Data Pre-Processing. Warsaw University. Available at URL http://www.mimuw.edu.pl/*son/datamining/DM/4-preprocess.pdf.

[32] Sun, Y. V and Hu, Y. (2017). Integrative Analysis of Multi-omics Data for Discovery and Functional Studies of Complex Human Diseases Yan.

[33] Upadhyayula, S. M. and Venkataramanan, K. (2020). Deep Learning Using Neural Networks. 49-61.

[34] van Karnebeek, C. D. M., Wortmann, S. B., Tarailo-Graovac, M., Langeveld, M., Ferreira, C. R., van de Kamp, J. M., Hollak, C. E., Wasserman, W. W., Waterham, H. R., Wevers, R. A., Haack, T. B., Wanders, R. J. A. and Boycott, K. M. (2018). The Role of the Clinician in the Multi-omics Era: Are You Ready? *Journal of Inherited Metabolic Disease*, 41(3), 571-582.

[35] Vincent, P., Larochelle, H., Lajoie, I., Bengio, Y., and Manzagol, P. (2010). Stacked Denoising Autoencoders: Learning Useful Representations in a Deep Network with a Local Denoising Criterion. *J. Mach. Learn. Res.*, 11, 3371-3408.

[36] Xie, R., Wen, J., Quitadamo, A., Cheng, J., and Shi, X. (2017). A Deep Auto-encoder Model for Gene Expression Prediction. *BMC Genomics*, 18(S9). Doi: 10.1186/s12864-017-4226-0.

[37] Zhang, L., Lv, C., Jin, Y., Cheng, G., Fu, Y., and Yuan, D. *et al.* (2018). Deep Learning-Based Multi-omics Data Integration Reveals Two Prognostic Subtypes in High-Risk Neuroblastoma. *Frontiers in Genetics*, 9. Doi: 10.3389/fgene.2018.00477.