Subject Section

Feature Selection May Improve Deep Neural Networks For The Bioinformatics Problems

Zheng Chen[#], Meng Pang[#], Zixin Zhao, Shuainan Li, Rui Miao, Yifan Zhang, Xiaoyue Feng, Xin Feng, Yexian Zhang, Meiyu Duan, Lan Huang and Fengfeng Zhou[*].

BioKnow Health Informatics Lab, College of Computer Science and Technology, and Key Laboratory of Symbolic Computation and Knowledge Engineering of Ministry of Education, Jilin University, Changchun, Jilin, China, 130012.

*To whom correspondence should be addressed.

#These authors contributed equally to this work.

Associate Editor: XXXXXXX

Received on XXXXX; revised on XXXXX; accepted on XXXXX

Abstract

Motivation: Deep neural network algorithms were utilized in predicting various biomedical phenotypes recently, and demonstrated very good prediction performances without selecting features. This study proposed a hypothesis that the deep neural network models may be further improved by feature selection algorithms.

Results: A comprehensive comparative study was carried out by evaluating 11 feature selection algorithms on three conventional deep neural network (DNN) algorithms, i.e., convolution neural network (CNN), deep belief network (DBN) and recurrent neural network (RNN), and three recent DNNs, i.e., MobilenetV2, ShufflenetV2 and Squeezenet. Five binary classification methylomic datasets were chosen to calculate the prediction performances of CNN/DBN/RNN models using feature selected by the 11 feature selection algorithms. Seventeen binary classification transcriptome and two multi-class transcriptome datasets were also utilized to evaluate how the hypothesis may generalize to different data types. The experimental data supported our hypothesis that feature selection algorithms may improve deep neural network models, and the DBN models using features selected by SVM-RFE usually achieved the best prediction accuracies on the five methylomic datasets.

Contact: FengfengZhou@gmail.com

Supplementary information: Supplementary data are available at *Bioinformatics* online.

1 Introduction

Modern biomedical technologies innovated and evolved comprehensively in recent years, and the biological OMIC data are accumulated at an accelerated speed (Stephens, et al., 2015). These data have been extensively utilized in investigating the genetic associations and biological mechanisms of human complex diseases like cancer and cardiovascular diseases (Bosse and Amos, 2018; Rosenson, et al., 2018; Seow, et al., 2017). Various gene panels have already been designed for scientific research (Chen, et al., 2019; Wang, et al., 2019) and even deployed in the commercial market (Mittendorf and King, 2019).

Feature selection played an essential role in establishing the predictive bioinformatics models based on the OMIC data (Alazmi, et al., 2018; Chen, et al., 2019; Mittendorf and King, 2019; Wang, et al., 2019; Xu, et al., 2018). There is a "large p small n" paradigm in the OMIC data based predictions, where p and n are the numbers of features and samples (Ye, et al., 2017; Zoh, et al., 2018). For example, a biomedical study usually does not have the luxury of collecting the similar number of samples to the number of human genes (Billatos, et al., 2018; Lim, et al., 2018). The literature demonstrated that not every OMIC feature was associated with the phenotype under investigation and some features may even decrease the model's predictive performances (Billatos, et al., 2018). So feature selection may be utilized to find a panel of features or biomarkers with

satisfying predictive performances (Chen, et al., 2017; Li, et al., 2014; Lin, et al., 2017; Qi, et al., 2019; Zeng, et al., 2016).

Deep neural network (DNN) facilitated applications in biomedical OMIC data only recently due to its intensive computation requirements (Dean, et al., 2012; Min, et al., 2017). DNN provides automatic feature abstraction and many DNN-based studies skipped the step of feature selection (Li, et al., 2018; Luo, et al., 2019). Various DNN frameworks were proposed in the last few years. Multi-layer convolution neural network (CNN) was used to predict protein's phosphorylation residues (Luo, et al., 2019). Deep belief network (DBN) is a multi-layer DNN with only inter-layer connections and demonstrated very good prediction accuracies using genomic and biophysiological data (Bu, et al., 2017; Lu, et al., 2018). Another type of DNN, the recurrent neural network (RNN), demonstrated its capability of accrately predicting genomic enhancers with cell-type specificities (Lim, et al., 2019).

Deep neural network (DNN) was only recently utilized to generate many bioinformatics prediction models (Di Lena, et al., 2012), but many of these studies did not fully utilize the power of feature selection to further improve the prediction models. Taking the top-tier bioinformatics journal Bioinformatics as an example, there were 195 papers matching the keywords "deep neural network", and only 36 of them matched the keywords "feature selection", too. As to the three popular DNN algorithms CNN/DBN/RNN, the journal Bioinformatics published 194, 31 and 132 papers, respectively. Moreover, there were only 28, 8 and 25 papers also matching "feature selection" for the three algorithms CNN/DBN/RNN, respectively. The above summary was updated on August 25, 2019.

This study hypothesized that feature selection may serve as a complementary step to improving a DNN model. We demonstrated that CNN, DBN and RNN may be improved by selecting a subset of features for the binary classification problems. Extensive evaluation was carried out about how eleven feature selection algorithms collaborated with the above three DNNs. Our experimental data suggested that feature selection improved the DNN's prediction accuracies in most cases.

2 Methods

2.1 Binary classification problem and its performance measurements

This study evaluated the predictive models for their binary classification performances. There are two groups of samples for a binary classification problem, i.e., positive (P) and negative (N) samples (Feng, et al., 2019; Senders, et al., 2019). Many machine learning algorithms were inherently designed for this simple model type, like the support vector machine (SVM) (Noble, 2006), etc. Many biomedical questions were based on this simple model type, e.g., the disease-vs-control samples in Genome-Wide Association Study (GWAS) and Epigenome-Width Association Study (EWAS) (Kupers, et al., 2019; Xie, et al., 2019).

This study evaluated a binary classification model by the metrics accuracy (Acc), F1-score, precision and recall, as similarly in (Feng, et al., 2019; Issarti, et al., 2019; Zhang, et al., 2018). The sizes of the positive and negative sample groups were P and N, respectivly. The number of correctly predicted positive samples was TP (true positive). The metrics TN (true negative) was the number of correctly predicted negative samples. The overall accuracy was the rate of the correctly predicted samples Acc=(TP+TN)/(P+N). Three more popular prediction performance metrics were also used to evaluate the binary prediction models, i.e., Precision=TP/(TP+FP), Recall=TP/(TP+FN), and F1-score=2×Precision×Recall/Precision+Recall) (Cogan, et al., 2019; Liu, et al., 2019).

All the algorithms were implemented and tested under the programming environment Python version 3.6.6.

2.2 Demonstrative datasets

This study chose five publicly available methylomic datasets to demonstrate how feature selection may improve a deep neural network model, as shown in Table 1. All the five methylome datasets were generated by the platform Illumina HumanMethylation450 BeadChip (GEO platform ID: GPL13534). Each methylome has 485,577 features, and all the five datasets have two groups of samples.

Table 1. Summary of the five binary classification datasets.

Dataset	Samples	Summary
GSE103186	191	intestinal metaplasia (130) vs normal(61)
GSE53045	111	smoker (77) vs control (34)
GSE66695	120	tumor (80) vs normal (40)
GSE74845	216	proximal (106) vs fimbrial (110)
GSE80970	286	Alzheimer's Disease (148) vs control (138)

These datasets were all methylomes generated by the Illumina HumanMethylation450 BeadChip.

2.3 Feature selection algorithms

Eleven feature selection algorithms were evaluated for their improvements over the deep neural network models. Firstly, we evaluated five filter algorithms. T-test (Trank) was a popular feature ranking algorithm and has been widely used to evaluate how significantly each feature was associated with the phenotype (Ye, et al., 2017). The association between a feature and the phenotype may also be evaluated from different criteria, e.g., linear SVM (lSVM) (Ozcift, 2012), chi-squared test (Chi2) (Pirooznia, et al., 2008), Mutual Information (MI) (Fernandez Rojas, et al., 2019), and SVM-based Recursive Feature Elimination (SVM-RFE) (Turewicz, et al., 2016). These feature selection algorithms were evaluated by the Incremental Feature Selection (IFS) strategy for the best feature subset (Ye, et al., 2017). The IFS strategy iteratively selected the next one feature that generated the current feature subset with the best prediction accuracy of the chosen DNN classifier. Due to the inherent randomization of the DNN training process, each random run may generate a different best feature subset.

Six more feature selection algorithms were evaluated on optimizing the deep neural network models. The criterion of Decision Tree (DTree) was utilized to optimize the chosen feature subset (Grabczewski and Jankowski, 2005). The algorithm Fast Correlation-Based Filter (FCBF) algorithm proposed three fast heuristic rules to identify redundant features without the pair-wise feature correlation calculations (Yu and Liu, 2003). Lasso (Lasso) utilized the L1 regularization to assign 0 as the weights of the features with minor phenotype associations (Guo, et al., 2019). McOne was another heuristic rule based feature removal algorithm (Ge, et al., 2016). The Random Forest (RF) was embedded as the evaluation classifier in the function SelectFromModel() to select a sub-optimal feature subset (Chen and Lin, 2006). A sampling-based meta random tree algorithm (MetaTree) was also evaluated for its performance on improving the deep neural network (He, et al., 2017).

2.4 Deep neural networks

The five datasets were pre-processed before applying the deep neural network algorithms. A feature was excluded if it has missing data.

A methylomic profile generated 485,577 features for one sample, and a Recurrent Neural Network (RNN) ran for over three days for only one epoch and 43GB of physical memory. An RNN was usually trained for a few hundred epochs. So this study used Variance to reduce the numbers of features. In the clinical setting, a measurement needs a very sensitive technology to detect if its variation range is small. So this study excluded features with small variances, and 500 features with the largest variances were kept for further analysis.

This study investigated three types of deep neural networks (DNNs), i.e., Convolution Neural Network (CNN) (Luo, et al., 2019), Deep Belief Network (DBN) (Bu, et al., 2017; Lu, et al., 2018), and Recurrent Neural Network (RNN) (Lim, et al., 2019). Unless otherwise specified, each of these three DNNs has three layers besides the input and output layers, and its learning rate was set to 0.001. The number of training epochs was 100. Additionally, CNN has a fully-connected first layer.

2.5 Stratified k-fold cross-validation

The stratified k-fold cross-validation (KFCV) strategy was utilized to calculate the overall classification performance. The groups of positive and negative samples in a dataset were randomly split into k equally-sized bins, respectively. For the iteration of k steps, one bin of positive and one bin of negative samples was chosen as the test dataset, and the other samples were used to train the classification model. This validation strategy ensured that each sample was used as the test sample for once and only once. The numbers of true positives and true negatives were summed over the iteration. The classification performance measurements Acc, F1-score, precision and recall were then calculated based on the definitions mentioned above.

2.6 Experimental procedure

The eleven feature selection algorithms were applied to generate feature subsets, and the three DNNs were evaluated for their classification performances over these feature subsets against the initial lists of 500 features. All the classification performances were calculated by the stratified 3-fold cross-validation strategy.

All the computational experiments were carried out in an Inspur Gene Server G100, with 256GB memory, 28 Intel Xeon® CPU cores (2.4GHz) and 30TB RISC1 disk space.

The overall procedure of this study was illustrated in Fig. 1.

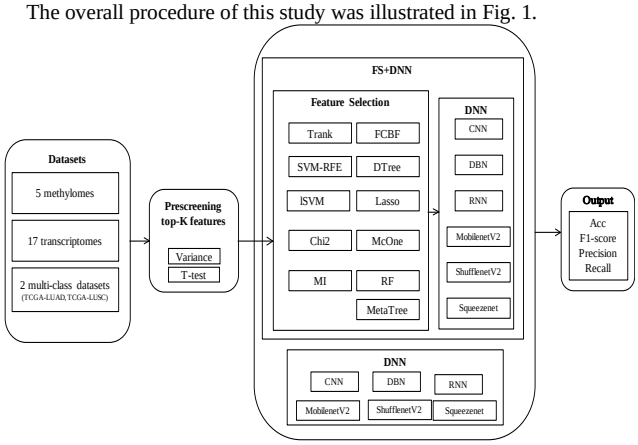


Fig. 1. The experimental flowchart of this study. There are several datasets of different types used to evaluate the performances of the feature selection algorithms, as in box of "Datasets". Eleven feature selection algorithms are shown in the box "Feature selection". "FS+DNN" is the DNN model using the features selected by a feature selection algorithm, while DNN could be a conventional neural network (CNN, DBN or RNN) or a recent deep neural network (MobilenetV2, ShufflenetV2 or Squeezenet). The classification performance is measured by the Acc (Accuracy), F1-score, Precision and Recall.

3 Results

3.1 Incremental selection of features ranked by t-test(Trank)

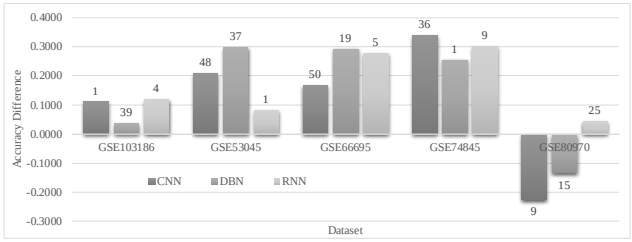


Fig. 2. Accuracy difference between FS+DNN and DNN. The measurement "Accuracy Difference" is the accuracy of FS+DNN minus that of DNN. DNN may be one of the three deep neural network algorithms CNN/DBN/RNN. "FS+DNN" is the model using the features selected by a feature selection algorithm, while "DNN" is the model using the initial list of 500 features with the maximal variances. The number of features selected by the t-test based incremental feature selection algorithm was labeled over the column, and the classification performance of each feature subset was evaluated by one of the three DNNs, i.e., CNN/DBN/RNN. The performance improvement of DNN by FS was also illustrated by F1-score, Precision and Recall in the Supplementary Figure S1.

Trank was widely used to evaluate how significantly a feature was associated with the phenotype (Ye, et al., 2017). The Incremental Feature Selection (IFS) strategy was usually used to find the top-k ranked features with the best classification performance, which was calculated by a user-defined classifier.

Fig. 2 demonstrated that the simple Trank based IFS strategy may improve the DNN-based model for most of the cases. Trank achieved the largest improvement for the dataset GSE74845. All the three DNNs did not perform very well on separating the fimbrial ovarian cancer patients (high risk) from the proximal counterparts (low risk) using the initial list of 500 features. Trank reduced the numbers of features to 36, 1 and 9 for the three DNNs, i.e., CNN, DBN and RNN. The largest overall accuracy 0.9417 was achieved by RNN with only nine features.

The second dataset GSE66695 has 80 breast cancer patients and 40 controls, and DBN achieved the best classification accuracies both before and after the step of feature selection. After the feature selection step, DBN achieved 0.2917 in improving the overall accuracy and the final prediction accuracy was 0.9833 using only 19 features. Reasonable improvements were also achieved for the other two datasets GSE53045 and GSE103186. Only a minor improvement was achieved for the dataset GSE80970. Both CNN and DBN decreased the model prediction accuracy by at least 10%. Only RNN improved the prediction model by 0.0420 in the overall accuracy.

Similar improvements were observed for the other three performance measurements, i.e., F1-score, Precision and Recall, as in the Supplementary Figure S1. Large improvements in all the three performance

measurements were observed for the three datasets GSE53045, GSE66695 and GSE74845. The dataset GSE103185 received minor improvements in the F1-score. The CNN's Precision and DBN's Recall were even slightly decreased after the step of feature selection. And feature selection didn't achieve much improvements for the dataset GSE80970.

3.2 Parameter tuning of CNN

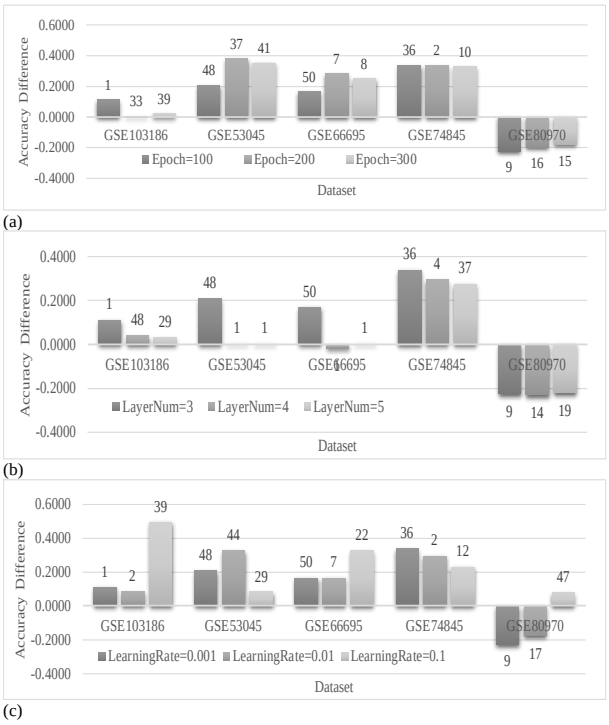


Fig. 3. How Trank improved CNN models with different parameter values. Accuracy differences between FS+CNN and CNN on the five datasets. FS+CNN is the CNN model trained over the features selected by Trank, and CNN is the CNN model trained over the initial list of 500 features. Numbers of features selected by Trank using CNN with different (a) Epoch values, (b) Layer numbers, and (c) Learning rates over the five datasets were labeled over the columns. The performance improvement of DNN by FS was also illustrated by F1-score, Precision and Recall in the Supplementary Figure S2.

The simple feature selection algorithm Trank improved most of the CNN models with different Epoch values, as shown in Fig. 3 (a). Trank achieved the largest improvements in accuracy (0.3349) of detecting ovarian cancers with high risk (dataset GSE74845). The data suggested that CNN did not handle well the information abstractions from the initial list of 500 features of the dataset GSE74845. CNN achieved a reasonable Acc=0.7417 with Epoch=300 on detecting breast cancer samples (GSE66695), but Trank chose only eight features to achieve Acc=0.9917. Among the five datasets, the best accuracy on the initial 500 features was achieved by CNN on the dataset GSE103186 (Acc=0.8534). Trank found 39 features to outperform the model by an improvement 0.0209.

The above figure suggested that the methylomic data in the ovarian cancer samples (GSE74845) may have noise patterns and even a simple Trank algorithm improved the CNN model by at least 0.2731 in Acc for different numbers of layers, as shown in Fig. 3 (b). The averaged number of features 25.6667 was much smaller than the initial number (500) of features, so that the training of the CNN model was much faster than using the 500 features. The best prediction accuracies over all the five datasets were achieved by three layers (LayerNum=3). And Trank did not improve the CNN model on predicting Alzheimers' diseases (GSE80970).

Learning rate played an essential role in optimizing a CNN model, but a larger learning rate did not always perform better, as shown in Fig. 3 (c). The feature selection algorithm Trank achieved at least 0.2000 in the averaged accuracy improvement for four of the five datasets and did not perform well on detecting the Alzheimer's disease (dataset GSE80970). The best averaged accuracy improvement 0.2886 was achieved on detecting the high-risk ovarian cancer (dataset GSE74845). Trank selected at most 50 features, which significantly reduced the time cost of training and testing.

So the simple feature selection algorithm Trank improved the CNN models for most cases. And the Alzheimer's disease seems to be challenging for Trank. The similar pattern may be observed for the other three performance measurements, i.e., F1-score, Precision and Recall, as in the Supplementary Figure S2.

3.3 Parameter tuning of DBN

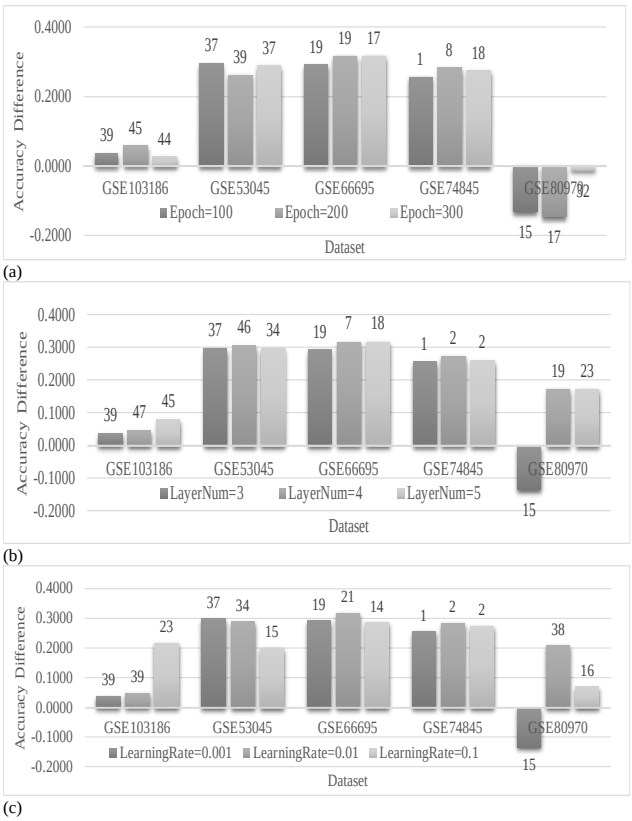


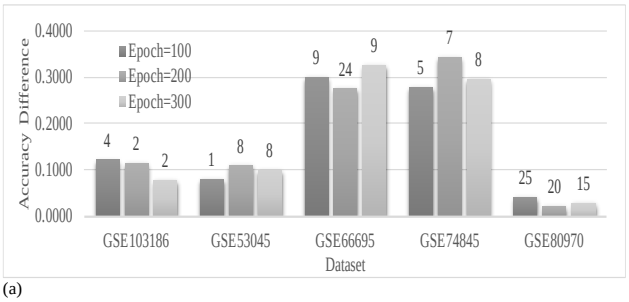
Fig. 4. How Trank improved DBN models with different parameter values. Accuracy differences between FS+DBN and DBN on the five datasets. FS+DBN is the DBN model trained over the features selected by Trank, and DBN is the DBN model trained over the initial list of 500 features. Numbers of features selected by Trank using DBN with different (a) Epoch values, (b) Layer numbers, and (c) Learning rates over the five datasets were labeled over the columns. The performance improvement of DNN by FS was also illustrated by F1-score, Precision and Recall in the Supplementary Figure S3.

The deep belief network (DBN) was also significantly improved by the feature selection algorithm Trank, as shown in Fig. 4 (a). The detection models of intestinal metaplasia (dataset GSE103186) and breast cancer (database GSE66695) were improved to achieve Acc>0.9100. Particularly, DBN achieved Acc=0.9917 with only 19 features for detecting breast cancers, while the initial list of 500 features only achieved Acc=0.6750 by the DBN model. The Alzheimer's disease (dataset GSE80970) seems to have many confounding factors, and the simple feature selection algorithm Trank did not improve the DBN model with different optimization Epochs. The parameter LayerNum=4 produced the best prediction accuracies for four datasets except for GSE80970, as shown in Fig. 4 (b). Breast cancer has a distinctive methylomic pattern against the normal samples, and seven methylation features generated a very good DBN model with Acc=0.9833. The initial list of 500 features did not perform very well on detecting breast cancer samples, suggesting the necessity of selecting the best subset of features for a DBN model. Smokers also demonstrated a specific methylomic pattern (dataset GSE53045), and 46 methylation features achieved the best prediction accuracy of 0.8559. The averaged improvement in accuracy 0.3003 suggested that some features may decrease the model's prediction performance. The least accuracy improvement 0.0471 was achieved by 47 features and LayerNum=4 on the dataset GSE103186.

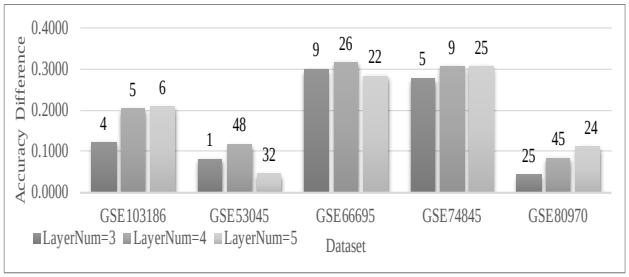
The default learning rate 0.001 achieved the best accuracies of predicting high-risk intestinal metaplasia (dataset GSE103186), smokers (dataset GSE53045) and breast cancers (dataset GSE66695), as shown in Fig. 4 (c). The DBN model achieved Acc=0.9833 with only 19 methylation features for predicting breast cancer patients, while the DBN model with the initial 500 features achieved Acc=0.6917. The best DBN model with the initial 500 features among all the five methylomic datasets achieved Acc=0.8796. And the 39 methylation features selected by Trank improved the DBN model to Acc=0.9162. Unfortunately, the Alzheimer's disease (dataset GSE80970) remained a difficult binary classification problem for DBN.

The other three performance measurements F1-score/Precision/Recall of DBN were also reasonably improved by the feature selection algorithm Trank, as shown in the Supplementary Figure S3. The same challenge remained in the dataset GSE80970.

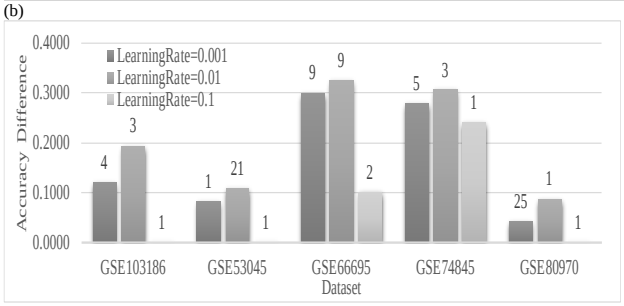
3.4 Parameter tuning of RNN



(a)



(b)



(c)

Fig. 5. How Trank improved RNN models with different parameter values. Accuracy differences between FS+RNN and RNN on the five datasets. FS+RNN is the RNN model trained over the features selected by Trank, and RNN is the RNN model trained over the initial list of 500 features. Numbers of features selected by Trank using RNN with different (a) Epoch values, (b) Layer numbers, and (c) Learning rates over the five datasets were labeled over the columns. The performance improvement of DNN by FS was also illustrated by F1-score, Precision and Recall in the Supplementary Figure S4.

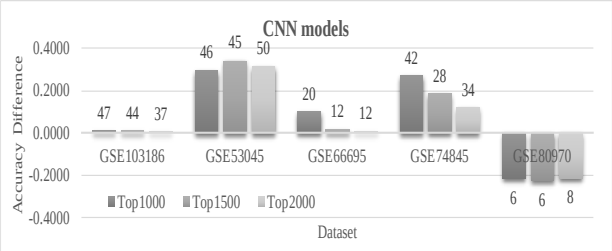
The simple feature selection algorithm Trank improved the RNN models with different Epoch numbers on all the five datasets, as shown in Fig. 5 (a). The best accuracy 0.9583 was achieved for detecting breast cancer patients (dataset GSE66695) by 200 Epochs, with an improvement 0.2750 against the initial 500 features. And the best accuracy using the initial 500 features (Acc=0.8115) was achieved with 300 Epochs for detecting the high-risk intestinal metaplasia (dataset GSE103186). The smallest averaged accuracy improvement 0.0303 was achieved by RNN for predicting the Alzheimer's disease (dataset GSE80970).

The parameter LayerNum=4 achieved the largest averaged accuracy improvement (0.2055) for the five methylomic datasets, as shown in Fig. 5 (b). The RNN models of all the five datasets were improved by the feature selection algorithms Trank, and the largest accuracy improvement 0.3167 was achieved by LayerNum=4 for detecting breast cancer patients. The 26 features selected by Trank achieved Acc=0.9667 by the RNN model. It is interesting to observe that RNN achieved the best Acc=0.7749 by 500 features and LayerNum=3. However, the feature selection algorithm Trank selected 4/5/6 features to achieve Acc=0.8953/0.8848/0.8901 for the parameter LayerNum=3/4/5, respectively.

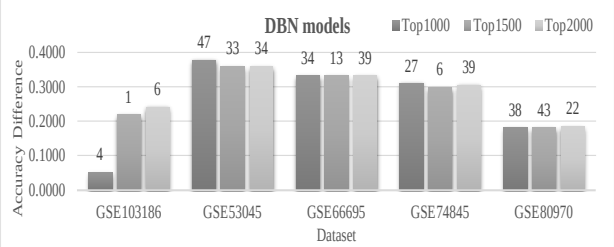
The feature selection algorithm Trank improved the RNN models with LearningRate=0.001/0.01/0.1 on all the five methylomic datasets, as shown in Fig. 5 (c). The largest accuracy improvement was achieved with LearningRate=0.01 for detecting breast cancer patients (dataset GSE66695), and the Trank selected nine features achieved Acc=0.9667, the best accuracy for the three LearningRate values on all the five datasets. And Fig. 5 (c) also demonstrated that at most 25 features were recommended by Trank to improve the RNN models. The RNN models received

similar improvements for the three measurements F1-score/Precision/Recall for different value choices of the two parameters Epoch and LayerNum, but the measurement Recall was not significantly improved for the three values of the parameter LearningRate, as shown in the Supplementary Figure S4.

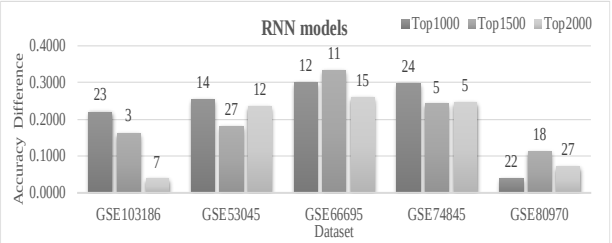
3.5 Different numbers of top-ranked features of DNN models



(a)



(b)



(c)

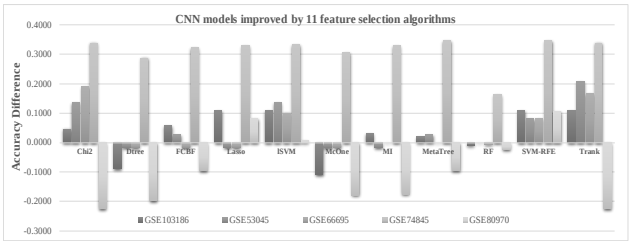
Fig. 6. How Trank improved the three DNN models with different numbers of top-ranked features. Accuracy differences between FS+DNN and DNN on the five datasets. FS+DNN is the DNN model trained over the features selected by Trank, and DNN is the DNN model trained over the initial list of the top-ranked features. Data labels on the top of each column were the numbers of features selected by Trank using DNN with different numbers of top-ranked features over the five datasets. DNN could be (a) CNN, (b) DBN and (c) RNN. The performance improvement of DNN by FS was also illustrated by F1-score, Precision and Recall in the Supplementary Figure S5.

We further evaluated how different numbers of top-ranked features may impact our hypothesis that feature selection may improve a deep neural network (DNN) model, as shown in Fig. 6. The feature selection algorithm Trank had varied performances among the five methylomic datasets, as shown in Fig. 6 (a). An averaged accuracy improvement 0.3183 was achieved by Trank for the three choices of top-ranked features, i.e., 1000, 1500 and 2000, for detecting the smoker's methylomic patterns (dataset GSE53045). But Trank decreased the CNN performance for predicting ovarian cancer patients at risk (dataset GSE74845), suggesting a necessity of optimizing the parameters of a machine learning algorithm.

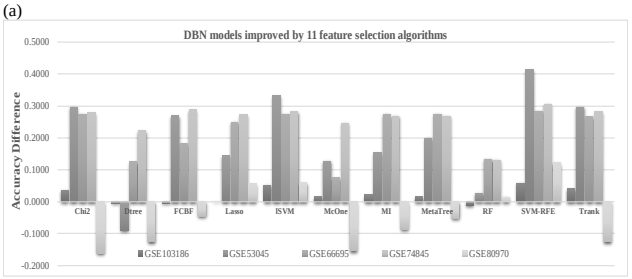
Trank improved the DBN models with top-ranked 1000/1500/2000 features on all the five methylomic datasets, as shown in Fig. 6 (b). The DBN models achieved at least Acc=0.9200 using the Trank-selected features for detecting the high-risk intestinal metaplasia (dataset GSE103186), smokers (dataset GSE53045) and breast cancer patients (dataset GSE66695). Particularly, the breast cancer patients could be detected at Acc=1.0000 using as small as 13 methylation features.

The RNN models on all the five methylomic datasets were also improved by the feature selection algorithm Trank, as shown in Fig. 6 (c). The dataset GSE66695 received the largest accuracy improvement from the feature selection algorithm Trank. Only 11 methylation features were needed to achieved Acc=1.0000 for detecting breast cancer patients. The high-risk intestinal metaplasia samples were detected at Acc=0.7435 using the top-ranked 1500 features, but Trank selected only three methylation feature to achieve a better accuracy 0.9058. Supplementary Figure S5 also supported our hypothesis on the prediction performance measurements F1-score/Precision/Recall.

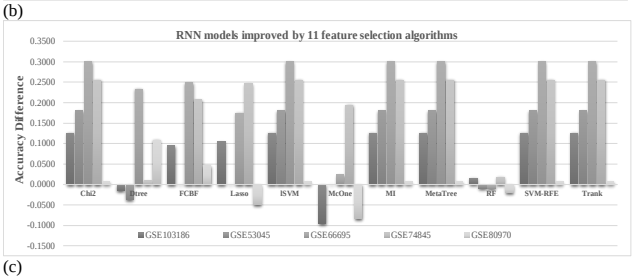
3.6 Evaluation of 11 feature selection algorithms on improving DNNs



(a)



(b)



(c)

Fig. 7. How 11 feature selection algorithms improved the three DNN models with on the five datasets. Accuracy differences were between FS+DNN and DNN on the five datasets. The DNN model was trained over the features selected by each feature selection algorithm denoted on the horizontal axis, and DNN is the DNN model trained over the initial list of 500 features. DNN could be (a) CNN, (b) DBN and (c) RNN. The other three

prediction performance measurements F1-score/Precision/Recall were also evaluated in the Supplementary Figure S6.

Fig. 7 demonstrated that the three DNN models may be improved by various feature selection algorithms. The CNN detection model of ovarian cancers at risk (dataset GSE74845) was improved by all the eleven feature selection algorithms, as shown in Fig. 7 (a). The largest accuracy improvement of 0.3472 was achieved by two feature selection algorithms MetaTree and SVM-RFE. However, MetaTree did not perform well on the other four datasets (averaged accuracy improvement -0.0116). SVM-RFE and lSVM were the only two feature selection algorithms that improved all the five methylomic datasets and achieved the top 2 largest averaged accuracy improvements. SVM-RFE performed the best with the averaged accuracy improvement 0.1453.

The two feature selection algorithms SVM-RFE and lSVM achieved the top 2 largest accuracy improvements for the DBN models on the five methylomic datasets, as shown in Fig. 7 (b). SVM-RFE selected features to achieve the prediction accuracy at least 0.9300 for four datasets, except for detecting the ovarian cancers at risk (dataset GSE74845). Another feature selection algorithm lSVM improved the DBN models of all the five datasets.

RNN collaborated well with many feature selection algorithms, and the largest averaged accuracy improvement 0.1735 were achieved by algorithms, as shown in Fig. 7 (c). Both SVM-RFE and lSVM were among these five feature selection algorithms. Similarly to the above sections, detecting the Alzheimer's disease remained the most challenging problem, with the best accuracy 0.6224 for the RNN models.

The observed best two feature selection algorithm SVM-RFE and lSVM were also supported by the other three prediction performance measurements F1-score/Precision/Recall in the Supplementary Figure S6.

3.7 Best feature selection algorithm for each DNN

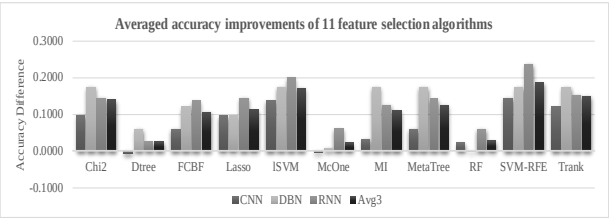


Fig. 8. Averaged accuracy improvements of the 11 feature selection algorithms on the three DNNs. Accuracy differences between FS+DNN and DNN on the five datasets. Supplementary Figure S7 presented the improvements of F1-scores, precisions and recalls of FS+DNN and DNN and also presented the accuracy differences, F1-score differences precisions, precision differences and recall differences between FS+DNN and DNN. The DNN model was trained over the features selected by each feature selection algorithm denoted on the horizontal axis, and DNN is the DNN model trained over the initial list of 500 features. The averaged accuracy improvement was averaged over the five methylomic datasets. The data series Avg3 gives the values averaged over the other three curves in this figure.

Fig. 8 and the above sections suggested that SVM-RFE was the best feature selection algorithm to work with the three DNNs on the five methylomic datasets. Another feature selection algorithm lSVM performed the second-best, with a slightly smaller Avg3 value in accuracy, F1-score, precision and recall than SVM-RFE, as shown in the Supplementary Figure S7. Trank was a popular and fast feature selection algorithm and achieved the third-best Avg3 value (accuracy=0.1485 and F1-score=0.1554) among the 11 algorithms. Trank performed very stably on these three DNNs and its standard deviation of the three accuracy improvement curves CNN/DBN/RNN (0.0272) was the smallest among the 11 feature selection algorithms. So the experimental data in this study suggested that SVM-RFE was the best feature selection algorithm to improve the deep neural networks CNN, DBN and RNN.

3.8 Best DNN working with features selected by SVM-RFE

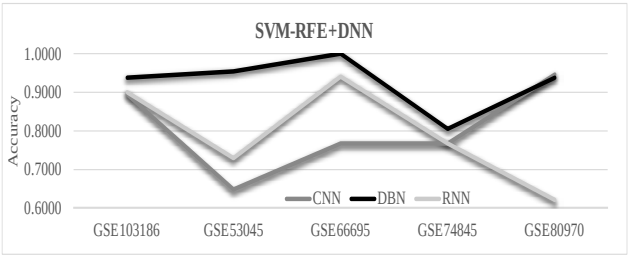
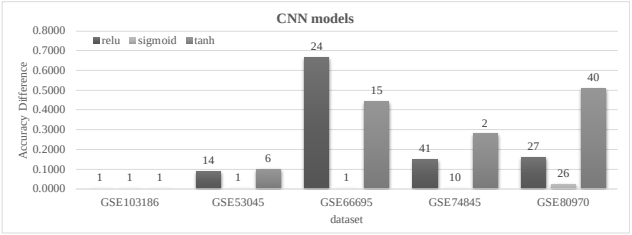


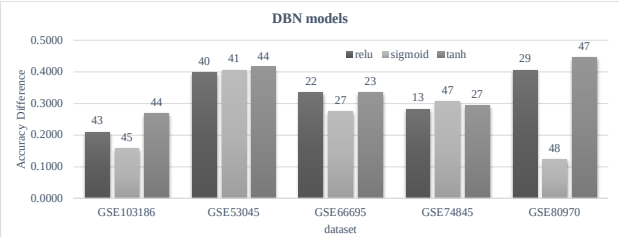
Fig. 9. Prediction accuracies of SVM-RFE working with the three DNNs on the five methylomic datasets. Supplementary Figure S8 demonstrated the other three prediction performance measurements, i.e., F1-score, precision and recall of SVM-RFE+DNN. DNN could be CNN, DBN or RNN. The horizontal axis was the dataset, and the vertical axis was the prediction performance of the deep neural network using the features selected by SVM-RFE.

DBN worked the best with the features selected by SVM-RFE, as shown in Fig. 9. The DBN models achieved the overall accuracy of 1.0000 and F1-score of 1.0000 for detecting breast cancer patients. The problem of detecting the ovarian cancers (GSE74845) at risk was very challenging since all the three deep neural networks did not work well on this problem. The best accuracy 0.8056, best F1-score 0.8151 and best precision 0.7578 were achieved by DBN. SVM-RFE selected features to achieve 0.9372 and 0.9550 in Acc by the DBN models on the dataset GSE103186 and the dataset GSE53045. The DBN model performed slightly worse than that of the CNN model for detecting Alzheimer's disease (GSE80970), and their difference in Acc was only 0.0105. Supplementary Figure S8 supported our observation that the feature selection algorithm SVM-RFE performed the best using the classifier DBN on most of the five datasets.

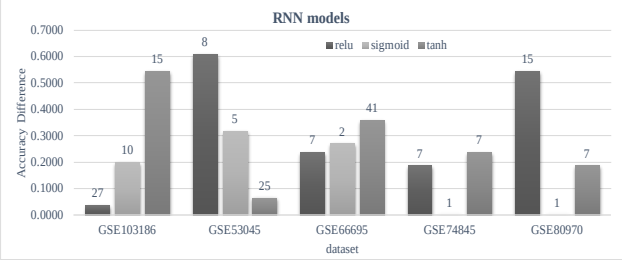
3.9 Evaluation of DNN models with different activation functions



(a)



(b)



(c)

Fig. 10. How SVM-RFE improved the three DNN models with different activation functions. Accuracy differences between FS+DNN and DNN on the five datasets are shown in the figure. Supplementary Figure S9 presented the plots for F1-scores, precisions and recalls of FS+DNN between DNN. FS+DNN is the DNN model trained over the features selected by the SVM-RFE based incremental feature selection algorithm, and DNN is the DNN model trained over the initial list of 500 features. Data labels on the top of each column were the numbers of features selected by SVM-RFE using DNN with different activation functions over the five datasets. DNN could be (a) CNN, (b) DBN and (c) RNN.

Different activation functions were evaluated for their impacts on our hypothesis that feature selection may improve a deep neural network (DNN) model, as shown in Fig. 10 and Supplementary Figure S9. In Fig. 10 (a), when activation function is tanh, CNN improved by SVM-RFE achieved largest accuracies and F1-score (Acc=0.6806/1.0000/0.7824/0.9895, F1-score=0.8099/1.0000/0.8127/0.9899) on the datasets GSE103186/GSE66695/GSE74845/GSE80970 using only 1/15/2/40 feature(s), respectively. CNN models with different activation functions only reached the largest accuracy 0.6396 after using SVM-RFE on the dataset GSE53045. It can be observed that it is challenging to improve the performance of predicting intestinal metaplasia patient (GSE103186) and the smoker people (GSE53045) for CNN models by feature selection.

In Fig. 10 (b), under the three activation functions (relu/sigmoid/tanh), SVM-RFE improved the accuracy of DBN models on all the five datasets. SVM-RFE improved DBN models by at least 0.2531 in averaged accuracy and at least 0.2462 in F1-score. In Fig. 10 (b), DBN training over SVM-RFE-selected features achieved accuracy at least 0.9424/0.9459/1.0000/0.9476 on the dataset GSE103186/GSE53045/GSE66695/GSE80970 when using relu, sigmoid or tanh. Besides, DBN training over SVM-RFE-selected features achieved accuracy 0.8009, which was 0.3056 more than DBN with the initial list of 500 features.

According to Fig. 10 (c), when the activation function is relu or tanh, SVM-RFE improved RNN on all the five datasets. When the activation function is sigmoid, SVM-RFE improved the accuracy and F1-score of the datasets GSE103186, GSE53045, and GSE66695. When the activation function is sigmoid, SVM-RFE did not improve the performance of GSE74845 and GSE80970, and the accuracies were only 0.5175 and 0.6806, respectively. But RNN with tanh training over SVM-RFE-selected features achieved the accuracy 0.7454 and F1-score 0.8122 on the dataset GSE74845 which made an accuracy improvement 0.2361 by only 7 features. The accuracy of RNN with relu trained over SVM-RFE-selected features reached 0.8639 and was 0.5445 more than that of RNN with relu using the initial list of 500 features. So DNN models with different activation functions may be improved by feature selections in most cases.

3.10 DNN models of multi-class prediction problems improved by SVM-RFE

This study further evaluated our hypothesis on two multi-class prediction problems LUSC and LUAD. The two transcriptome datasets were generated for the lung squamous cell carcinoma and lung adenocarcinoma samples using the microarray technology and were publicly available from the database of the cancer genome atlas (TCGA) (Cancer Genome Atlas Research, 2011; Feng, et al., 2019; Huo, et al., 2017). Each sample was annotated with a lung cancer developmental stage (I/II/III/IV). So these two datasets were both 4-class prediction problems. F1-scores, Precision and Recall were defined for the binary classification problems, so this section didn't evaluate these three measurements.

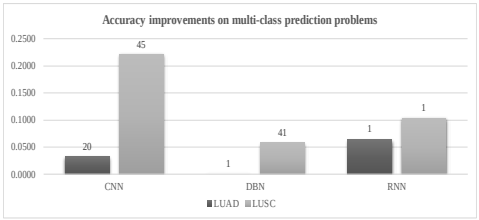


Fig. 11. How SVM-RFE improved the three DNN models on the multi-class prediction problems. The accuracy differences between FS+DNN and DNN on the two datasets were shown above. FS+DNN is the DNN model trained over the features selected by SVM-RFE and DNN is the DNN model trained over the initial list of 500 features. Data labels on the top of each column were the numbers of features selected by SVM-RFE. DNN could be CNN, DBN and RNN.

Fig. 11 illustrated that SVM-RFE improved the prediction accuracies of both CNN and RNN models on the two datasets LUAD and LUSC. The DBN model of the dataset LUSC was improved by 0.0584 in Acc by SVM-RFE, but achieved the same prediction accuracy 0.7419 on the dataset LUAD. The major contribution of SVM-RFE was that only one feature was selected to achieve the same prediction accuracy as the model using all the 500 top-ranked features. The RNN models on the two datasets LUAD and LUSC were significantly simplified to use only one feature to achieved 0.0645 and 0.1039 in the accuracy improvements, as shown in Fig. 11.

3.11 Evaluation of three recent DNN models

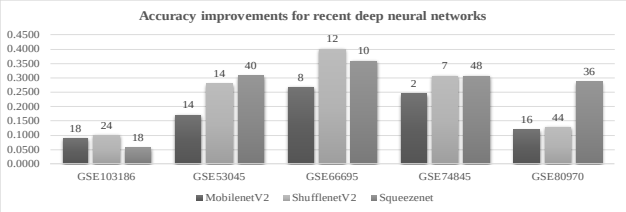


Fig. 12. Three recent DNN models were improved by SVM-RFE on the five datasets. Overall accuracy differences between FS+DNN and DNN on the two datasets are shown in the figure above. Supplementary Figure S10 presented accuracies, F1-scores, precisions and recalls of FS+DNN and DNN and also presented the accuracy differences, F1-score differences precisions, precision differences and recall differences between FS+DNN and DNN. FS+DNN is the DNN model trained over the features selected by the SVM-RFE based incremental feature selection algorithm, and DNN is the DNN model trained over the initial list of 500 features. Data labels on the top of each column were the numbers of features selected by SVM-RFE using DNN. DNN could be MobilenetV2, ShufflenetV2 and Squeezenet.

This study also evaluated whether more recent deep neural networks may be improved by feature selection and chose these three networks for this purpose, i.e., MobilenetV2 (Sandler, et al., 2018), ShufflenetV2 (Ma, et al., 2018) and Squeezenet (Santos, et al., 2018). SVM-RFE improved all the four performance measurements (Acc, F1-score, precision and recall) of all the three recent DNN models (MobilenetV2, ShufflenetV2 and Squeezenet) on all the five datasets, as shown in Fig. 12 and Supplementary Figure S10. SVM-RFE improved MobilenetV2, ShufflenetV2 and Squeezenet by at least 0.0576 in Acc and at least 0.0410 in F1-score with only 48 features at most. MobilenetV2, ShufflenetV2 and Squeezenet performed very well on the dataset GSE103186, GSE66695 and GSE80970 and all of them achieved accuracy at least 0.8671. Fig. 12 demonstrated that Squeezenet trained over the SVM-RFE-selected features achieved accuracies 0.9162, 1.0000 and 0.9755 on the datasets GSE103186, GSE66695 and GSE80970, respectively. SVM-RFE achieved at least 0.2522 and 0.2855 in the averaged accuracy improvement for these three models on the dataset GSE53045 and GSE74845, respectively, although the three models training over SVM-RFE-selected features achieved accuracy no more than 0.8468. The experimental data illustrated in Fig. 12 and Supplementary Figure S10 suggested that feature selection could improve both conventional DNNs (CNN, DBN and RNN) and more recent DNNs (MobilenetV2, ShufflenetV2 and Squeezenet).

3.12 How feature selection may improve the transcriptomic DNN models

Transcriptome is a popular OMIC technology and was also evaluated for our hypothesis. This study chose the 17 publicly-available transcriptome datasets to test whether a DNN model may be improved by the feature selection algorithm SVM-RFE (Ge, et al., 2016). All the 17 transcriptome datasets were binary classification problems. Fifteen of them described cancer-related problems, and were denoted as ALL1, ALL2, ALL3, ALL4, Adeno, CNS, Colon, DLBCL, Gas1, Gas2, Gas3, Leuk, Lym, Mye, and Pros, respectively. The other two datasets were about type I diabetes (T1D) and ischemic stroke (Stroke), respectively.

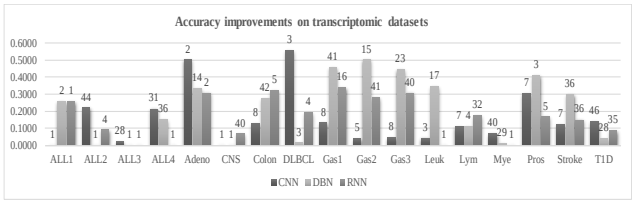


Fig. 13. Prediction accuracy improvements of CNN/DBN/RNN on the 17 transcriptomic datasets by feature selection. Overall accuracy differences between FS+DNN and DNN on the 17 datasets are shown in the figure above. Supplementary Figure S11 demonstrated the improvements of F1-score, precision and recall of FS+DNN and DNN. FS+DNN is the DNN model trained over the features selected by the SVM-RFE based incremental feature selection algorithm, and DNN is the DNN model trained over the initial list of 500 features. Data labels on the top of each column were the numbers of features selected by SVM-RFE using DNN. DNN could be CNN, DBN or RNN.

As shown in Fig. 13, SVM-RFE improved the prediction accuracies of all the three DNN models (CNN, DBN and RNN) on 10 datasets of 17 datasets (i.e. Adeno, Colon, DLBCL, Gas1, Gas2, Gas3, Lym, Pros, Stroke and T1D). At least one model (CNN, DBN or RNN) on the other 7 datasets (i.e. ALL1, ALL2, ALL3, ALL4, CNS, Leuk, Mye) was improved. Some DNN models were not improved due to that both DNN and FS+DNN achieved the prediction accuracies 1.0000, and theoretically there was no space for improvements. SVM-RFE achieved such prediction accuracies using only 15.8824 features on average.

CNN models were improved in Acc by at least 0.0240 on 15 datasets of the 17 datasets except for the datasets ALL1 and CNS. An averaged improvement in accuracy 0.2150 was achieved for the DBN models on the 17 transcriptome datasets using only 17.4118 features on averaged out the original list of 500 features. The prediction accuracies of the 17 RNN models were also improved by 0.1618 on average after selecting an averaged number of features 15.5882.

And it can be easily found that SVM-RFE improved the accuracy and F1-score or keep the same accuracy and the same F1-score on all models for the 17 datasets using only averaged 15.88 features.

Supplementary Figure S11 demonstrated a similar observation that the F1-score, precision and recall of the three DNN models on the 17 datasets were significantly improved or remained the same after the numbers of features were significantly reduced by SVM-RFE.

3.13 An alternative selection of initial 500 features by Trank

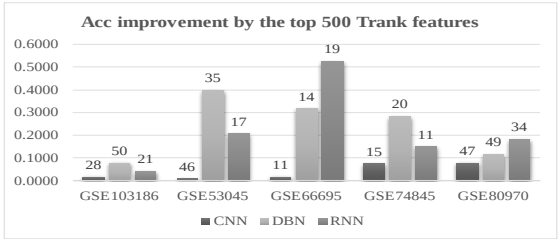
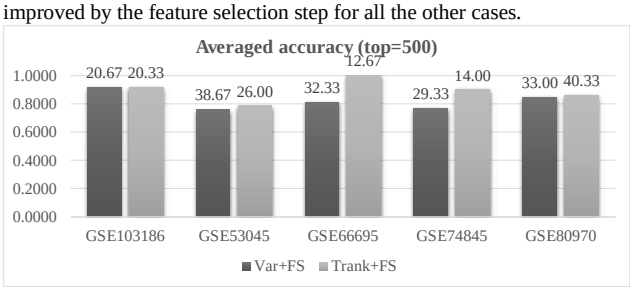


Fig. 14. The CNN/DBN/RNN models improved by the SVM-RFE based IFS strategy using the top 500 Trank features. Overall accuracy differences between FS+DNN and DNN on the same dataset are shown in the figure above. Supplementary Figure S12 demonstrated the improvements in F1-scores, precisions and recalls of FS over each DNN. The

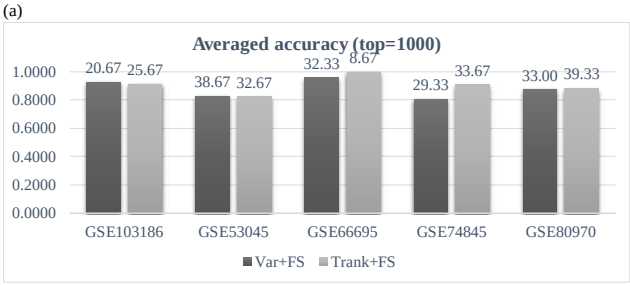
DNN model was trained over the top 500 Trank features. Data labels on the top of each column were the numbers of features selected by SVM-RFE using DNN. DNN could be CNN, DBN or RNN.

The methylomic features were screened for those with the largest variance and may introduce bias before any feature selection algorithm. So we evaluated our hypothesis over the top 500 features ranked by the popular t-test algorithm, as shown in Fig. 14 and Supplementary Figure S12. Our hypothesis was still supported by the observation that SVM-RFE improved all the three DNN models on the five datasets in the performance measurement Acc, as shown in Fig. 14. DNN models trained over the SVM-RFE-selected features achieved at least 0.0090 in Acc for the five methylomic datasets using at most 10% features.

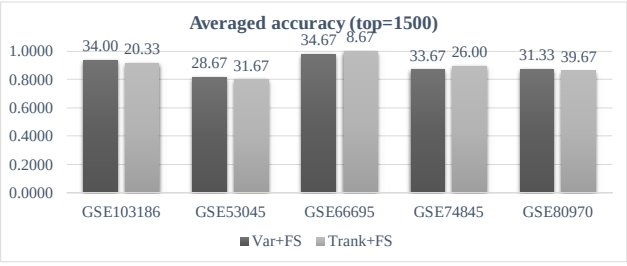
Supplementary Figure S12 also demonstrated that feature selection improved the three other classification performance measurements in most cases, i.e., F1 score, precision and recall. F1 scores were improved by at least 0.0121 for all the three DNN algorithms over all the five datasets, as in the Supplementary Figure S12 (a). Minor decreases in precision were observed for the RNN models over the two datasets GSE103186 (-0.0123) and GSE53045 (-0.0625), respectively (Supplementary Figure S12 (b)). DBN generated a minor decrease in recall (-0.0154) for the dataset GSE80970 (Supplementary Figure S12 (c)). The precision and recall were improved by the feature selection step for all the other cases.



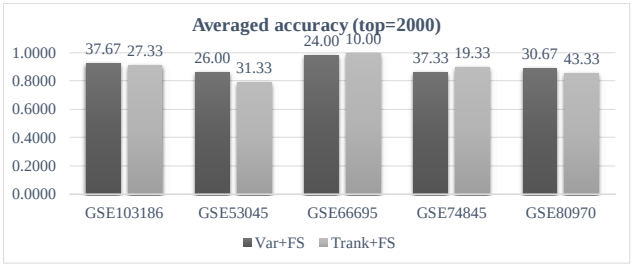
(a)



(b)



(c)



(d)

Fig. 15. Performance comparison between Variance and Trank over the five datasets. SVM-RFE was used to select features from the top features ranked by Variance (Var+FS) and Trank (Trank+FS). The prediction accuracy averaged over the three algorithms CNN/DBN/RNN was compared for different numbers of features, i.e., (a) 500, (b) 1000, (c) 1500 and (d) 2000. The Supplementary Figure S13 gave the comparison in the other three performance measurements, i.e., F1-score, Precision and Recall.

Fig. 15 (a) demonstrated that Trank+FS outperformed Var+FS in the averaged accuracies of CNN, DBN and RNN on all the five datasets using the top 500 features ranked by Trank and Var, respectively. Trank+FS only outperformed Var+FS on two or three out of the five datasets if more top-ranked features were kept for feature selection, as shown in Fig. 15 (b), (c) and (d). Supplementary Figure S13 demonstrated that the Var+FS models achieved comparable prediction performances as to Trank+FS in the three other performance measurements, including F1-score, precision, and recall.

So overall, feature selection improved the CNN/DBN/RNN models using the top Trank features, too. The performance comparison between Var+FS and Trank+FS in Fig. 15 and Supplementary Figure S13 suggested that similar prediction performances may be achieved for the features prescreened by Variance or Trank, and no evidence was observed that Trank+FS consistently outperformed Var+FS.

3.14 Biological inferences of the detected biomarkers

Table 2. SVM-RFE selected 33 methylation biomarkers for detecting breast cancer patients.

ID	HG	Chr	Location	RefGene	ID	HG	Chr	Location	RefGene
cg17520909	37	1	159683855	CRP	cg01811882	37	13	22493739	
cg12680131	37	2	1417153	TPO	cg02262167	37	13	112157919	
cg08281123	37	3	19791304		cg08049519	37	15	81412880	
cg00584840	37	3	112903054		cg21829038	37	15	101593831	LRRK1
cg17155524	37	4	2305734	ZFYVE28	cg05280794	37	17	1040653	ABR;ABR
cg03779973	37	4	151001971	DCLK2	cg01600516	37	17	6904263	ALOX12
cg23943944	37	5	116197864		cg08002427	37	17	11503508	DNAH9
cg08812108	37	6	2515318		cg15290312	37	17	76897139	TIMP2
cg16620382	37	6	43211213	TTBK1	cg23749482	37	17	77901317	
cg06231385	37	7	4743056	FOXK1	cg17714010	37	17	78968450	CHMP6
cg10454879	37	8	602158		cg04817870	37	17	79963651	ASPSCR1
cg07298985	37	8	22133076	PIWIL2	cg21410293	37	18	46386361	KIAA0427
cg27024127	37	8	27522576	SCARA3	cg13782274	37	20	62097681	KCNQ2
cg09533869	37	8	97747124	PGCP	cg11767442	37	21	45584432	
cg21211688	37	9	136403935	ADAMTSL2	cg00949980	37	21	46237117	SUMO3
cg12682323	37	10	132883127		cg00014152	37	X	48457128	WDR13
cg09540629	37	10	134981759	KNDC1					

ID is the probeset ID in the platform Illumina HumanMethylation450 (GPL13534). HG is the human genome build number. Chr and Location give the chromosome and position of each biomarker. And RefGene is the UCSC RefGene name where the methylation biomarker locates.

The feature selection algorithm selected 33 methylation biomarkers for detecting breast cancers, and many of them were known to be associated with the breast cancer onset and development, as shown in Table 2.

Chromosome 17 harbored the largest number of biomarkers and was known to receive frequent mutations in sporadic breast carcinomas (Carrozzo and Ledbetter, 1993; Coles, et al., 1990). Among the 7 methylation biomarkers, 6 resided within the bodies of protein-coding genes. Three genes (ABR, ALOX12 and TIMP2) were known to be associated with breast cancer (Chien, et al., 2018; Huang, et al., 2019; Liscia, et al., 1999). The biomarker ABR encodes a RhoGEF and GTPase Activating Protein and plays an essential role in suppressing tumor development (Liscia, et al., 1999). The inhibition of the arachidonate lipoxygenase12 (Alox12) sensitized breast cancer to chemotherapy (Huang, et al., 2019). Triple-negative breast cancer would be very invasive if the gene TIMP2 was repressed (Chien, et al., 2018).

Besides TIMP2, there were three other tumor suppressor genes among our 33 methylation biomarkers. The methylation biomarker cg06231385 resides within the gene Foxk1, which was transcribed at a significantly low level in breast cancer than the adjacent tissue (Sun, et al., 2016). Another methylation biomarker cg07298985 was within the transcription start site (TSS) of the gene P-Element-induced wimpy testis (PIWIL2), which is involved in the epigenetic regulation process in germline cells (Litwin, et al., 2018). PIWIL2 was observed to have a statistically significantly high expression level in the invasive ductal breast cancer (Litwin, et al., 2018). The expression status of PIWIL2 alone served well as a prognostic biomarker for various clinical factors, including tumor stage and size (Sarvestani, et al., 2016). The third methylation biomarker cg12680131 was within the TSS region of another gene Thyroid peroxidase (TPO), which was the key enzyme involved in thyroid hormone synthesis and a protective role in breast cancer development (Godlewska and Banga, 2019).

Table 3. SVM-RFE selected 48 methylation biomarkers for detecting Alzheimer's patients.

ID	HG	Chr	Location	RefGene
cg26422465	37	1	3561008	WDR8
cg17279365	37	1	217168635	ESRRG
cg12466610	37	1	220950205	MOSC2
cg00345083	37	1	4725584	AJAP1
cg03221390	37	1	247803637	
cg00017157	37	1	167127353	
cg11418303	37	1	71511514	PTGER3
cg24007926	37	2	206842760	
cg05023192	37	2	240965916	NDUFA10
cg01957222	37	2	1426786	TPO
cg04814784	37	3	10182561	VHL
cg03965172	37	4	3683268	
cg08506672	37	5	3959743	
cg10140678	37	5	2864736	
cg04922606	37	6	170703742	FAM120B
cg00035449	37	6	169539646	
cg15567368	37	7	563891	
cg11420142	37	8	92570895	
cg02299007	37	8	1140574	
cg14114910	37	9	124924045	MORN5
cg21211688	37	9	136403935	ADAMTSL2
cg14877834	37	10	34151366	
cg02533724	37	10	130279912	
cg19377607	37	10	72138124	LRRC20
cg24697097	37	10	131912837	
cg05059349	37	13	112849476	
cg22274196	37	13	95958190	
cg07456585	37	14	74704714	VSX2
cg01543583	37	14	59947673	C14orf149
cg21193926	37	14	76443578	TGFB3
cg08049519	37	15	81412880	
cg11418607	37	15	67323243	
cg02074316	37	15	94147555	
cg02856402	37	16	4560300	HMOX2
cg07128503	37	16	85747424	C16orf74
cg08024264	37	17	75084281	SCARNA16;C17orf86
cg26536949	37	17	57053	
cg13989295	37	17	57187728	SKA2
cg10771931	37	19	34972145	WTIP
cg26365090	37	20	42574362	TOX2
cg23712855	37	X	153991009	DKC1
cg14295915	37	X	48334557	FTSJ1
cg03905640	37	Y	19625308	FAM41AY2;FAM41AY1
rs133860				
rs1945975				
rs2235751				
rs10882854				
rs845016				

ID is the probeset ID in the platform Illumina HumanMethylation450 (GPL13534). HG is the human genome build number. Chr and Location give the chromosome and position of each biomarker. And RefGene is the UCSC RefGene name where the methylation biomarker locates.

As shown in Table 3, 48 methylation biomarkers for detecting Alzheimer's disease were selected by SVM-RFE, and some of them were known to be associated with the Alzheimer's disease onset and development. The methylation biomarker cg05023192 resided within the bodies of protein-coding gene NDUFA10 from chromosome 2. NDUFA10 is one of the critical hippocampal genes and pathways that are involved in the pathogenesis of Alzheimer's Disease (AD). Liang, et al. proposed that an energy deficiency in the brain might be the commonest etiological agent for AD. NDUFA10 and NDUFA9 encode subunits of NADH (Zhang, et al., 2015). As one protein complex, NADH dehydrogenase is considered the vital energy producer (Zhang, et al., 2015). And quite a few other members (NDUFB3, NDUFA9, NDUFV1, NDUFV2, NDUFS3) of the NDUF family were also observed to be involved in the development of Alzheimer's disease (Zhang, et al., 2015).

The methylation biomarker cg04814784 was within the gene VHL. Mutations in the tumor suppressor gene VHL cause von Hippel-Lindau disease (VHL). VHL encodes a protein, which is a part of a multiprotein complex involved in ubiquitination and degradation of the transcription factor HIF leading to the dysregulation of a variety of genes involved in growth control (Rosner, et al., 2008). Hypoxia-inducible transcription factor-1 (HIF-1) regulated the cellular responses that are activated by reduction of blood supply leading to hypoxic condition. And cerebrovascular deficiencies (such as cerebral ischemia/stroke) were generally thought to contribute to AD (Lonati, et al., 2014).

Another gene thyroid peroxidase (TPO) harbored the methylation biomarker cg01957222 and was also closely associated with AD. The experimental data of the TPO serum antibodies (TPO-Abs) suggested that the subclinical hyperthyroidism in the elderly increases the risk of dementia and Alzheimer's disease (Kalmijn, et al., 2000). Such abnormal TPO metabolism was further supported by the imaging modality of both MRI and PET (Yamaguchi, et al., 1997).

4 Conclusions

This study investigated the hypothesis that deep neural network (DNN) models may be improved by feature selection algorithms. Our experimental data supported the hypothesis that even the simple t-test based feature ranking algorithm may significantly improve the three deep neural networks CNN/DBN/RNN in most cases. The comparative data visualizations suggested that the DBN model using features selected by SVM-RFE achieved the best prediction performances for the five methylomic datasets.

The data in this study demonstrated that although DNNs performed well on feature abstraction, there is still a large performance improvement space by selecting a good subset of features.

Acknowledgements

The results shown here are in whole or part based upon data generated by the TCGA Research Network: https://www.cancer.gov/tcga. The insightful and constructive comments from the anonymous reviewers were greatly appreciated.

Funding

This work was supported by the Strategic Priority Research Program of the Chinese Academy of Sciences (XDB13040400), Jilin Provincial Key Laboratory of Big Data Intelligent Computing (20180622002JC), the Education Department of Jilin Province (JJKH20180145KJ), and the startup grant of the Jilin University. This work was also partially supported by the BioknowMedAI Institute (BMCPP-2018-001), the High Performance Computing Center of Jilin University, and by the Fundamental Research Funds for the Central Universities, JLU.

Conflict of Interest: The authors declared no conflicts of interest.

References

Alazmi, M., et al. Systematic selection of chemical fingerprint features improves the Gibbs energy prediction of biochemical reactions. Bioinformatics 2018.

Billatos, E., et al. The Airway Transcriptome as a Biomarker for Early Lung Cancer Detection. Clin Cancer Res 2018;24(13):2984-2992.

Bosse, Y. and Amos, C.I. A Decade of GWAS Results in Lung Cancer. Cancer Epidemiol Biomarkers Prev 2018;27(4):363-379.

Bu, H., et al. A new method for enhancer prediction based on deep belief network. BMC Bioinformatics 2017;18(Suppl 12):418.

Cancer Genome Atlas Research, N. Integrated genomic analyses of ovarian carcinoma. Nature 2011;474(7353):609-615.

Carrozzo, R. and Ledbetter, D.H. Dinucleotide repeat polymorphism mapping to the critical region for lissencephaly (17p13.3). Hum Mol Genet 1993;2(5):615.

Chen, X., et al. Long non-coding RNAs and complex diseases: from experimental results to computational models. Brief Bioinform 2017;18(4):558-576.

Chen, Y.-W. and Lin, C.-J. Combining SVMs with various feature selection strategies. In, Feature extraction. Springer; 2006. p. 315-324.

Chen, Y.L., et al. A 17 gene panel for non-small cell lung cancer prognosis identified through integrative epigenomic-transcriptomic analyses of hypoxia-induced epithelial-mesenchymal transition. Mol Oncol 2019.

Chien, Y.C., et al. EZH2 promotes migration and invasion of triple-negative breast cancer cells via regulating TIMP2-MMP-2/-9 pathway. Am J Cancer Res 2018;8(3):422-434.

Cogan, T., Cogan, M. and Tamil, L. MAPGI: Accurate identification of anatomical landmarks and diseased tissue in gastrointestinal tract using deep learning. Comput Biol Med 2019;111:103351.

Coles, C., et al. Evidence implicating at least two genes on chromosome 17p in breast carcinogenesis. Lancet 1990;336(8718):761-763.

Dean, J., et al. Large scale distributed deep networks. In, Advances in neural information processing systems. 2012. p. 1223-1231.

Di Lena, P., Nagata, K. and Baldi, P. Deep architectures for protein contact map prediction. Bioinformatics 2012;28(19):2449-2457.

Feng, X., et al. Age Is Important for the Early-Stage Detection of Breast Cancer on Both Transcriptomic and Methylomic Biomarkers. Front Genet 2019;10:212.

Fernandez Rojas, R., Huang, X. and Ou, K.L. A Machine Learning Approach for the Identification of a Biomarker of Human Pain using fNIRS. Sci Rep 2019;9(1):5645.

Ge, R., et al. McTwo: a two-step feature selection algorithm based on maximal information coefficient. BMC Bioinformatics 2016;17:142.

Godlewska, M. and Banga, P.J. Thyroid peroxidase as a dual active site enzyme: Focus on biosynthesis, hormonogenesis and thyroid disorders of autoimmunity and cancer. Biochimie 2019;160:34-45.

Grabczewski, K. and Jankowski, N. Feature selection with decision tree criterion. In, Fifth International Conference on Hybrid Intelligent Systems (HIS'05). IEEE; 2005. p. 6 pp.

Guo, Y., et al. Combining Sparse Group Lasso and Linear Mixed Model Improves Power to Detect Genetic Variants Underlying Quantitative Traits. Front Genet 2019;10:271.

He, D., et al. Software-defined-networking-enabled traffic anomaly detection and mitigation. 2017;4(6):1890-1898.

Huang, Z., et al. ALOX12 inhibition sensitizes breast cancer to chemotherapy via AMPK activation and inhibition of lipid synthesis. Biochem Biophys Res Commun 2019.

Huo, D., et al. Comparison of Breast Cancer Molecular Features and Survival by African and European Ancestry in The Cancer Genome Atlas. JAMA Oncol 2017;3(12):1654-1662.

Issarti, I., et al. Computer aided diagnosis for suspect keratoconus detection. Comput Biol Med 2019;109:33-42.

Kalmijn, S., et al. Subclinical hyperthyroidism and the risk of dementia. The Rotterdam study. Clin Endocrinol (Oxf) 2000;53(6):733-737.

Kupers, L.K., et al. Meta-analysis of epigenome-wide association studies in neonates reveals widespread differential DNA methylation associated with birthweight. Nat Commun 2019;10(1):1893.

Li, M., et al. Detecting tissue-specific early warning signals for complex diseases based on dynamical network biomarkers: study of type 2 diabetes by cross-tissue analysis. Brief Bioinform 2014;15(2):229-243.

Li, Y., et al. DEEPre: sequence-based enzyme EC number prediction by deep learning. Bioinformatics 2018;34(5):760-769.

Lim, A., Lim, S. and Kim, S. Enhancer prediction with histone modification marks using a hybrid neural network model. Methods 2019.

Lim, S.B., et al. A merged lung cancer transcriptome dataset for clinical predictive modeling. Sci Data 2018;5:180136.

Lin, Y., et al. Computer-aided biomarker discovery for precision medicine: data resources, models and applications. Brief Bioinform 2017.

Liscia, D.S., et al. Prognostic significance of loss of heterozygosity at loci on chromosome 17p13.3-ter in sporadic breast cancer is evidence for a putative tumour suppressor gene. Br J Cancer 1999;80(5-6):821-826.

Litwin, M., et al. Aberrant Expression of PIWIL1 and PIWIL2 and Their Clinical Significance in Ductal Breast Carcinoma. Anticancer Res 2018;38(4):2021-2030.

Liu, G., Mao, S. and Kim, J.H. A Mature-Tomato Detection Algorithm Using Machine Learning and Color Analysis. Sensors (Basel) 2019;19(9).

Lonati, E., et al. Pin1, a new player in the fate of HIF-1alpha degradation: an hypothetical mechanism inside vascular damage as Alzheimer's disease risk factor. Front Cell Neurosci 2014;8:1.

Lu, P., et al. Research on Improved Depth Belief Network-Based Prediction of Cardiovascular Diseases. J Healthc Eng 2018;2018:8954878.

Luo, F., et al. DeepPhos: prediction of protein phosphorylation sites with deep learning. Bioinformatics 2019.

Ma, N., et al. Shufflenet v2: Practical guidelines for efficient cnn architecture design. In, Proceedings of the European Conference on Computer Vision (ECCV). 2018. p. 116-131.

Min, S., Lee, B. and Yoon, S. Deep learning in bioinformatics. Brief Bioinform 2017;18(5):851-869.

Mittendorf, E.A. and King, T.A. Routine Use of Oncotype DX Recurrence Score Testing in Node-Positive Hormone Receptor-Positive HER2-Negative Breast Cancer: The Time Has Come. Ann Surg Oncol 2019;26(5):1173-1175.

Noble, W.S.J.N.b. What is a support vector machine? 2006;24(12):1565.

Ozcift, A. SVM feature selection based rotation forest ensemble classifiers to improve computer-aided diagnosis of Parkinson disease. J Med Syst 2012;36(4):2141-2147.

Pirooznia, M., et al. A comparative study of different machine learning methods on microarray gene expression data. BMC Genomics 2008;9 Suppl 1:S13.

Qi, X., et al. Decoding competing endogenous RNA networks for cancer biomarker discovery. Brief Bioinform 2019.

Rosenson, R.S., et al. HDL and atherosclerotic cardiovascular disease: genetic insights into complex biology. Nat Rev Cardiol 2018;15(1):9-19.

Rosner, M., et al. The mTOR pathway and its role in human genetic diseases. Mutat Res 2008;659(3):284-292.

Sandler, M., et al. Mobilenetv2: Inverted residuals and linear bottlenecks. In, Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2018. p. 4510-4520.

Santos, A.G., et al. Reducing squeezenet storage size with depthwise separable convolutions. In, 2018 International Joint Conference on Neural Networks (IJCNN). IEEE; 2018. p. 1-6.

Sarvestani, F.M., et al. The Evaluation of Cancer Testis Gene PIWIL2 Expression Levels as a New Prognostic Biomarker for Breast Cancer. Clin Lab 2016;62(8):1469-1475.

Senders, J.T., et al. Natural Language Processing for Automated Quantification of Brain Metastases Reported in Free-Text Radiology Reports. JCO Clin Cancer Inform 2019;3:1-9.

Seow, W.J., et al. Association between GWAS-identified lung adenocarcinoma susceptibility loci and EGFR mutations in never-smoking Asian women, and comparison with findings from Western populations. Hum Mol Genet 2017;26(2):454-465.

Stephens, Z.D., et al. Big Data: Astronomical or Genomical? PLoS Biol 2015;13(7):e1002195.

Sun, T., et al. Forkhead box protein k1 recruits TET1 to act as a tumor suppressor and is associated with MRI detection. Jpn J Clin Oncol 2016;46(3):209-221.

Turewicz, M., et al. PAA: an R/bioconductor package for biomarker discovery with protein microarrays. Bioinformatics 2016;32(10):1577-1579.

Wang, Z., et al. Assessment of Blood Tumor Mutational Burden as a Potential Biomarker for Immunotherapy in Patients With Non-Small Cell Lung Cancer With Use of a Next-Generation Sequencing Cancer Gene Panel. JAMA Oncol 2019.

Xie, T., et al. Epigenome-Wide Association Study (EWAS) of Blood Lipids in Healthy Population from STANISLAS Family Study (SFS). Int J Mol Sci 2019;20(5).

Xu, C., et al. An OMIC biomarker detection algorithm TriVote and its application in methylomic biomarker detection. Epigenomics 2018;10(4):335-347.

Yamaguchi, S., et al. Decreased cortical glucose metabolism correlates with hippocampal atrophy in Alzheimer's disease as shown by MRI and PET. J Neurol Neurosurg Psychiatry 1997;62(6):596-600.

Ye, Y., et al. RIFS: a randomly restarted incremental feature selection algorithm. Sci Rep 2017;7(1):13013.

Yu, L. and Liu, H. Feature selection for high-dimensional data: A fast correlation-based filter solution. In, Proceedings of the 20th international conference on machine learning (ICML-03). 2003. p. 856-863.
Zeng, T., et al. Big-data-based edge biomarkers: study on dynamical drug sensitivity and resistance in individuals. Brief Bioinform 2016;17(4):576-592.
Zhang, L., et al. Potential hippocampal genes and pathways involved in Alzheimer's disease: a bioinformatic analysis. Genet Mol Res 2015;14(2):7218-7232.
Zhang, R., et al. pyHIVE, a health-related image visualization and engineering system using Python. BMC Bioinformatics 2018;19(1):452.
Zoh, R.S., et al. A Powerful Bayesian Test for Equality of Means in High Dimensions. J Am Stat Assoc 2018;113(524):1733-1741.