Journal of proteome research

Article

Cite This: J. Proteome Res. 2019, 18, 3195−3202

pubs.acs.org/jpr

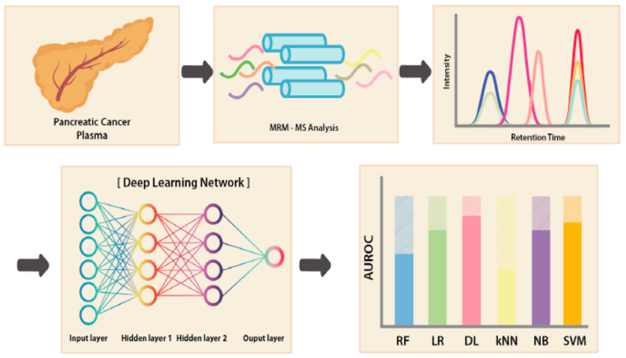
Clinically Applicable Deep Learning Algorithm Using Quantitative Proteomic Data

Hyunsoo Kim,[†,‡,§,⊥] Yoseop Kim,[§,⊥] Buhm Han,[‡] Jin-Young Jang,[*,∥] and Youngsoo Kim[*,†,‡,§]

[†]Institute of Medical and Biological Engineering, Medical Research Center,[‡]Department of Biomedical Sciences, [§]Department of Biomedical Engineering, and [∥]Department of Surgery, Seoul National University College of Medicine, Yongon-Dong, Seoul 110-799, Republic of Korea

Supporting Information

ABSTRACT: Deep learning (DL), a type of machine learning approach, is a powerful tool for analyzing large sets of data that are derived from biomedical sciences. However, it remains unknown whether DL is suitable for identifying contributing factors, such as biomarkers, in quantitative proteomics data. In this study, we describe an optimized DL-based analytical approach using a data set that was generated by selected reaction monitoring−mass spectrometry (SRM−MS), comprising SRM−MS data from 1008 samples for the diagnosis of pancreatic cancer, to test its classification power. Its performance was compared with that of 5 conventional multivariate and machine learning methods: random forest (RF), support vector machine (SVM), logistic regression (LR), k-nearest neighbors (k-NN), and naïve Bayes (NB). The DL method yielded the best classification (AUC 0.9472 for the test data set) of all approaches. We also optimized the parameters of DL individually to determine which factors were the most significant. In summary, the DL method has advantages in classifying the quantitative proteomics data of pancreatic cancer patients, and our results suggest that its implementation can improve the performance of diagnostic assays in clinical settings.



KEYWORDS: *deep learning, machine learning, SRM−MS, targeted proteomics, mass spectrometry*

■ INTRODUCTION

Deep learning (DL) has received much attention in relation to many real-world tasks. In this era of big data, the transformation of large quantities of data into valuable knowledge has become increasingly important in various domains—specifically, biomedical data.[1] Significant amounts of biomedical data, including omics and imaging data, have been accumulated, and their potential for application in biological and health care research has garnered the interest of industry and academia. For instance, IBM developed Watson for Oncology, a platform that analyzes patients' medical information and assists clinicians in considering treatment options.[2] In addition, Google DeepMind, having achieved great success with AlphaGo in the game of Go,[3,4] recently launched DeepMind Health to develop effective health care technologies.[5]

There are many advantages of DL over shallow machine learning methods. Conventional machine learning methods require engineering domain knowledge to create features from raw data,[6−8] whereas DL automatically extracts simple features from the input data using a general-purpose learning procedure.[9] These simple features are mapped into outputs using a complex architecture that is composed of a series of nonlinear functions—hierarchical representations—to maximize the predictive accuracy of the model. By increasing the number of layers and neurons per layer, robust features can be constructed, and error signals can be diminished as they pass through multiple layers. Thus, DL assembles high-level transformed features from input data, rendering it more desirable than shallow machine learning algorithms in this respect.[10]

Omics data constitute one of the most prominent examples of feature-rich and high-dimensional data with complex multilevel structures.[11−13] Hence, DL, through its multilayer representation learning models, is a promising new approach for the analysis of omics data. In particular, DL is a new class of machine learning method that has been applied to various areas of genomic research, including inferring expression profiles of genes and predicting the functional activity of genomic sequences.[14] In another study that was based on gene expression data, DL outperformed linear regression with regard to inference of the expression of target genes from that of landmark genes.[15]

Compared with their growing application in various genomic fields, DL-based machine learning methods have only been

Received: April 25, 2019
Published: July 17, 2019

DOI: 10.1021/acs.jproteome.9b00268

applied to qualitative proteomics,[16−18] not quantitative proteomics. To address this issue, we used a DL method for quantitative proteomic data sets in the diagnosis of pancreatic cancer and analyzed then by selected reaction monitoring−mass spectrometry (SRM−MS).[19] We developed an optimized DL-based analytical approach that enabled us to classify patients with pancreatic cancer. The performance of our DL approach was then compared with that of typical machine learning methods: random forest (RF), support vector machine (SVM), logistic regression (LR), k-nearest neighbor (k-NN), and naïve Bayes (NB).

Unlike other machine learning algorithms, our DL tool has several hyperparameters, and its classification and prediction depends on how they are tuned.[9,20] To this end, we optimized 10 hyperparameters: the number of epochs, the number of nodes and hidden layers, activation function, rho, epsilon, L1 & L2 regularization, hidden dropout ratio, input dropout ratio, train samples per iteration, and max w2. Consequently, we determined the appropriate hyperparameter for the implementation of our DL method and identified important parameters that affect its performance.

In this study, we applied a DL method to classify quantitative proteomics data. We examined 5 metrics (recall, precision, F_1 score, accuracy, and the area under the receiver operating characteristic curve) in analyzing the performance of our approach and other machine learning methods in predicting pancreatic cancer patients from an SRM−MS data set. We found that our DL method performs better than many machine learning methods. Our results also demonstrate the major advantages of new data-centric approaches that are powered by DL over traditional models. Our application of DL to quantitative proteomics data suggests its applicability to new domains. The goal of this work is to facilitate the clinical application of DL in quantitative proteomics.

■ EXPERIMENTAL SECTION

Experimental Design

This study used quantitative proteomic data sets that were derived from a previous study,[19] from which we obtained 1008 samples (300 normal control, NC; 109 pancreatic benign, PB; 49 other benign, OB; 149 other cancers, OC; and 401 pancreatic adenocarcinoma, PDAC). The entire data set was split into two subsets—a training and test data set—at a ratio of approximately 0.7 (691 samples; 322 PDAC, 41 OB, 88 PB, and 240 NC) to 0.3 (317 samples; 79 PDAC, 8 OB, 149 OC, 21 PB, and 60 NC), respectively.

The entire data set was split into 2 subsets—a training and test data set—by random partitioning at a fixed ratio of approximately 0.7 (691 samples; 322 PDAC, 41 OB, 88 PB, and 240 NC) to 0.3 (317 samples; 79 PDAC, 8 OB, 149 OC, 21 PB, and 60 NC), respectively. These proportions are common for moderately sized samples in machine learning applications and similar to what has been used elsewhere.[21,22] We included a cohort of OCs in only the test data set, including thyroid, colon, and breast, to determine whether our model was affected by tumor heterogeneity.

To designate the output of each algorithm as the ability to discriminate pancreatic cancer (PDAC) from nonpancreatic cancer, the data set was reconfigured into a control group by combining NC, PB, OB, and OC; PDAC was defined as the case group. A total of 34 peptides (derived from 26 proteins) were measured by SRM−MS for all plasma samples. The protocols and SRM−MS procedure are detailed in a previous study.[19]

Deep Learning Method

In this study, a feedforward deep neural network model for class prediction was established using the SRM−MS data set. Ten-fold crossvalidation was used in processing the training data set as the technique for constructing the model to avoid sampling bias. In each iteration, approximately 622 data points (691 × 0.9) were randomly selected from the subtraining data set and input into the model; the remaining 69 values (691 × 0.1) were used to as the subtest data set to evaluate errors in the model while maintaining equal proportions of the control and case groups for each selected data point (stratified sampling). To construct the model, we applied the stepwise method to reduce the computational burden of testing all possible feature sets.

The training data set was used to fine-tune the model and optimize the parameters. The trained models were then tested on an independent test data set, and their performance with regard to classification was evaluated. The performance of the model was further validated with an independent test data set. The test data set was not used to construct the model. The performance of the test data set was used to guide the optimization of the parameters. To reduce the likelihood of sample selection bias and model overfitting, we also performed bootstrapping validation in addition to crossvalidation, with 100 iterations (0.9 ratio for the training data set).

The training and test data sets were processed using the H2O package, version 3.10.3.6 in R language, ver. 3.3.3 (R Foundation for Statistical Computing, Vienna, Austria). Choosing the appropriate hyperparameters for DL is crucial in applying it properly. Hyperparameter tuning was performed on the 10 most important parameters in the DL method (number of epochs, number of nodes and hidden layers, activation function, rho, epsilon, L1 & L2 regularization, hidden dropout ratio, input dropout ratio, train samples per iteration, and max w2).[23] We could have conducted a grid search to optimize the values for each parameter simultaneously. However, to determine the parameters that were significant, we optimized them one by one with the values that are commonly used for each parameter. All procedures in the optimization of the parameters are detailed in the Supporting Information (SI).

Other Machine Learning Methods

We compared the performance of the DL method with that of alternative approaches. Although many machine learning methods are used for classification, only 5 representative methods are recommended by the proteomics community and applied widely: logistic regression (LR), naïve Bayes (NB), k-nearest neighbor (kNN), support vector machine (SVM), and random forest (RF).[24] In modeling these methods, the training and test data sets were processed as with the DL method. However, to obtain the optimal hyperparameter, we performed a grid search through a predefined hyperparameter space in R package to tune the parameters in the 5 methods.[25]

Data Analysis

Five traditional measures of model performance were used: recall, precision, F_1 score, accuracy, and the area under the receiver operating characteristic curve (AUROC). For the metric definitions, we used the following abbreviations: the number of true positives (TP), the number of false positives

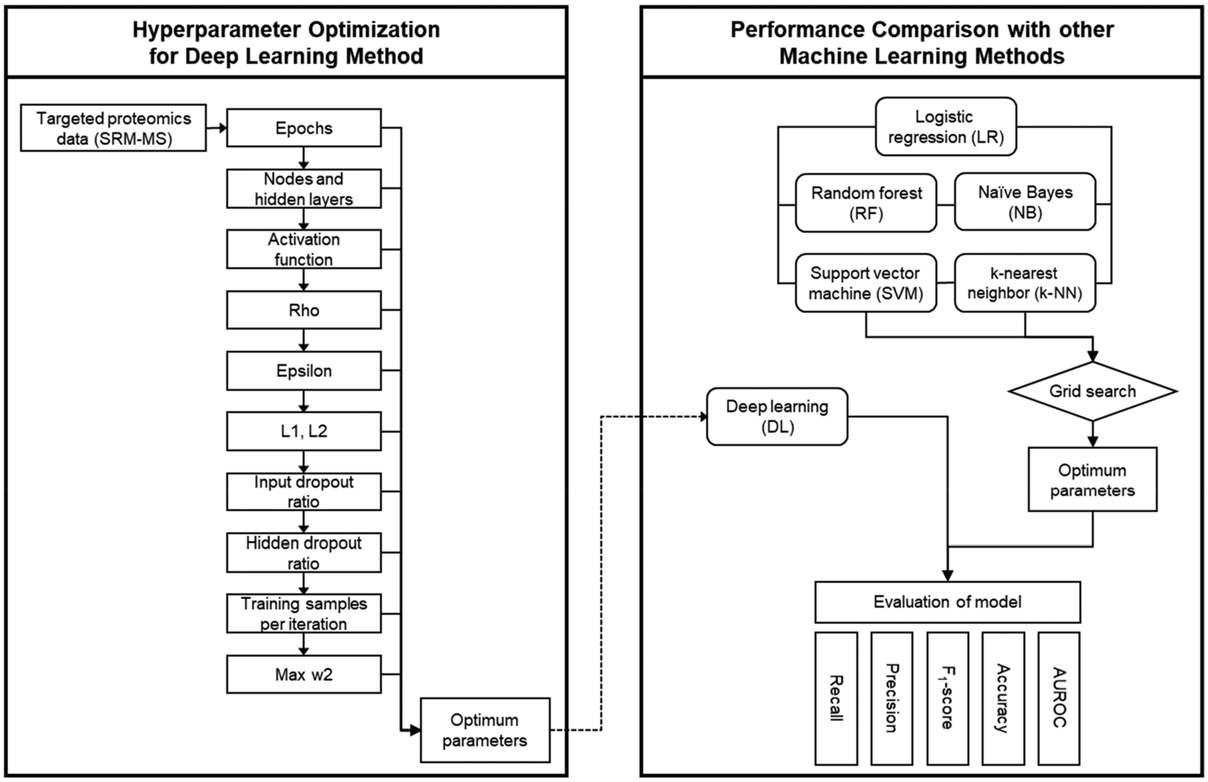


Figure 1. Overall scheme of the study for tuning hyperparameters and assessing model performance. Targeted proteomic data were obtained through selected reaction monitoring−mass spectrometry (SRM−MS) analysis in a previous study. The optimization of the parameters of deep learning was performed one by one for 10 important factors. Five representative machine learning algorithms (logistic regression, LR; random forest, RF; naive Bayes, NB; support vector machine, SVM; k-nearest neighbor, k-NN) used grid search to obtain the optimal hyperparameters. Then, model performance was evaluated and compared based on 5 metrics (recall, precision, F_1 score, accuracy, and AUROC).

(FP), the number of true negatives (TN), and the number of false negatives (FN). Model recall (also known as the true positive rate or sensitivity) can be considered the percentage of true class labels that are correctly identified by the model as true and is defined as Recall = $\frac{\mathrm{TP}}{\mathrm{TP}+\mathrm{FN}}$. Similarly, model precision (also known as the positive predictive value) is the probability that a predicted true label is true and is defined as Precision = $\frac{\mathrm{TP}}{\mathrm{TP}+\mathrm{FP}}$. The F_1 score is simply the harmonic mean of the recall and precision. Thus, this value is defined as F_1 score = $2 \times \frac{\mathrm{Recall} \times \mathrm{Precision}}{\mathrm{Recall} + \mathrm{Precision}}$. Accuracy is another measure of the all-around robustness of a model and is the percentage of correctly identified labels of the entire population: Accuracy = $\frac{\mathrm{TP}+\mathrm{TN}}{\mathrm{TP}+\mathrm{TN}+\mathrm{FP}+\mathrm{FN}}$. AUROC can be computed by plotting the recall versus the false positive rate (FPR) at various decision thresholds, where FPR = $\frac{\mathrm{FP}}{\mathrm{FP}+\mathrm{TN}}$. In this study, all models could assign a probability estimate of a sample that belonged to the true class. Thus, we constructed an AUROC curve by measuring the recall and FDR from this plot, where 1.0 denotes perfect separation and 0.5 is random classification. All computations were performed on a dual-core (Intel Core i7−7500U CPU; 3.5 GHz) server with 16 GB memory that ran local Windows 10.

■ RESULTS

Hyperparameter Optimization for Deep Learning Method

We determined the predictive ability of the DL method to separate pancreatic cancer patients from controls using quantitative proteomics data; the overall workflow of our study is shown in Figure 1. We trained the DL model using a wide range of parameters and selected the best model with the optimal values for 5 metrics (specifically with regard to AUROC). We evaluated the parameters that affected the performance of the DL-based model. The hyperparameter optimization is detailed in the SI.

Of the 10 parameters, epoch, activation function, epsilon, and input dropout ratio affected the classification performance of the DL method (Figure 2). One parameter—the epoch, ranging from 1−1000—was tested. As a result, the difference in AUROC for the test data set, based on the epochs, revealed a maximum-minimum of 0.3104. The best performance (AUROC of 0.9453) was found at epoch 400. Among the 3 types of activation function, the rectifier performed best, based on the AUROC of the test data set. Between 1.0×10^{-4} to 1.0×10^{-10}, the best performance was observed at epsilon 1.0×10^{-8}. An input dropout ratio of 0.0 was found to be the most performance.

Six parameters—nodes and hidden layers, rho, L1 & L2 regularization, hidden dropout ratio, train samples per iteration, and max w2—barely affected the classification performance of the DL method. The difference in maximum−minimum AUROC in the test data set for these parameters was 0.0182, 0.0205, 0.0266, 0.0054, 0.0183, and 0.0005, respectively, and there was almost no change in performance (Tables S1 and S2 and Figure S1).

In the hyperparameter optimization, the execution time increased by up to approximately 14 times compared with the number of nodes and hidden layers, which was larger than the

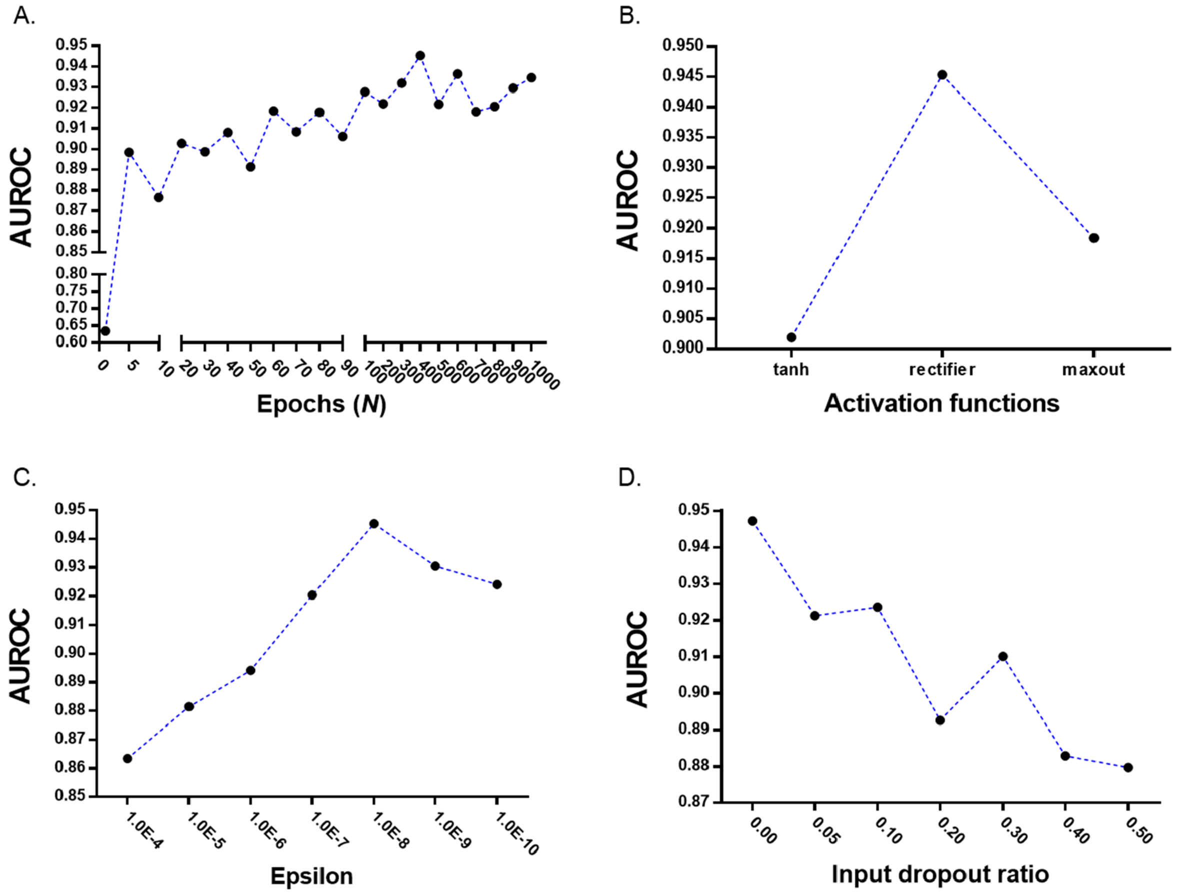


Figure 2. Optimization results for hyperparameters that were completely affected by model performance of deep learning. The hyperparameters of deep learning were optimized, and their performance was evaluated based on the AUROC value in the independent test dataset (the *x*-axis corresponds to parameter values, and the *y*-axis refers to AUROC values). Epoch showed the best performance with AUROC at epochs 400 (A). Activation function showed the best performance with AUROC in rectifier (B). The epsilon value (C), and input dropout ratio (D) were optimized at 1.0 × 10^{-8} and 0.0 (no applied), respectively.

epoch-induced changes in execution time on the values of the tested parameters (Figure S2A,B). In addition, the rectifier executes faster than other activation functions.[26] In this study, the rectifier had a faster execution rate that was 1.32−3.07 times that of other activation functions (Figure S2C).

DL-Based Method Outperforms Alternative Approaches

We compared the performance of the DL with that of 5 popular machine learning approaches: LR, NB, k-NN, SVM, and RF (Table 1). The execution time of the DL tool was approximately 37.5 times longer versus the other 5 methods; there was no significant difference in execution time between the 5 methods. The DL method selected 14 features for the optimal model, similar to the LR and SVM methods, which selected 15 features, and the NB approach (13 features). However, the k-NN and RF methods selected 6 features—far fewer than the other machine learning methods.

For the description of features, we used the gene symbol of the protein, followed by the peptide sequence. Thirteen features were shared between DL and 1 or more of the machine learning methods. Among them, 1 peptide (IGFBP2.LIQGAPTIR) was shared by all methods, 4 peptides (TTHY.AADDTWEPFASGK, LRG1.DLLLPQPDLR, KLKB1.DSVTGTLPK, and C5.NADYSYSVWK) were shared by 4 methods, 3 peptides (ITIH4.AGFSWIEVTFK, CLU.ASSIIDELFQDR, and PROS1.NNLELSTPLK) were common to 3 methods, 3 peptides (SEPP1.CINQLLCK, C1R.DYFIATCK, and CFI.VFSLQWGEVK) were shared by 2 approaches, and 2 peptides (BTD.ILSGDPYCEK and BTD.LSSGLVTAALYGR) were shared by 1 method. Additionally, DL identified 1 unique peptide (IPSP.GFQQLLQELNQPR).

With regard to classification performance, the DL method had the best overall performance for 5 metrics—recall, precision, F_1 score, accuracy, and AUROC—in the training and independent test data sets (Figure 3). In summary, the DL method yielded a recall of 0.9114, a precision of 0.6923, an F_1 score of 0.7869, an accuracy of 0.8770, and an AUROC value of 0.9472 in the test data set. Consequently, the performance of the DL method improved by 5.88% for recall, 6.90% for precision, 6.46% for F_1 score, 3.35% for accuracy, and 3.73% for AUROC compared with the LR method, the next-best performing approach. In addition, compared with RF method, which had the worst performance, DL improved its performance by 60.0% for recall, 21.54% for precision, 38.14% for F_1 score, 11.65% for accuracy, and 16.05% for AUROC.

By bootstrapping validation, the deep learning algorithm performed best on the test data set compared with the other algorithms but slightly worse than the crossvalidation (Table S3). Further, because having fewer features will be more practical (lower cost) in clinical applications, the performance of the algorithm was assessed with the number of features limited to 3, 6, and 10 (Table S4). Consequently, DL outperformed 5 other popular machine learning methods, even

Table 1. Comparison of Performance between Deep Learning and Five Machine Learning Methods by Crossvalidation

	Machine Learning Method	Logistic Regression (LR)	Naïve Bayes (NB)	k-Nearest Neighbor (k-NN)	Support Vector Machine (SVM)	Random Forest (RF)	Deep Learning (DL)
Training dataset	Execution time (h:m:s)	< 00:02:00	< 00:02:00	< 00:02:00	< 00:02:00	< 00:02:00	01:25:05
	Number of optimized features (N)	15	13	6	15	6	14
	Detailed information of optimal model	Coefficient of feature and intercept: TTHY.AADDTWEPFASGK = -1.138 ITIH4.AGFSWIEVTFK = -0.584 CLU.ASSIIDELFQDR = -0.412 LRG1.DLLLPQPDLR = 1.630 KLKB1.DSVTGTLPK = -4.796 C1R.DYFIATCK = 2.244 MBL2.FQASVATPR = 1.246 ITIH4.LALDNGGLAR = 0.249 IGFBP2.LIQGAPTIR = 6.334 C5.NADYSYSVWK = 0.530 PROS1.NNLELSTPLK = 1.369 CLU.TLLSNLEEAK = -0.857 LRG1.VAAGAFQGLR = -1.081 CFI.VFSLQWGEVK = 0.667 IGFBP3.YGQPLPGYTTK = 0.828 Intercept = -0.174	Mean parameter of control vs case: TTHY.AADDTWEPFASGK = 2.844 vs 1.889 IGFBP3. ALAQCAPPPAVCAELVR = 0.545 vs 0.470 CLU.ASSIIDELFQDR = 2.163 vs 1.950 SEPP1.CINQLLCK = 0.940 vs 0.820 HRG.DGYLFQLLR = 3.417 vs 3.160 LRG1.DLLLPQPDLR = 1.214 vs 2.452 KLKB1.DSVTGTLPK = 0.817 vs 0.741 MBL2.FQASVATPR = 0.280 vs 0.516 IGFBP2.LIQGAPTIR = 0.147 vs 0.273 C5.NADYSYSVWK = 5.099 vs 7.468 PROS1.NNLELSTPLK = 0.740 vs 0.836 CLU.TLLSNLEEAK = 2.513 vs 2.341 SERPINC1.VAEGTQVLELPFK = 2.351 vs 2.304	[a]Configuration of model: ITIH4.AGFSWIEVTFK, IGFBP3.ALAQCAPPPAVCAELVR, LRG1.DLLLPQPDLR, KLKB1.DSVTGTLPK, BTD.ILSGDPYCEK, and IGFBP2.LIQGAPTIR	Configuration of (kernel) model: TTHY.AADDTWEPFASGK = -0.459 ITIH4.AGFSWIEVTFK = -0.361 CLU.ASSIIDELFQDR = -0.286 SEPP1.CINQLLCK = -0.232 LRG1.DLLLPQPDLR = 0.307 KLKB1.DSVTGTLPK = -0.720 C1R.DYFIATCK = 0.427 MBL2.FQASVATPR = 0.366 ITIH4.LALDNGGLAR = 0.074 IGFBP2.LIQGAPTIR = 0.610 BTD.LSSGLVTAALYGR = 0.223 C5.NADYSYSVWK = 1.003 PROS1.NNLELSTPLK = 0.326 CLU.TLLSNLEEAK = -0.532 CFI.VFSLQWGEVK = 0.357	Weight of features: LRG1.VAAGAFQGLR = 0.178 ECM1.ELLALIQLER = 0.251 CLU.TLLSNLEEAK = 0.100 IGFBP2.LIQGAPTIR = 0.103 C5.NADYSYSVWK = 0.169 TTHY.AADDTWEPFASGK = 0.199	[a]Configuration of model: TTHY.AADDTWEPFASGK, ITIH4.AGFSWIEVTFK, CLU.ASSIIDELFQDR, SEPP1.CINQLLCK, LRG1.DLLLPQPDLR KLKB1.DSVTGTLPK, C1R.DYFIATCK, IPSP.GFQQLLQELNQPR, BTD.ILSGDPYCEK, IGFBP2.LIQGAPTIR, BTD.LSSGLVTAALYGR, C5.NADYSYSVWK, PROS1.NNLELSTPLK, and CFI.VFSLQWGEVK
	TN (*N*)	327	336	317	336	327	324
	FN (*N*)	59	116	89	72	131	52
	FP (*N*)	42	33	52	33	42	45
	TP (*N*)	263	206	233	250	191	270
	Recall	.6970	.5152	.6667	.6970	.3939	.8182
	Precision	.9200	.9444	.7586	.9583	.8125	.9643
	F_1 score	.7931	.6667	.7097	.8070	.5306	.8852
	Accuracy	.8261	.7536	.7391	.8406	.6667	.8986
	AUROC	.9184	.8973	.8552	.9310	.8047	.9301
Test dataset	**TN (*N*)**	201	210	193	206	204	206
	FN (*N*)	11	27	19	18	34	7
	FP (*N*)	37	28	45	32	34	32
	TP (*N*)	68	52	60	61	45	72
	Recall	.8608	.6582	.7595	.7722	.5696	.9114
	Precision	.6476	.6500	.5714	.6559	.5696	.6923
	F_1 score	.7391	.6541	.6522	.7093	.5696	.7869
	Accuracy	.8486	.8265	.7981	.8423	.7855	.8770
	AUROC	.9131	.8788	.8882	.9064	.8162	.9472

[a]TN, true negative; FN, false negative; FP, false positive; TP, true positive. [b]Inner workings are difficult to understand and explain.

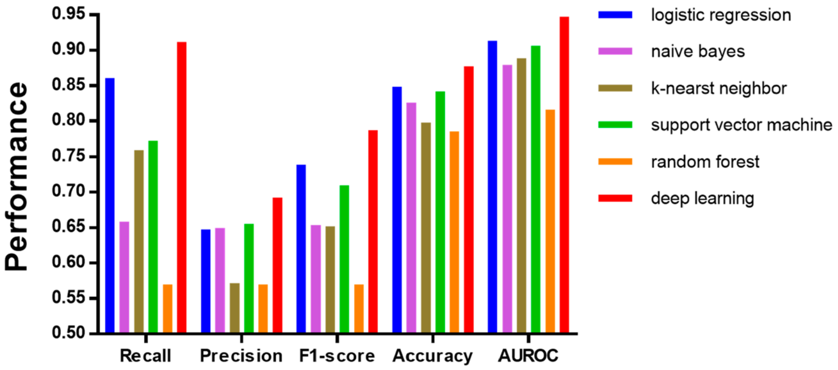


Figure 3. Bar graph of model performance in the independent test data set. Performance was compared between deep learning (DL, in red) and five traditional machine learning methods: logistic regression (LR, in blue), naïve Bayes (NB, in purple), k-nearest neighbor (k-NN, in brown), support vector machine (SVM, in green), and random forest (RF, in orange). Measurements of model performance were evaluated based on recall, precision, F_1 score, accuracy, and AUROC. DL method showed the best performance for all five metrics in the independent test dataset.

with fewer features on a limited scale. These findings indicate that the DL method is reliable and can be applied to the prediction of novel biomarkers using targeted proteomics data.

DISCUSSION

Big data can be examined by DL, a state-of-the-art machine learning technique that is based on deep multilayered neural networks.[9] Omics has naturally become an application field of DL, specifically in genomics. However, there has been no systematic examination of DL in the (quantitative) proteomics space. In this report, we aimed to address this issue and assessed the performance of DL in classifying proteomic biomarkers. We are the first group to use the DL framework to integrate targeted proteomics information in biomarker research. In this study, we used an independent test data set to compare several algorithms. In addition, we compared 5 metrics to assess performance, unlike most studies in this field, which have incorporated on 1 or 2. We found that DL outperforms 5 other popular machine learning methods in classifying pancreatic cancer patients using SRM−MS data.

Despite the outstanding performance of the DL method, there are several caveats in its application in quantitative proteomics research. Because DL is a time-consuming computation relative to other machine learning methods (Tables 1 and S2), it requires high-performance computer hardware. Also, proteomic data sets are generally small compared with genomic and image data.[27,28] The major limitation of our study is that although DL performed better with 34 features and 1008 samples, it is generally implemented with more features and samples, requiring further evaluation of its performance.

Finally, DL must tune various parameters for its proper use. The grid search tunes the hyperparameters to determine the optimal values for a given model. During training, the model is tuned using a grid search strategy to identify the tuned parameters that effect the best performance.[29] Generally, a grid search is used to determine the optimal combination of hyperparameters in DL,[30] but it is difficult to know which element is important.

A chief criticism against DL is that it is used as a "black box" (Table 1): although it produces outstanding results, we know little about how they are derived internally. In bioinformatics, particularly in biomedical domains, it is insufficient to simply produce good outcomes. Because many studies are connected to patient health, it is crucial to providing logical reasoning, just as clinicians do for medical decisions.

However, we believe that the ability of DL to predict different groups with high accuracy will engender entirely new data processing options, supporting and enhancing future targeted proteomics-based biomarker research. DL is also expected to be valuable for discovery proteomics data sets that have many features, which will be useful in biomarker studies with comprehensive proteome wide predictions of bioinformatic workflows. Thus, we anticipate that the DL algorithm will benefit the scientific community and have an impact on proteomics research.

CONCLUSIONS

DL is a rapidly advancing field with constantly emerging methods and technologies that might be readily adaptable to biomedicine. Our results demonstrate that the implementation of DL using targeted proteomics data is a powerful tool for biomarker validation. The high degree of various metrics indicates the potential for improved standardization and interrater reliability for disease classification tasks in the clinical laboratory. Future work should optimize its performance in a clinical environment toward full implementation of the DL method as a clinical tool.

ASSOCIATED CONTENT

Supporting Information

The Supporting Information is available free of charge on the ACS Publications website at DOI: 10.1021/acs.jproteome.9b00268.

Supplementary Methods: Hyperparameter optimization of deep learning method; Table S1. Hyperparameter settings for optimization of deep learning method; Table S2. Performance of deep learning method according to hyperparameter settings; Table S3. Comparison of performance between deep leaning method and five machine learning methods by bootstrapping validation; Table S4. Comparison of performance between deep leaning method and five machine learning methods according to number of features; Figure S1. Optimization of hyperparameters that were barely affected by the performance of deep learning; and Figure S2. Execution time according to optimization of 3 hyperparameters of the deep learning method (PDF)

AUTHOR INFORMATION

Corresponding Authors

*Phone: 82-10-8338-6719. E-mail: jangjy4@snu.ac.kr.
*Phone: 82-10-5351-1959. E-mail: biolab@snu.ac.kr.

ORCID

Youngsoo Kim: 0000-0001-8881-0662

Author Contributions

H.K. contributed to study concept and statistical analysis; H.K., Y.K., B.H. contributed to administrative, technical, or material support; J.-Y.J. and Y.K. contributed to obtained funding; H.K. and Y.K. contributed to drafting of the manuscript.

Author Contributions

⊥These authors contributed equally to this work.

Notes

The authors declare no competing financial interest.

ACKNOWLEDGMENTS

This work was supported by the Industrial Strategic Technology Development Program (#10079271 and #2000134), the Collaborative Genome Program for Fostering New Post-Genome Industry (NRF-2017M3C9A5031597), funded from the Korean Government and the Korea Health Technology R&D Project (No. HI14C2640).

ABBREVIATIONS

DL, deep leaning; SRM−MS, selected reaction monitoring−mass spectrometry; RF, random forest; SVM, support vector machine; LR, logistic regression; k-NN, k-nearest neighbor; NB, naïve Bayes; AUROC, area under the receiver operating characteristic curve; TP, true positives; FP, false positives; TN, true negatives; FN, false negatives; FPR, false positive rate

REFERENCES

(1) Topol, E. J. High-performance medicine: the convergence of human and artificial intelligence. *Nat. Med.* **2019**, *25* (1), 44−56.

(2) Schmidt, C. M. D. Anderson Breaks With IBM Watson, Raising Questions About Artificial Intelligence in Oncology. *J. Natl. Cancer Inst* **2017**, *109* (5), No. djx113-djx113.

(3) Silver, D.; Huang, A.; Maddison, C. J.; Guez, A.; Sifre, L.; van den Driessche, G.; Schrittwieser, J.; Antonoglou, I.; Panneershelvam, V.; Lanctot, M.; Dieleman, S.; Grewe, D.; Nham, J.; Kalchbrenner, N.; Sutskever, I.; Lillicrap, T.; Leach, M.; Kavukcuoglu, K.; Graepel, T.; Hassabis, D. Mastering the game of Go with deep neural networks and tree search. *Nature* **2016**, *529* (7587), 484−9.

(4) Silver, D.; Schrittwieser, J.; Simonyan, K.; Antonoglou, I.; Huang, A.; Guez, A.; Hubert, T.; Baker, L.; Lai, M.; Bolton, A.; Chen, Y.; Lillicrap, T.; Hui, F.; Sifre, L.; van den Driessche, G.; Graepel, T.; Hassabis, D. Mastering the game of Go without human knowledge. *Nature* **2017**, *550* (7676), 354−359.

(5) Powles, J.; Hodson, H. Google DeepMind and healthcare in an age of algorithms. *Health Technol. (Berl)* **2017**, *7* (4), 351−367.

(6) Kim, H.; Yu, S. J.; Yeo, I.; Cho, Y. Y.; Lee, D. H.; Cho, Y.; Cho, E. J.; Lee, J. H.; Kim, Y. J.; Lee, S.; Jun, J.; Park, T.; Yoon, J. H.; Kim, Y. Prediction of Response to Sorafenib in Hepatocellular Carcinoma: A Putative Marker Panel by Multiple Reaction Monitoring-Mass Spectrometry (MRM-MS). *Mol. Cell. Proteomics* **2017**, *16* (7), 1312−1323.

(7) Yu, S. J.; Kim, H.; Min, H.; Sohn, A.; Cho, Y. Y.; Yoo, J. J.; Lee, D. H.; Cho, E. J.; Lee, J. H.; Gim, J.; Park, T.; Kim, Y. J.; Kim, C. Y.; Yoon, J. H.; Kim, Y. Targeted Proteomics Predicts a Sustained Complete-Response after Transarterial Chemoembolization and Clinical Outcomes in Patients with Hepatocellular Carcinoma: A Prospective Cohort Study. *J. Proteome Res.* **2017**, *16* (3), 1239−1248.

(8) Kim, H.; Park, J.; Kim, Y.; Sohn, A.; Yeo, I.; Jong Yu, S.; Yoon, J. H.; Park, T.; Kim, Y. Serum fibronectin distinguishes the early stages of hepatocellular carcinoma. *Sci. Rep* **2017**, *7* (1), 9449.

(9) LeCun, Y.; Bengio, Y.; Hinton, G. Deep learning. *Nature* **2015**, *521* (7553), 436−44.

(10) Mamoshina, P.; Vieira, A.; Putin, E.; Zhavoronkov, A. Applications of Deep Learning in Biomedicine. *Mol. Pharmaceutics* **2016**, *13* (5), 1445−54.

(11) Kim, S.; Oesterreich, S.; Kim, S.; Park, Y.; Tseng, G. C. Integrative clustering of multi-level omics data for disease subtype discovery using sequential double regularization. *Biostatistics* **2017**, *18* (1), 165−179.

(12) Kim, H.; Sohn, A.; Yeo, I.; Yu, S. J.; Yoon, J. H.; Kim, Y. Clinical Assay for AFP-L3 by Using Multiple Reaction Monitoring-Mass Spectrometry for Diagnosing Hepatocellular Carcinoma. *Clin. Chem.* **2018**, *64* (8), 1230−1238.

(13) Sohn, A.; Kim, H.; Yeo, I.; Kim, Y.; Son, M.; Yu, S. J.; Yoon, J. H.; Kim, Y. Fully validated SRM-MS-based method for absolute quantification of PIVKA-II in human serum: Clinical applications for patients with HCC. *J. Pharm. Biomed. Anal.* **2018**, *156*, 142−146.

(14) Kelley, D. R.; Snoek, J.; Rinn, J. L. Basset: learning the regulatory code of the accessible genome with deep convolutional neural networks. *Genome Res.* **2016**, *26* (7), 990−9.

(15) Chen, Y.; Li, Y.; Narayan, R.; Subramanian, A.; Xie, X. Gene expression inference with deep learning. *Bioinformatics* **2016**, *32* (12), 1832−9.

(16) Tran, N. H.; Zhang, X.; Xin, L.; Shan, B.; Li, M. De novo peptide sequencing by deep learning. *Proc. Natl. Acad. Sci. U. S. A.* **2017**, *114* (31), 8247−8252.

(17) Kim, M.; Eetemadi, A.; Tagkopoulos, I. DeepPep: Deep proteome inference from peptide profiles. *PLoS Comput. Biol.* **2017**, *13* (9), No. e1005661.

(18) Zhou, X. X.; Zeng, W. F.; Chi, H.; Luo, C.; Liu, C.; Zhan, J.; He, S. M.; Zhang, Z. pDeep: Predicting MS/MS Spectra of Peptides with Deep Learning. *Anal. Chem.* **2017**, *89* (23), 12690−12697.

(19) Park, J.; Choi, Y.; Namkung, J.; Yi, S. G.; Kim, H.; Yu, J.; Kim, Y.; Kwon, M. S.; Kwon, W.; Oh, D.-Y.; et al. Diagnostic performance enhancement of pancreatic cancer using proteomic multimarker panel. *Oncotarget* **2017**, *8* (54), 93117−93130.

(20) Koutsoukas, A.; Monaghan, K. J.; Li, X.; Huan, J. Deep-learning: investigating deep neural networks hyper-parameters and comparison of performance to shallow methods for modeling bioactivity data. *J. Cheminform.* **2017**, *9* (1), 42.

(21) Shi, L.; Campbell, G.; Jones, W. D.; Campagne, F.; Wen, Z.; Walker, S. J.; Su, Z.; Chu, T. M.; Goodsaid, F. M.; Pusztai, L.; et al. The Micro Array Quality Control (MAQC)-II study of common practices for the development and validation of microarray-based predictive models. *Nat. Biotechnol.* **2010**, *28* (8), 827−838.

(22) Consortium, S. M.-I. A comprehensive assessment of RNA-seq accuracy, reproducibility and information content by the Sequencing Quality Control Consortium. *Nat. Biotechnol.* **2014**, *32* (9), 903−914.

(23) Cook, D. Practical Machine Learning with H2O. *Powerful, Scaleable Technique for Deep Learning and AI*, 1st ed.; O'REILLY Media Press: Sebastopol, CA, 2016; pp 187−224.

(24) Zhang, Y.; Xin, Y.; Li, Q.; Ma, J.; Li, S.; Lv, X.; Lv, W. Empirical study of seven data mining algorithms on different characteristics of datasets for biomedical classification applications. *Biomed Eng. Online* **2017**, *16* (1), 125.

(25) Kuhn, M. Building Predictive Models in R Using the caret Package. *Journal of Statistical Software* **2008**, *28* (5), 1−26.

(26) Wang, P.; Ge, R.; Xiao, X.; Cai, Y.; Wang, G.; Zhou, F. Rectified-Linear-Unit-Based Deep Learning for Biomedical Multi-label Data. *Interdiscip. Sci.: Comput. Life Sci.* **2017**, *9* (3), 419−422.

(27) Wang, M.; Tai, C.; E, W.; Wei, L. DeFine: deep convolutional neural networks accurately quantify intensities of transcription factor-DNA binding and facilitate evaluation of functional non-coding variants. *Nucleic Acids Res.* **2018**, *46* (11), No. e69.

(28) Ting, D. S. W.; Cheung, C. Y.; Lim, G.; Tan, G. S. W.; Quang, N. D.; Gan, A.; Hamzah, H.; Garcia-Franco, R.; San Yeo, I. Y.; Lee, S. Y.; Wong, E. Y. M.; Sabanayagam, C.; Baskaran, M.; Ibrahim, F.; Tan, N. C.; Finkelstein, E. A.; Lamoureux, E. L.; Wong, I. Y.; Bressler, N. M.; Sivaprasad, S.; Varma, R.; Jonas, J. B.; He, M. G.; Cheng, C. Y.; Cheung, G. C. M.; Aung, T.; Hsu, W.; Lee, M. L.; Wong, T. Y. Development and Validation of a Deep Learning System for Diabetic Retinopathy and Related Eye Diseases Using Retinal Images From Multiethnic Populations With Diabetes. *JAMA* **2017**, *318* (22), 2211−2223.

(29) Welchowski, T.; Schmid, M. A framework for parameter estimation and model selection in kernel deep stacking networks. *Artif Intell Med.* **2016**, *70*, 31−40.

(30) James Bergstra, Y. B. Random Search for Hyper-Parameter Optimization. *J. Machine Learning Res.* **2012**, *13*, 281−305.