We are IntechOpen, the world's leading publisher of Open Access books
Built by scientists, for scientists

6,900
Open access books available

186,000
International authors and editors

200M
Downloads

Our authors are among the

154
Countries delivered to

TOP 1%
most cited scientists

12.2%
Contributors from top 500 universities

CLARIVATE ANALYTICS
BOOK CITATION INDEX
INDEXED

WEB OF SCIENCE™

Selection of our books indexed in the Book Citation Index in Web of Science™ Core Collection (BKCI)

Interested in publishing with us?
Contact book.department@intechopen.com

Numbers displayed above are based on latest data collected.
For more information visit www.intechopen.com

Chapter 17

Deep Learning Models for Predicting Phenotypic Traits and Diseases from Omics Data

Md. Mohaiminul Islam, Yang Wang and Pingzhao Hu

Additional information is available at the end of the chapter

http://dx.doi.org/10.5772/intechopen.75311

Abstract

Computational analysis of high-throughput omics data, such as gene expressions, copy number alterations and DNA methylation (DNAm), has become popular in disease studies in recent decades because such analyses can be very helpful to predict whether a patient has certain disease or its subtypes. However, due to the high-dimensional nature of the data sets with hundreds of thousands of variables and very small number of samples, traditional machine learning approaches, such as support vector machines (SVMs) and random forests, have limitations to analyze these data efficiently. In this chapter, we reviewed the progress in applying deep learning algorithms to solve some biological questions. The focus is on potential software tools and public data sources for the tasks. Particularly, we show some case studies using deep neural network (DNN) models for classifying molecular subtypes of breast cancer and DNN-based regression models to account for interindividual variation in triglyceride concentrations measured at different visits of peripheral blood samples using DNAm profiles. We show that integration of multi-omics profiles into DNN-based learning methods could improve the prediction of the molecular subtypes of breast cancer. We also demonstrate the superiority of our proposed DNN models over the SVM model for predicting triglyceride concentrations.

Keywords: deep learning, omics data, phenotypic traits, data integration, support vector machine, random forest

IntechOpen

© 2018 The Author(s). Licensee IntechOpen. This chapter is distributed under the terms of the Creative Commons Attribution License (http://creativecommons.org/licenses/by/3.0), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

1. Introduction

1.1. Omics data

Omics refers to use high-throughput experimental technologies to examine genomics, transcriptomics, metabolomics and proteomics for understanding biological and disease mechanisms. The omics data generated from these technologies are high-dimensional and correlated. Different computational and statistical analyses of these data can be used to identify risk factors for different diseases or to build autonomous disease prediction models. The technological development allows researchers to have a huge amount of high-dimensional biological data. The omics technologies generate such high-throughput data by detecting numerous alterations in molecular components [1]. These technologies also generate additional biological data to comprehend different types of correlations and dependencies among the molecular components. Bioinformatics is a discipline which emerges to perform computational analysis with the high-throughput biological data. Bioinformatics offers tools and methodologies for analyzing different omics data to understand the underlying information about different diseases. Such analyses will help physicians to provide early and patient-specific treatment. Schneider and Orchard [2] listed the state-of-the-art available technologies to generate omics data. They also provided the list of different available bioinformatics resources to analyze omics data and discussed the bioinformatics challenges to handle the high-throughput data.

For example, one of the key questions in medical science is to identify the specific mutations for a particular disease in individual patients. To do this, we need to first collect blood samples from the patients with the disease and healthy individuals without diseases. DNA will be then extracted from these blood samples. DNA microarray (also named as DNA chip) or DNA sequencing technologies will be then used to generate genome-wide genomic data for identifying mutated genes for the disease [2]. Similar technologies and procedures can also be used to generate other types of omics data such as RNA gene expression data and DNA methylation (DNAm) data.

1.2. Phenotypic traits prediction

A living biological organism can show a number of observable characteristics such as morphology, growth and behavior of the organism. Phenotypes are the product of different genetic expressions of an organism. These expressions are known as the genotype of that organism. However, phenotypic traits are the alternatives of a phenotype of a particular organism. For example, hair is a phenotype but different hair colors are the phenotypic traits. Study of the phenotypic traits prediction is very important as it gives us the knowledge about how genotype impacts upon an individual's diseases or traits. Lippert et al. [3] used the whole-genome sequencing data to identify individuals by predicting their biometric traits. Genome sequencing data were also used by Chen et al. [4] to build a probabilistic Bayesian model to predict the dichotomous traits (e.g., glaucoma, Corn's disease, prostate cancer). This model incorporates annotated information about different variant genotypes and genes, which are associated with diseases. There are other phenotypic trait prediction models such as eye color [5, 6], skin color [6] or facial structures [7].

1.2.1. Breast cancer and its molecular subtypes

Cancer is a collection of diseases that are characterized by uncontrolled cell growth in organs, that is, the sites the cells originate from. For example, breast cancer begins in the breast tissue and may start in the duct or lobe of the breast. When the breast cells are not working properly, they divide continually and a tumor is formed. Breast cancer is a complex and heterogeneous disease with differing prognostic and clinical outcomes.

In clinical practice, estrogen-receptor expression can be used to classify breast cancer patients into estrogen-receptor positive (ER+) or negative (ER−). A patient is classified as ER+ if the stimulation of growth of her cancer cells depends on the receptor for hormone estrogen. Stratification of breast cancer patients into ER+ or ER− is very important as physicians will use the information to determine whether the patients need chemotherapy or hormonal treatments. Statistics show that approximately 67% of breast cancer tests are positive for hormone receptors [8].

Perou et al. discovered the intrinsic subtypes of breast cancer using gene expression profiles of frozen tissue samples through unsupervised analysis [9]. They classified breast cancers into five groups such as normal-like, luminal A and B (Lumb A and B), basal-like and Her2-enriched (Her2). They established a polymerase chain reaction (PCR)-based test for 50 genes, and these genes are well-known as PAM50 signature. Gene expression levels of these 50 genes can be used to classify patients into one of these four subtypes: luminal A and B, basal-like and Her2-enriched. These subtypes are also known as PAM50 subtypes. This classification has been shown to be prognostically independent of clinicopathologic factors and can identify patients who are more likely to benefit from adjuvant chemotherapy [10]. It can also help identifying the fundamental differences among the PAM50 subtypes at the molecular level [11].

1.2.2. Prediction of triglyceride concentration in blood

Triglyceride is a type of fat in the human blood. Having a high concentration of triglycerides in human blood can increase our risk of heart diseases, stroke and other disorders. Many genetic loci have been identified by genome-wide association studies, but only a small proportion of interindividual variability of triglycerides has been explained by the genetic determinants. It is known that the level of triglycerides is heritable. Consequently, the development of new high-throughput genomic technologies makes it natural to extend these phenotypic prediction models to complex traits such as triglyceride. Using DNAm profiles to predict disease phenotypic courses has not yet been explored in detail.

1.3. Machine learning for disease phenotype prediction

Artificial intelligence (AI) is an area of computer science which demonstrates its necessity in our everyday life by machine learning (ML) methods. ML methods can automate the data analysis and can find the hidden intrinsic patterns from big data, which is impossible for a human being. ML methods use these patterns to build predictive models without any explicit programming. These predictive ML models are improving our daily life in various ways such as recommendations of different products during online shopping based on our searches of products, stock price prediction, classification of different objects from images, real-time language translation, and so on.

Conventional ML methods, such as support vector machine (SVM), random forest (RF), Bayesian network (BN), and so on, are dependent on the well-defined, engineered and robust hand-tuned features (or feature vectors) as inputs from the raw input data to make reasonable predictions. A domain human expertise is required to develop these engineered features. However, real-time biomedical data are often high-dimensional and noisy. These conventional ML methods are not capable enough to provide suitable techniques to handle such natural raw data (i.e., normalized gene expression data).

Different ML-based methods were developed to classify breast cancer patients into one of the PAM50 subtypes using gene expression profiles [12]. However, a new class of ML methods called deep learning (DL) can handle such high-dimensional, noisy and natural raw data by following representation learning or hierarchical data-driven approaches.

2. Deep learning and its application in bioinformatics

DL is a family of artificial neural network (ANN)-based ML methods which have been inspired by the working principles of a human brain. In a DL network architecture, a series of hidden layers are connected in a cascade fashion between input and output of the network. Each of these layers takes input from its previous layer and transforms the data into a more abstract form. Nonlinear layers allow DL methods to model complex relations between input and output of the network like shallow ANNs. DL is a representation learning method which means it can be fed with raw data and then, it will automatically extract necessary representation for predictions. A DL network provides representations at different levels. The output of each of hidden layers is considered as the representation at that level. The higher layers the data belong, the more abstract representations we get for these data. In different studies, these higher-level representations of raw data prove to be very effective for classification or detection problems. The most important thing here is that these representations, alternatively called feature vectors, are not learned by human engineering rather from the raw input data directly.

Unlike other ML methods, DL methods have been shown to efficiently handle high-dimensional and noise data in many domains such as computer vision, language processing (**Table 1**). These qualities of DL attract biomedical researchers to use DL instead of conventional ML methods because biomedical data (e.g., omics data) often suffer from high-dimensionality and noisiness.

2.1. How deep learning evolved?

Technological advancement and availability of large data sets allow researchers to rekindle their interest in deep neural networks. Recently, DNN-based models have achieved the state-of-the-art prediction performances at the cost of immense computational power. For example, Krizhevsky et al. [13] trained a deep convolutional neural network (DCNN)-based model with 1.2 million images to classify 1000 different classes, which took approximately 6 days to complete the training. They built this model by optimizing about 60 million learnable parameters. They won the ILSVRC-2012 competition with their state-of-the-art prediction performances in image classification.

Characteristics	Conventional machine learning (including shallow neural network)	Deep learning
Feature engineering	Handcrafted features created by experts are required	Perform automatic feature extraction
Problem-solving approach	Break the problems into small parts and solve them separately. Combine the results for the final output.	Can solve the problems in an end-to-end fashion
Execution time	Relatively much less time is required (ranging from a few seconds to a few hours)	Training requires long time for optimizing its thousands of parameters.
Interpretability	Relatively easy to interpret the reasoning	Hard to interpret the reasoning
High-throughput data	It is challenging to handle high-throughput data directly	Techniques like dropout enable high-throughput data processing relatively possible

Table 1. Comparison of conventional machine learning and deep learning.

In computer vision, there has always been a growing need to train visual recognition systems more generically so that a system trained on one visual recognition task (e.g., classification) could be easily adapted to another task (e.g., detection). To handle such a challenge of adapting the source domain to different target domains, many domain adaptation methods have been proposed [14–16]. For example, Donahue et al. [17] extracted "deep features" from a deep neural network trained on one computer vision task and shown the state-of-the-art performance on a variety of other tasks. The DCNN-based features are very powerful, generic and, thus, can be well adapted to different visual domains, which was also evidenced by Sharif Razavian et al. [18]. They proposed a DCNN-based pretrained model called OverFeat, which was applied to different tasks for which OverFeat was not trained. They achieved state-of-the-art prediction performances for object detection [19]. After having feature descriptors extracted from the first fully connected layer of OverFeat, they applied a linear SVM classifier to these features for image classification, scene recognition, fine-grained recognition and attribute detection on different data sets. A pretrained CNN is usually followed by domain-specific fine-tuning on data from the target domains, especially when training data are scarce. Following this approach, Girshick et al. [20] fine-tuned a CNN pretrained on ILSVRC2012 classification data set and achieved substantially better object detection performance on PASCAL VOC as compared to the standard models based on simple hand-engineered features.

DL-based models trained using a small data set usually show poor prediction performances over the test data. To tackle such kind of overfitting problem, a technique called dropout was proposed by Hinton et al. [21]. Dropout is a regularization technique, which randomly drops out a predefined ratio of neuron connections within a DL-based network's layer. This helps the model to learn a general weight for each of the neuron connections of that layer. This procedure allows the network to perform well over the test data. In addition, another technique called DropConnect was proposed as a generalization of dropout [22]. It randomly drops out of neuron connections from the network instead of randomly dropping neural connection from a layer. DropConnect provided improved prediction performances for different image recognition benchmark data sets over the dropout [22].

There is another class of DNN known as deep belief network (DBN). A DBN is composed of multiple layers where different layers have neural connections between them but not within the hidden units of a layer. Each of these layers is a restricted Boltzmann machines (RBMs). DBN follows an unsupervised layer-wise pretraining approach for these RBMs using a method called contrastive divergence [23]. DBN is useful to fine-tune a DNN when the number of training samples is small. In this case, a DBN is first trained and the optimized weights from this DBN are then used to fine-tune a DNN. Therefore, this DNN will start the training with these learned weights instead of training from the scratch. This leads the network to be convergent at early stage and improve prediction performances because these weights are neighboring to the best value of a converged model [24].

Stacked autoencoder is another type of DL-based method, which is used to create a vigorous high-level representation of its input using unsupervised ML approaches [24]. A stacked autoencoder can be built using a DNN by stacking layers on top of others. With this approach, we can increase or decrease the dimension of the input data. Vincent et al. showed that we can feed a high-level representation of input data learned from a stacked autoencoder into a SVM with improved prediction performance [25].

DL-based methods have already achieved state-of-the-art prediction performances in diverse fields such as image classification [26], object detection [13], speech recognition [27], and so on. However, DL methods also allow us to build state-of-the-art prediction models for sequential data [28, 29]. A series of functionally powerful deep learning tools have been implemented (**Table 2**), which can be publicly downloaded to perform these applications.

Software	Website	Open Source	Interface	Deep learning algorithms
Caffe	http://caffe.berkeleyvision.org/	Yes	Python, MATLAB	CNN, RNN
Deeplearning4j	https://deeplearning4j.org/	Yes	Java, Python	CNN, RNN, RBN, DBN
Keras	https://keras.io/	Yes	Python, R	CNN, RBN, DBN
Theano	http://deeplearning.net/software/theano	Yes	Python	CNN, RNN, RBN, DBN
Torch	http://torch.ch/	Yes	Lua, LuaJIT, C, OpenCL	CNN, RNN, RBN, DBN
Tensorflow	https://www.tensorflow.org/	Yes	Python, Java	CNN, RNN, RBN, DBN
MatConvNet	http://www.vlfeat.org/matconvnet/	Yes	C++, MATLAB	CNN, RNN, RBN, DBN
MXNet	https://mxnet.incubator.apache.org/	Yes	R, C++, Python, Julia, Scala, MATLAB	CNN, RNN, RBN, DBN

Table 2. A list of commonly used deep learning tools.

CNN, convolutional neural network; RNN, recurrent neural network; RBN, radial basis network; and DBN, deep brief network.

2.2. Application of deep learning to solve different bioinformatics applications

In recent years, deep learning has been successfully applied to answer many biological questions using diverse biological data sources (**Figure 1**). We briefly review these advancements in different bioinformatics applications in the following paragraphs.

Maienschein-Cline et al. used a conventional unsupervised machine learning approach to identify targeted genes which were regulated by transcription factors (TFs) by integrating high-throughput binding data and gene expression data [30]. Their experimental results showed that the conventional method faced challenges to handle such high-throughput data even though they may be very helpful for clinical diagnosis. Therefore, Dananee et al. used a DL-based method called stacked denoising autoencoder (SADE) [25] for such kind of analysis [31]. They first took a high-level representation of the input gene expression profiles. They then fed this higher-level representation of the data to a shallow neural network and a SVM. They demonstrated the improved prediction performances than the baseline principal component analysis (PCA) or kernel principal component analysis (KPCA) approaches. They also identified a set of genes from the connectivity matrices of their DL-based model. These genes can be investigated further to improve clinical diagnosis.

Somatic point mutation-based cancer classification (SMCC) becomes an attractive research problem as DNA sequencing technology allows us to generate a huge volume of sequencing data. Efficient analysis of the somatic point mutation data can lead to a better patient-specific personal therapy to cancer patients. This kind of somatic point mutation data usually suffer from large sparsity. Therefore, previous SMCC approaches were not able to provide clinically acceptable cancer classification results. However, recently a DL-based classifier known as DeepGene tried to overcome this limitation [32]. This model first filtered the gene data by mutation rate to remove irrelevant genes from them. Then, it indexed the gene data by their nonzero elements which let DeepGene overcome the data sparsity problem. Finally, the outputs of these two steps were fed into a DNN which performed automatic extraction of features for SMCC. DeepGene achieved ~67% prediction accuracy which is much better than the prediction performances of most baseline classifiers (i.e., SVM ~67%, k-nearest neighbors (KNN) ~42% and naïve Bayes (NB) ~9%).

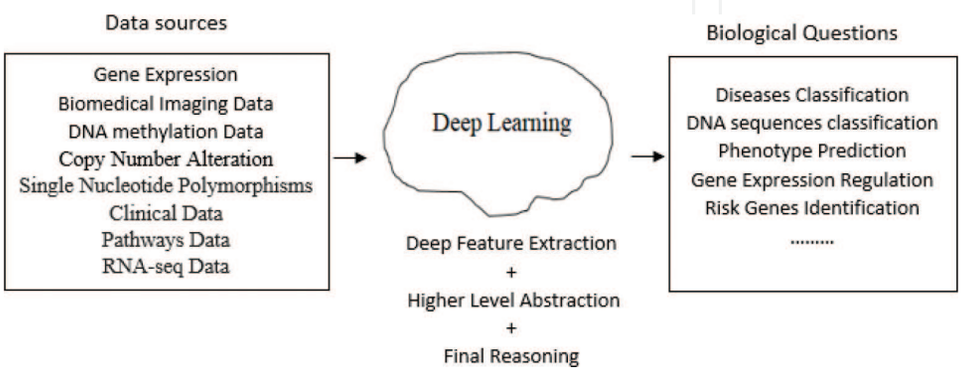

Figure 1. Scopes of deep learning in bioinformatics.

DBN was used to cluster cancer patients by integrating their gene expression and clinical data [33]. This model can capture intra- and cross-modality correlations (i.e., correlation among genomic data from different platforms) and learn a unified representation of the input. Therefore, this model outperformed existing methods in clustering cancer patients. This model can also be used to predict missing values in the data and identify key target genes of miRNAs responsible for different cancer subtypes. Moreover, preliminary clinical screening of a patient with skin disease usually begins with a visual diagnosis by a dermatologist. Since this is a very common malignancy in a human being [34, 35], an automatic system to classify skin diseases will be very helpful for the clinical purpose. Kelley et al. introduced a DCNN model to learn the functional activity of DNA sequences for 164 cell-specific DNA accessibility multitask prediction, and this model achieved the best result among earlier methods [36]. Recently, Esteva et al. collected 129,450 clinical images of skin diseases and built a DCNN model to classify the disease [37]. This model achieved better prediction performance than dermatologists. Nonetheless, a large number of nuclei and the variability in their sizes in histopathological images of breast cancer pose a great difficulty to build an automated system for nucleus detection. Xu et al. overcame this challenge by using a deep learning approach called stacked sparse autoencoder (SSAE). This model outperformed nine previous state-of-the-art nuclear detection methods [38].

Conventional machine learning approaches have been applied to analyze high-content microscopy data to protein subcellular localization from yeast cell images [39]. However, these approaches were not able to perform such analysis without human expert's intervention and yet did not provide accurate classification. Kraus et al. [40] came up with a model called DeepLoc which is a DCNN-based approach to overcome these limitations. DeepLoc outperforms the model ensLOC [39] by 71.4% according to mean average precision using fewer number of images. However, ensLOC uses binary SVM ensemble approach to assign single cells to subcellular compartment classes. Kraus et al. [40] also investigated the reason behind their success over ensLOC by performing 2D visualization of their network's components. They found out that DeepLoc generates a unique signal for different inputs. The structure of a protein and its functions can be studied further by protein contact map prediction from sequences. Wang et al. [41] treated this problem as a pixel-level labeling by considering a protein contact as an image. They proposed a novel deep learning-based protein contact map prediction model with extremely unbalanced positive and negative labels. Their model integrates two evolutionary couplings (EC) and sequence conservation information into their network. Their model gives the state-of-the-art performance result in protein contact map prediction. Furthermore, the predicted protein contacts by this model can generate an improved 3D structure model than the previous best models: CCMpred [42] and MetaPSICOV [43]. Besides, many biological processes such as signal transduction and cellular organization can be affected by different protein-protein interactions (PPI). Hence, it is very important to build a PPI prediction model in order to provide a better design for the therapy of a disease. Sun et al. [44] were the first one to build a deep learning-based model that is a stacked autoencoder for the sequence-based PPI prediction. They achieved an accuracy of 97.19% with 10-fold cross-validation which is better than any existing PPI predictors.

Genomics becomes rich with many different types of functional genomic data because of the latest sequencing technology. Eser et al. introduced flexible integration of data with deep learning

(FIDDLE), which is an open source DCNN-based data integration framework [45]. FIDDLE can predict yeast transcription start site sequencing (TSS-seq) [46] by the integration of heterogeneous genomic data such as RNA-seq and DNA sequence data. The excellent performance of FIDDLE demonstrated that a model built on integrating multiple data sets can provide better prediction performances than the models which were built using only one data set.

Chen et al. [47] proposed a deep learning system (D-GEX) which takes a gene's expression profile as input and infers the expression profile of a target gene. D-GEX has the ability to show cross-platform generalization. This model archives 15.33% improvement in gene expression prediction than a linear regression approach. D-GEX proves its cross-platform generalization when the learned D-GEX is used in RNA-seq-based database for gene expression prediction for each target gene and still outperforms LR by 6.57%.

Existing methods for the classification of cellular phenotypes from cellular images consist of multiple steps. Each of these steps is required with manual modifications and the tuning of different parameter settings. Godinez et al. [48] introduced a new multiscale CNN (M-CNN) network which uses microscopic images to classify them into phenotypes. The prediction performances of the M-CNN in terms of accuracy over eight benchmark data sets are significantly higher than the previous state-of-the-art methods including CNN-based approaches.

Gene expression can be regulated using transcription factors (TFs). Therefore, the cell-specific TF binding predictions using gold standard Chip-seq data are very important. Qin and Feng [49] introduced a DNN model termed TFImpute to achieve the abovementioned goal. TFImpute can determine whether a specific TF would bind to a given DNA sequence in a specific cell line. The prediction performance of TFImpute proves its superiority from the comparison with another latest DNN-based approach called DeepBind [50]. Therefore, biologists can use TFImpute to understand how TF binding can be included by a specific cell line.

Zhou et al. [51] were the first one to propose a DCNN-based approach to predict the effects of noncoding-variants from large-scale chromatin-profiling data and achieved state-of-the-art predictive performance. They call their method as deep learning-based sequence analyzer (DeepSea). Experimental results show that DeepSea can also precisely predict the consequence of specific SNPs on TF binding.

Obtaining precise knowledge about a patient's health condition is crucial to provide early and better treatment. Discovery of good imaging biomarkers can lead clinical research into achieving this goal. Oakden-Rayner et al. [52] provided a proof-of-concept research which proves that computer-based cross-sectional chest CT image analysis is able to predict 5 year mortality in adult (age > 60 years) person. Their framework includes deep learning model, and the predictive performances of this model are better than those who use human-generated features. Besides, visualization of different components of this deep learning-based model can provide an explanation about the better prediction performances [53].

Gene expression can be controlled by enhancer elements and cis-acting DNA regulatory elements [54]. However, existing enhancer predictors face a challenge, that is, the lack of availability of huge and experimentally confirmed enhancers for humans or other species. Yang et al. [55] developed a DNN-based hybrid architecture termed as BiRen which takes only DNA sequence as input to predict enhancers. Experimental results proved that BiRen can

predict common enhances more accurately than previous state-of-the-art methods, which are based on DNA sequence only.

Analysis of high-dimensional single-cell RNA-seq data is very important to answer several biological questions such as the amount of heterogeneity of cells in a population, the discovery of a biomarker for explicit cells and retrieving analogous cell types. Lin et al. [56] introduced an NN-based model to address all these queries without integrating any prior knowledge into the model. This method can deduce cell type more properly using a database of tens of thousands of single cell profiles than any existing methods.

Although the significant advancements have been made in applying DNN models to different bioinformatics applications as described earlier, here, we introduce several DNN-based classification frameworks which take either CNA profiles or gene expression profiles or both of them as input for the prediction of molecular subtypes of breast cancer. In addition, we also present a DNN-based regression model which takes high-dimensional DNAm data as input to predict triglyceride concentrations (before and after treatment) in the human blood.

3. Case studies

We discussed three case studies where deep learning-based approaches were used to address two biological problems such as classification of molecular subtypes of breast cancer (**Figure 2(a)** and **(b)**) and prediction of triglyceride concentration in human blood (**Figure 2(c)**).

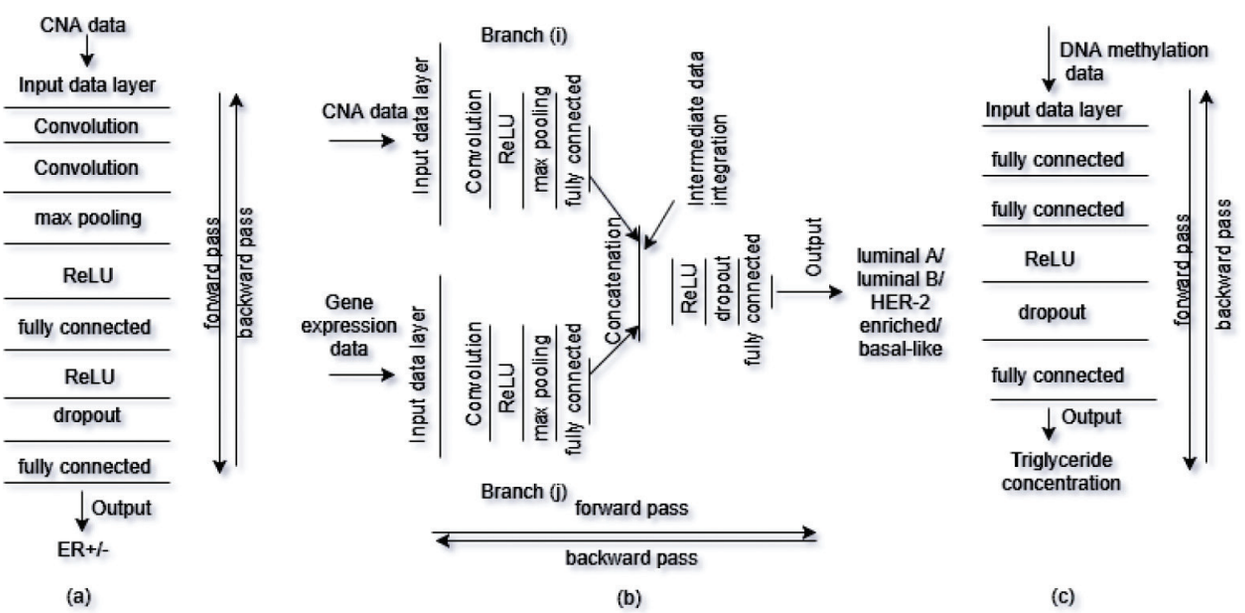

Figure 2. Deep learning architectures of three different case studies: (a) classification of molecular subtypes of breast cancer using single data source; (b) classification of molecular subtypes of breast cancer using multiple data sources; and (c) prediction of triglyceride concentration in human blood.

3.1. Prediction of molecular subtypes of breast cancer

Classification of molecular subtypes of breast cancer using high-throughput genomics data, such as gene expressions and copy number alterations (CNAs), is a challenging task because we have much smaller number of training samples than the number of genomic features. Many traditional machine learning methods often get overfitted over training data to handle such a problem. Here, we explored to use DNN methods to overcome this problem.

We conducted our experiments with two data sets: gene expression and CNA. Both data sets had approximate 2000 patients and larger than 16,000 genes. We collected these data sets from a project called Molecular Taxonomy of Breast Cancer International Consortium (METABRIC) [57]. We built a patient by gene mutation matrix from the CNA data where we had only three different values to represent three different copy number statuses of a gene in a patient: 1 = copy number gain, −1 = copy number loss and 0 = no change.

Using the CNA profiles of 991 and 984 patients as training and test sets, respectively, which was separated in original study [56] based on the patients collected at different periods, we classified the status of estrogen-receptor as a supervised binary classification problem. Furthermore, using the CNA profiles of 935 and 842 patients as training and test sets, respectively, we classified the status of PAM50 subtypes (luminal A, luminal B, HER-2 enriched and basal-like) as a supervised multiclass classification problem (it should be noted that some patients have no status of the PAM50 subtypes). Our DCNN-based deep learning architecture used for the tasks is shown in **Figure 2(a)** [58].

Our network took a CNA profile with approximate 16,000 genes of a breast cancer patient as the input in the input data layer. The input layer passed this input data to a convolutional layer. In the convolutional layer, we passed a kernel with weights over the input data and performed an element-wise multiplication with the input data. Then, we took the sum of the elements from this operation to represent the input data. This operation helped us capture the potential correlations among the neighbor genes. Output from the convolutional layers went as an input into max pooling layer. Here, we also passed a kernel without any weights over the input of this layer. We then took the maximum value from the input data under the kernel to represent the input data. This operation helped the model to achieve three invariant properties such as scaling, translation, and rotation. We introduced nonlinearity into our model by passing the output from max pooling layer through a rectified linear unit (ReLu) layer. ReLu layer performed a thresholding operation by converting all values less than zero to zero. At this point, we passed the output from the ReLu layer to a fully connected layer to get the higher level abstraction. We then passed the output from this fully connected to a dropout layer where we used another ReLu between them. Dropout layer randomly dropped out a number of neurons from the network. Dropout strategy enforced the network to learn a general weight for each of the neurons of the network. This helped the model to achieve better generalization ability over the test data as well as prevented the model from the overfitting problem. Then, we used another fully connected layer to get the prediction score for each of the classes. We converted this class score into class probability using softmax function. We trained our network in backpropagation style.

We evaluated the prediction performances of our network by two metrics such as overall accuracy and area under the curve (AUC). We achieved 84.1% accuracy and 0.904 AUC for binary classification. On the other hand, for multiclass classification, we achieved 58.19% accuracy and 0.79 AUC. We compared our experimental outcomes with other two supervised machine learning methods such as support vector machine (SVM) and random forest (RF). Prediction performances of SVM for the binary classification were 76.5% accuracy and 0.702 AUC while 45.0% accuracy and 0.78 AUC for multiclass classification. With RF, we had 82.7% accuracy and 0.817 AUC for binary classification and 49.5% accuracy and 0.729 AUC for multiclass classification. Our experimental outcomes showed that deep learning-based models have better performance than the traditional machine learning methods such as SVM and RF.

We also explored to build DCNN-based network (**Figure 2(b)**) to classify the status of PAM50 subtypes using both gene expression and CNA data. One of these data sets may have biological knowledge which is absent in another one. Our network shown in **Figure 2(b)** had two branches: i and j branches which took CNA data and the gene expression data of the same patient as inputs, respectively. Each of these branches performed the same series of operations over the input data. Outputs of the last fully connected layers from both branches were concatenated to be fed into the third part of the network which provided the final predictions. Here, we also used overall accuracy and AUC as the performance measurement metric. In terms of prediction performances, our model achieved 79.2% accuracy and 0.85 AUC. We compared our prediction performances with two baseline models: SVM and RF. Their prediction performances in terms of overall accuracy were 69.5 and 70.1%, respectively. Besides, our baseline models provided AUC 0.804 and 0.781, respectively. From these comparisons, we can clearly see that our deep learning-based data integration model (**Figure 2(b)**) achieved improved prediction performances than all of our baseline models. Our data integration model also outperformed the models (**Figure 2(a)**) which were built using only one data source as input to the architecture.

3.2. Prediction of triglyceride concentration

Triglyceride is a kind of fat which is found in human blood. Excessive triglyceride concentration can cause different heart-related diseases such as stroke. In this experiment, we built a DNN-based regression model [59], which took epigenome-wide DNA methylation profiles as input to predict triglyceride concentration at multiple draw of peripheral human blood samples.

For our experiments, we collected the epigenome-wide DNAm profiles and triglyceride concentrations (mg/dL) measured at the baseline level (pretreatment) of visit 2 and changes in response to treatment with fenofibrate (posttreatment) at visit 4 from Genetic Analysis Workshop 20 (GAW20). The DNAm profiles were generated using the Illumina Infinium HumanMethylation450 BeadChip array. The beta value measuring the methylation level was expressed between 0 and 1 in 993 samples of the Genetics of Lipid Lowering Drugs and Diet Network (GOLDN) study. It should be noted that there were only 499 samples with the posttreatment DNAm data. Here, we built three regression models to solve three tasks: (1) we used pretreatment data to predict triglyceride concentration before the medication. We built this model using 900 randomly selected training samples, and we had 93 test samples for

the purpose of testing the model's prediction performances; (2) we used posttreatment data to predict triglyceride concentration after the medication. In this case, we had 400 randomly selected training samples and 99 test samples; (3) we used pretreatment data to predict triglyceride concentration after the medication. Here, we had 620 randomly selected training samples and 94 test samples.

Our DNN-based regression model (**Figure 2(c)**) is a fully connected neural network with four hidden layers between input and output layers. This network took a vector of size larger than 450,000 which represented an epigenome-wide DNA methylation profile of a sample. This vector went through two full connected layers, and the output is the higher level abstraction of the input. Then, we used a ReLU layer to introduce nonlinearity into the model, and the output of ReLU went into a dropout layer to achieve the generalization ability to test the data. The output of dropout layer then went into the final output layer and provided the predicted triglyceride concentration.

We used root mean square error (RMSE) and Pearson correlation coefficient (PCC) to measure our prediction performances. Prediction performances of our DNN-based regression models are as follows: for Task 1, we had RMSE 88.5 and PCC 0.27; for Task 2, we had RMSE 48.1 and PCC 0.22 and for Task 3, we had RMSE 47.4 and PCC 0.29. We used SVM as our baseline model. SVM provided prediction performances are as follows: for Task 1, the RMSE was 90.3 and PCC was 0.13; for Task 2, the RMSE was 48.7 and PCC was 0.19 and for Task 3, the RMSE was 46.9 and PCC was 0.13.

Experimental results showed that our DNN-based triglyceride concentration regression models provided improved prediction performances for all three tasks than the baseline SVM-based regression method. Both our DNN-based and SVM-based regression models achieve best prediction performances when we used the pretreatment data to predict triglyceride concentration after the medication. This outcome shows that there is a long-term epigenetic effect on the phenotypic traits.

4. Conclusion

Classification of molecular subtypes of a disease using omics profiles is a challenging problem since the data sets are quite high-dimensional and highly correlated. The curse of high-dimensionality also affects the performance of predicting a phenotype using DNAm data. Traditional machine learning algorithms, such as SVM and RF, have potential challenges to handle high-dimensional and highly correlated data sets. Recently, DNN learning has been demonstrated advantages over these methods since it does not require any hand-crafted features. DNN learning automatically extracts features from the raw data and efficiently analyzes high-dimensional and correlated data. In this chapter, we reviewed the status of applying DNN in bioinformatics and showed some case studies which introduced several DNN frameworks for classifying molecular subtypes of breast cancer using only one data source or two heterogeneous data sources. In addition, we also presented a DNN-based regression framework which took epigenome-wide DNAm data as input to predict triglyceride concentrations

in human blood. In summary, our and others' works have demonstrated that DNN is a promising tool to predict phenotypic traits and diseases from genome-wide omics data.

Acknowledgements

The authors thank the Molecular Taxonomy of Breast Cancer International Consortium (METABRIC) and the organizer of genetic analysis workshop 20 (GAW20) for providing the data sets used in this study. This work was supported in part by Canadian Breast Cancer Foundation – Prairies/NWT Region, Natural Sciences and Engineering Research Council of Canada, Manitoba Research Health Council and University of Manitoba.

Conflict of interest

The authors declare that they have no competing interests.

Author details

Md. Mohaiminul Islam[1,2,3], Yang Wang[2] and Pingzhao Hu[1,2,3,4]*

*Address all correspondence to: pingzhao.hu@umanitoba.ca

1 Department of Biochemistry and Medical Genetics, University of Manitoba, Winnipeg, Manitoba, Canada

2 Department of Computer Science, University of Manitoba, Winnipeg, Manitoba, Canada

3 George and Fay Yee Centre for Healthcare Innovation, University of Manitoba, Winnipeg, Manitoba, Canada

4 Department of Electrical and Computer Engineering, University of Manitoba, Winnipeg, Manitoba, Canada

References

[1] Knasmüller S, Nersesyan A, Mišík M, Gerner C, Mikulits W, Ehrlich V, Hoelzl C, Szakmary A, Wagner KH. Use of conventional and-omics based methods for health claims of dietary antioxidants: A critical overview. British Journal of Nutrition. 2008;**99**(E-S1):ES3-E52. DOI: 10.1017/S000711450896575

[2] Schneider MV, Orchard S. Omics technologies, data and bioinformatics principles. In: Bioinformatics for Omics Data: Methods and Protocols. 2011. pp. 3-30

[3] Lippert C, Sabatini R, Maher MC, Kang EY, Lee S, Arikan O, Harley A, Bernal A, Garst P, Lavrenko V, Yocum K. Identification of individuals by trait prediction using whole-genome sequencing data. Proceedings of the National Academy of Sciences. 2017 Sep 19;**114**(38):10166-10171

[4] Chen YC, Douville C, Wang C, Niknafs N, Yeo G, Beleva-Guthrie V, Carter H, Stenson PD, Cooper DN, Li B, Mooney S. A probabilistic model to predict clinical phenotypic traits from genome sequencing. PLoS Computational Biology. 2014;**10**(9):e1003825. DOI: 10.1371/journal.pcbi.1003825

[5] Liu F, Wen B, Kayser M. Colorful DNA polymorphisms in humans. Seminars in Cell & Developmental Biology. 2013;**24**(6):562-575. DOI: 10.1016/j.semcdb.2013.03.013

[6] Hart KL, Kimura SL, Mushailov V, Budimlija ZM, Prinz M, Wurmbach E. Improved eye- and skin-color prediction based on 8 SNPs. Croatian Medical Journal. 2013;**54**(3):248-256. DOI: 10.3325/cmj.2013.54.248

[7] Claes P, Liberton DK, Daniels K, Rosana KM, Quillen EE, Pearson LN, McEvoy B, Bauchet M, Zaidi AA, Yao W, Tang H. Modeling 3D facial shape from DNA. PLoS Genetics. 2014;**10**(3):e1004224. DOI:/10.1371/journal.pgen.1004224

[8] Breast Cancer Information and Awareness. Available from: http://www.breastcancer.org. [Accessed: 2017-10-20]

[9] Perou CM, Sørlie T, Eisen MB, van de Rijn M, Jeffrey SS, Rees CA, Pollack JR, Ross DT, Johnsen H, Akslen LA, Fluge Ø. Molecular portraits of human breast tumours. Nature. 2000;**406**(6797):747-752. DOI: 10.1038/35021093

[10] Parker JS, Mullins M, Cheang MC, Leung S, Voduc D, Vickery T, Davies S, Fauron C, He X, Hu Z, Quackenbush JF. Supervised risk predictor of breast cancer based on intrinsic subtypes. Journal of Clinical Oncology. 2009;**27**(8):1160-1167. DOI: 10.1200/JCO.2008.18.1370

[11] Sorlie T, Tibshirani R, Parker J, Hastie T, Marron JS, Nobel A, Deng S, Johnsen H, Pesich R, Geisler S, Demeter J. Repeated observation of breast tumor subtypes in independent gene expression data sets. Proceedings of the National Academy of Sciences of the United States of America. 2003;**100**(14):8418-8423. DOI: 10.1073/pnas.0932692100

[12] Milioli HH, Vimieiro R, Tishchenko I, Riveros C, Berretta R, Moscato P. Iteratively refining breast cancer intrinsic subtypes in the METABRIC dataset. BioData Mining. 2016;**9**(1):2. DOI: 10.1186/s13040-015-0078-9

[13] Krizhevsky A, Sutskever I, Hinton GE. Imagenet classification with deep convolutional neural networks. In: Advances in Neural Information Processing Systems. 2012. pp. 1097-1105

[14] Khosla A, Zhou T, Malisiewicz T, Efros AA, Torralba A. Undoing the damage of dataset bias. In: European Conference on Computer Vision. 2012. pp. 158-171. DOI: 10.1109/CVPR.2011.5995347

[15] Saenko K, Kulis B, Fritz M, Darrell T. Adapting visual category models to new domains. Computer Vision–ECCV. 2010;**2010**:213-226

[16] Aytar Y, Zisserman A. Tabula rasa: Model transfer for object category detection. In: IEEE International Conference on Computer Vision (ICCV); 2011 Nov 6; IEEE. pp. 2252-2259

[17] Donahue J, Jia Y, Vinyals O, Hoffman J, Zhang N, Tzeng E, Darrell T. Decaf: A deep convolutional activation feature for generic visual recognition. In: International Conference on Machine Learning; 2014; p. 647-655

[18] Sharif Razavian A, Azizpour H, Sullivan J, Carlsson S. CNN features off-the-shelf: An astounding baseline for recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops. 2014. pp. 806-813

[19] Sermanet P, Eigen D, Zhang X, Mathieu M, Fergus R, LeCun Y. Overfeat: Integrated recognition, localization and detection using convolutional networks. arXiv preprint arXiv:1312.6229. 2013

[20] Girshick R, Donahue J, Darrell T, Malik J. Rich feature hierarchies for accurate object detection and semantic segmentation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition; 2014; p. 580-587

[21] Hinton GE, Srivastava N, Krizhevsky A, Sutskever I, Salakhutdinov RR. Improving neural networks by preventing co-adaptation of feature detectors. arXiv preprint arXiv:1207.0580; 2012

[22] Wan L, Zeiler M, Zhang S, Cun YL, Fergus R. Regularization of neural networks using dropconnect. In: Proceedings of the 30th International Conference on Machine Learning (ICML-13). 2013. pp. 1058-1066

[23] Hinton GE. Training products of experts by minimizing contrastive divergence. Training. Neural Computation. 2006;**14**(8):1771-1800. DOI: 10.1162/089976602760128018

[24] Larochelle H, Erhan D, Courville A, Bergstra J, Bengio Y. An empirical evaluation of deep architectures on problems with many factors of variation. In: Proceedings of the 24th International Conference on Machine Learning 2007. pp. 473-480

[25] Vincent P, Larochelle H, Lajoie I, Bengio Y, Manzagol PA. Stacked denoising autoencoders: Learning useful representations in a deep network with a local denoising criterion. Journal of Machine Learning Research. 2010;**11**:3371-3408

[26] Russakovsky O, Deng J, Su H, Krause J, Satheesh S, Ma S, Huang Z, Karpathy A, Khosla A, Bernstein M, Berg AC. Imagenet large scale visual recognition challenge. International Journal of Computer Vision. 2015;**115**(3):211-252. DOI: 10.1007/s11263-015-0816-y

[27] Abdel-Hamid O, Mohamed AR, Jiang H, Deng L, Penn G, Yu D. Convolutional neural networks for speech recognition. IEEE/ACM Transactions on Audio, Speech, and Language Processing. 2014;**22**(10):1533-1545. DOI: 10.1109/TASLP.2014.2339736

[28] Ren S, He K, Girshick R, Sun J. Faster r-cnn: Towards real-time object detection with region proposal networks. IEEE Transactions on Pattern Analysis and Machine Intelligence. 2017;**39**(6):1137-1149. DOI: 10.1109/TPAMI.2016.2577031

[29] Collobert R, Weston J, Bottou L, Karlen M, Kavukcuoglu K, Kuksa P. Natural language processing (almost) from scratch. Journal of Machine Learning Research. 2011; **12**:2493-2537

[30] Maienschein-Cline M, Zhou J, White KP, Sciammas R, Dinner AR. Discovering transcription factor regulatory targets using gene expression and binding data. Bioinformatics. 2011;**28**(2):206-213. DOI: 10.1038/srep20649

[31] Danaee P, Ghaeini R, Hendrix DA. A deep learning approach for cancer detection and relevant gene identification. Pacific Symposium on Biocomputing. Pacific Symposium on Biocomputing. 2016;**22**:219. DOI: 10.1142/9789813207813_0022

[32] Yuan Y, Shi Y, Li C, Kim J, Cai W, Han Z, Feng DD. DeepGene: An advanced cancer type classifier based on deep learning and somatic point mutations. BMC Bioinformatics. 2016;**17**(17):476. DOI: 10.1186/s12859-016-1334-9

[33] Liang M, Li Z, Chen T, Zeng J. Integrative data analysis of multi-platform cancer data with a multimodal deep learning approach. IEEE/ACM Transactions on Computational Biology and Bioinformatics (TCBB). 2015;**12**(4):928-937. DOI: 10.1109/TCBB.2014.2377729

[34] American Cancer Society. Cancer facts & figures 2016. Atlanta: American Cancer Society; 2016. Available from: http://www.cancer.org/acs/groups/content/@research/documents/document/acspc047079.pdf. [Accessed: 2017-10-10]

[35] Stern RS. Prevalence of a history of skin cancer in 2007: Results of an incidence-based model. Archives of Dermatology. 2010;**146**(3):279-282. DOI: 10.1001/archdermatol.2010.4

[36] Kelley DR, Snoek J, Rinn JL. Basset: Learning the regulatory code of the accessible genome with deep convolutional neural networks. Genome Research. 2016;**26**(7):990-999. DOI: 10.1101/gr.200535.115

[37] Esteva A, Kuprel B, Novoa RA, Ko J, Swetter SM, Blau HM, Thrun S. Dermatologist-level classification of skin cancer with deep neural networks. Nature. 2017;**542**(7639):115-118. DOI: 10.1038/nature21056

[38] Xu J, Xiang L, Liu Q, Gilmore H, Wu J, Tang J, Madabhushi A. Stacked sparse autoencoder (SSAE) for nuclei detection on breast cancer histopathology images. IEEE Transactions on Medical Imaging. 2016;**35**(1):119-130. DOI: 10.1109/TMI.2015.2458702

[39] Chong YT, Koh JL, Friesen H, Duffy SK, Cox MJ, Moses A, Moffat J, Boone C, Andrews BJ. Yeast proteome dynamics from single cell imaging and automated analysis. Cell. 2015;**161**(6):1413-1424. DOI: 10.1016/j.cell.2015.04.051

[40] Kraus OZ, Grys BT, Ba J, Chong Y, Frey BJ, Boone C, Andrews BJ. Automated analysis of high-content microscopy data with deep learning. Molecular Systems Biology. 2017;**13**(4):924. DOI: 10.15252/msb.20177551

[41] Wang S, Sun S, Li Z, Zhang R, Xu J. Accurate de novo prediction of protein contact map by ultra-deep learning model. PLoS Computational Biology. 2017;**13**(1):e1005324

[42] Seemayer S, Gruber M, Söding J. CCMpred—Fast and precise prediction of protein residue-residue contacts from correlated mutations. Bioinformatics. 2014;**30**(21):3128-3130. DOI: 10.1371/journal.pcbi.1005324

[43] Jones DT, Singh T, Kosciolek T, Tetchner S. MetaPSICOV: Combining coevolution methods for accurate prediction of contacts and long range hydrogen bonding in proteins. Bioinformatics. 2014;**31**(7):999-1006. DOI: 10.1093/bioinformatics/btu79

[44] Sun T, Zhou B, Lai L, Pei J. Sequence-based prediction of protein protein interaction using a deep-learning algorithm. BMC Bioinformatics. 2017;**18**(1):277. DOI: 10.1186/s12859-017-1700-2

[45] Eser U, Churchman LSFIDDLE. An integrative deep learning framework for functional genomic data inference. Cold Spring Harbor Laboratory. Biorxiv. 2016. DOI: 10.1101/081380

[46] Malabat C, Feuerbach F, Ma L, Saveanu C, Jacquier A. Quality control of transcription start site selection by nonsense-mediated-mRNA decay. eLife. 2015;**4**:e06722. DOI: 10.7554/eLife.06722

[47] Chen Y, Li Y, Narayan R, Subramanian A, Xie X. Gene expression inference with deep learning. Bioinformatics. 2016;**32**(12):1832-1839. DOI: 10.1093/bioinformatics/btw074

[48] Godinez WJ, Hossain I, Lazic SE, Davies JW, Zhang X. A multi-scale convolutional neural network for phenotyping high-content cellular images. Bioinformatics. 2017;**33**(13):2010-2019. DOI: 10.1093/bioinformatics/btx069

[49] Qin Q, Feng J. Imputation for transcription factor binding predictions based on deep learning. PLoS Computational Biology. 2017;**13**(2):e1005403. DOI: 10.1371/journal.pcbi.1005403

[50] Alipanahi B, Delong A, Weirauch MT, Frey BJ. Predicting the sequence specificities of DNA-and RNA-binding proteins by deep learning. Nature biotechnology. 2015;**33**(8):831-838. DOI: 10.1038/nbt.3300

[51] Zhou J, Troyanskaya OG. Predicting effects of noncoding variants with deep learning-based sequence model. Nature Methods. 2015;**12**(10):931-934. DOI: 10.1038/nmeth.3547

[52] Oakden-Rayner L, Carneiro G, Bessen T, Nascimento JC, Bradley AP, Palmer LJ. Precision radiology: Predicting longevity using feature engineering and deep learning methods in a radiomics framework. Scientific Reports. 2017;**7**(1):1648. DOI: 10.1038/s41598-017-01931-w

[53] Zeiler MD, Fergus R. Visualizing and understanding convolutional networks. In: European Conference on Computer Vision. 2014. pp. 818-833

[54] Calo E, Wysocka J. Modification of enhancer chromatin: What, how, and why? Molecular Cell. 2013;**49**(5):825-837. DOI: 10.1016/j.molcel.2013.01.038

[55] Yang B, Liu F, Ren C, Ouyang Z, Xie Z, Bo X, Shu W. BiRen: Predicting enhancers with a deep-learning-based model using the DNA sequence alone. Bioinformatics. 2017 Feb 17;**33**(13):1930-1936. DOI: 10.1093/bioinformatics/btx105

[56] Lin C, Jain S, Kim H, Bar-Joseph Z. Using neural networks for reducing the dimensions of single-cell RNA-Seq data. Nucleic Acids Research. 2017;**45**(17):e156. DOI: 10.1093/nar/gkx681

[57] Curtis C, Shah SP, Chin SF, Turashvili G, Rueda OM, Dunning MJ, Speed D, Lynch AG, Samarajiwa S, Yuan Y, Gräf S. The genomic and transcriptomic architecture of 2,000 breast tumours reveals novel subgroups. Nature. 2012;**486**(7403):346-352. DOI: 10.1038/nature10983

[58] Islam MM, Ajwad R, Chi C, Domaratzki M, Wang Y, Hu P. Somatic copy number alteration-based prediction of molecular subtypes of breast cancer using deep learning model. 30th Canadian Conference on Artificial Intelligence. 2017 May 16:57-63

[59] Islam MM, Tian Y, Cheng Y, Wang Y, Hu P. A deep neural network regression model for triglyceride concentrations prediction using epigenome-wide methylation profiles. BMC Proceedings. 2018 (In press)

IntechOpen

IntechOpen

IntechOpen