DeepUEP: Prediction of Urine Excretory Proteins Using Deep Learning

WEI DU¹, (Member, IEEE), RAN PANG¹, GAOYANG LI¹,³, HUANSHENG CAO⁴, YING LI¹, and YANCHUN LIANG¹,²

¹Key Laboratory of Symbol Computation and Knowledge Engineering, Ministry of Education, College of Computer Science and Technology, Jilin University, Changchun 130012, China
²Zhubai Laboratory of Key Laboratory of Symbol Computation and Knowledge Engineering, Ministry of Education, Zhubai College of Jilin University, Zhubai 519041, China
³School of Life Science and Technology, Tongji University, Shanghai 200092, China
⁴Center for Fundamental and Applied Microbiomics, Biodesign Institute, Arizona State University, Tempe, AZ 85287, USA

Corresponding authors: Ying Li (liying@jlu.edu.cn) and Yanchun Liang (ycliang@jlu.edu.cn)

This work was supported in part by the National Natural Science Foundation of China under Grant 61872418, Grant 61972174, and Grant 61972175, in part by the Natural Science Foundation of Jilin Province under Grant 20180101050JC and Grant 20180101331JC, in part by the Guangdong Key-Project for Applied Fundamental Research under Grant 2018KZDXM076, in part by the Guangdong Premier Key-Discipline Enhancement Scheme under Grant 2016GDYSZDXK036, and in part by the Jilin Provincial Education Department Project under Grant JJKH20180467KJ.

ABSTRACT Urine excretory proteins are among the most commonly used biomarkers in body fluids. Computational identification of urine excretory proteins can provide very useful information for identifying targeted disease biomarkers in urine by linking transcriptome or proteomics data. There are few methods based on conventional machine learning algorithms for predicting urine excretory proteins, and most of these methods strongly depend on the extraction of features from urine excretory proteins. An end-to-end model for urine excretory protein prediction, called DeepUEP, is presented using deep neural networks relying on only amino acid sequence information. The model achieves good performance and outperforms existing methods on training and testing sets. By comparing known urinary protein biomarkers with the results of the model, we find that the model can achieve a true-positive rate of over 80% for urinary protein biomarkers that have been detected in more than one study. We also combine our model with transcriptome and proteomics data from lung cancer patients to predict the potential urinary protein biomarkers of lung cancer. A web server is developed for the prediction of urine excretory proteins, and it can be accessed at the following URL: http://www.csbg-jlu.info/DeepUEP/. We believe that our prediction model and web server are useful for biomedical researchers who are interested in identifying urinary protein biomarkers, especially for candidate proteins in transcriptome or proteomics analyses of diseased tissues.

INDEX TERMS Bioinformatics, biomedical computing, machine learning, data mining.

I. INTRODUCTION Disease biomarkers play important roles in detecting diseases as well as the mechanisms of disease occurrence and development, especially for early diagnosis, which is essential for controlling and preventing disease [1]. These diagnostic biomarkers can be objectively measured and evaluated to differentiate normal samples from disease samples [2]. The rapid advancement of omics techniques make it possible to detect biomarkers in body fluid for specific human diseases in a systematic and comprehensive manner [3]. Therefore, the use of bioinformatics technology to systematically and efficiently identify clinically detectable, the use of highly sensitive and highly specific clinical biomarkers from big data has become a popular topic of current research. At present, the detection of biomarkers in body fluids such as blood, urine and saliva is an effective way to diagnose diseases [4]–[6]. Because there are many signals for various physiological and pathophysiological conditions in the blood, most studies on body fluid biomarkers focus on blood [7]–[10].

With the improvement of the proteomic analyses of urine samples in the last few years, researchers have found that urine is also an ideal source for detecting human diseases [5]. Compared to blood, urine is a better source of biomarkers...
because the composition of urine is relatively simple and it can be easily and noninvasively obtained [11]. Urine excretory proteins contain soluble proteins and protein components of solid-phase elements [12]. The soluble proteins consist of plasma proteins involved in glomerular filtration and soluble proteins secreted by epithelial cells, which account for ∼49% of all urine excretory proteins. Currently, ∼40% of these soluble proteins are secreted by epithelial cells, and the remaining ∼60% are plasma proteins involved in glomerular filtration [13]. Most solid-phase components, which account for ∼48% of all urine excretory proteins, are generated from epithelial cells by whole-cell shedding and plasma membrane and intracellular component shedding [12], [14], [15]. Only ∼3% of urine excretory proteins are generated by exosome secretion [16]. The urine excretory proteins in this article are defined as in the literature [3] and mainly include plasma proteins involved in glomerular filtration, soluble proteins secreted by epithelial cells and proteins associated with plasma membrane and intracellular component shedding.

These urine excretory proteins can be used as biomarkers in urological diseases and reflect the statuses of other organs that are distant from the urinary system [17]. Most current studies focus on using urinary proteomics to discover the biomarkers of urogenital diseases, including different kinds of urogenital cancers, such as renal cell carcinoma [18], prostate cancer [19] and bladder cancer [20], but several recent studies have shown that some proteins in urine are related to systemic diseases such as ovarian carcinoma [21], lung cancer [22], hepatocellular carcinoma [23] and gastric cancer [3].

At present, numerous studies have identified a variety of disease biomarkers in urine through comparative proteomic analyses of urinary samples from patients with a specific disease and from control groups. Because there are many low-abundance proteins in urine and the dynamic ranges of these proteins span several orders of magnitude, comparative and quantitative analysis of proteomic data from urine samples is a very challenging process [19]. Due to the limitations of proteomic experimental techniques, the discovery of urinary protein biomarkers faces many problems [12]. Therefore, developing a computational method for accurately predicting urine excretory proteins could solve these problems to a certain extent. However, there is very little research focused on building a computational model for predicting urine excretory proteins [3], [24]. The existing methods can be carried out in two main steps: (1) selecting features from constructed feature sets; and (2) applying a machine learning algorithm for training by using the selected features. The processes of feature engineering and feature selection may result in incomplete or biased features [25]. End-to-end deep-learning algorithms can automatically learn feature representations and predict results from data, going directly from input to output, which enables deep learning to surpass conventional machine learning techniques [26]–[29].

Here, we present DeepUEP, a novel deep-learning model for predicting urine excretory proteins. Different from existing urine excretory protein prediction methods, DeepUEP makes predictions directly based on amino acid sequences that are encoded as profile matrices by using the Position-Specific Iterative Basic Local Alignment Search Tool (PSI-BLAST) [30]. The framework of DeepUEP mainly consists of a convolutional neural network (CNN) module that can extract short motifs from input profiles, a recurrent neural network (RNN) module with long short-term memory (LSTM) cells that can extract the long spatial dependencies between amino acids, and an attention module that assigns higher importance to amino acids that are relevant for the prediction. We validate that our model achieves a good accuracy on the training dataset and the independent test dataset (91.25% on the training dataset; 88.98% on the independent test dataset). Our experiments also show that DeepUEP outperformed existing methods based on conventional machine learning.

DeepUEP, in combination with transcriptome and proteomics data, can provide very useful information for the detection of targeted disease biomarkers in urine. By comparing known urinary protein biomarkers with our model results, we find that our model can achieve a true-positive rate of over 80% for urinary protein biomarkers that have been detected in more than one study. We also combine our model with transcriptome and proteomic data from lung cancer patients to predict the potential urinary protein biomarkers of lung cancer. A web server is developed for predicting urinary excretion proteins, and it can be accessed at the following URL: http://www.csbg-jlu.info/DeepUEP/. We believe that our predictive model and web server are useful for biomedical researchers interested in identifying urinary protein biomarkers, especially when they have candidate proteins for analysing diseased tissues using transcriptome or proteomics data. The main contributions of this paper are as follows: (1) A novel deep-learning model is proposed that achieves good performance and outperforms existing methods for urine excretory protein prediction using only the amino acid sequence; and (2) a user-friendly web server is developed for biomedical researchers.

II. METHODS

A. DATA COLLECTION

There are several existing datasets of proteins that can be detected in urine. First, we collected proteins that are detected in urine from the Sys-BodyFluid database [31] and the Human Proteome Project (HPP) [32]. The Sys-BodyFluid database includes 1,941 proteins that have been detected experimentally in nine urinary proteomic studies. The goal of the HPP is to experimentally observe all the proteins produced from the human genome. Currently, more than 17,000 proteins have been observed via mass spectrometry (MS) or non-MS experiments [33]. Meanwhile, we also used urine excretory proteins generated by other urinary proteomic studies [34]–[36]. We find a total of 3133 different urine excretory proteins, and 1638 of these proteins are observed in only one study. These results indicate that
there are great differences in the urine excretory proteins of different experiments. Therefore, the 1495 proteins detected in more than one study were retained. Then, to avoid learning bias due to information redundancy, we remove proteins with a mutual sequence similarity higher than 30% using the tool CD-HIT [37]. Finally, 1350 proteins are used as the positive data; 1000 of these proteins are used as positive training data, while the remaining 350 proteins are used as positive test data.

Because experiments that can verify which proteins are not urine excretory proteins do not exist, generating a negative dataset is a challenge. In this study, we apply a process similar to that proposed by Cui et al. [38] and choose human proteins from Pfam families that do not include any proteins that have been detected in urine. To decrease the influence of protein families with small protein numbers, we choose proteins from families with at least ten proteins. For such families, ten members are selected to compose the negative data. Then, we remove the proteins with a mutual sequence similarity higher than 30% using the tool CD-HIT. As a result, 1362 proteins are selected for the negative data, 1000 of which are used as negative training data, while the remaining 362 proteins are used as the negative test data.

B. ENCODING PROTEINS

Given protein sequences, the model predicts whether these proteins are urine excretory proteins. This problem can be formulated as a binary classification problem, and each protein can be classified as either a urine excretory protein or a non-urine excretory protein. In this study, the protein sequences are encoded using profiles constructed by PROFILpro [39] to search for the sequence in the UniRef50 database [40]. The profile is actually a position-specific scoring matrix [41] based on the amino acid frequencies at every position of a multiple alignment using PSI-BLAST [30].

In practice, the sequence lengths of different proteins are different. To reduce the training time, we set the maximum protein length to 1000. For proteins with lengths less than 1000, the end of the matrix is filled with zeros. For proteins longer than 1000, we selected 500 amino acids from the beginning (N-terminus) and 500 amino acids from the end (C-terminus) of the protein. According to this rule, only 11.74% of human proteins are truncated. Because most information on urine excretory proteins is stored at the beginning (N-terminus) and the end (C-terminus) of the sequence, this selection will retain the most information [3], [42], [43].

C. PREDICTION MODEL

The prediction model based on a deep-learning framework is mainly composed of a CNN module, an RNN module with LSTM cells and an attention module that is the same as that in [44]. The detailed architecture of the prediction model is shown in Fig. 1.

The constructed prediction model can predict urine excretory proteins using the profiles generated by protein sequences as described above as input. The specific objective function is defined as follows:

\[ y = f(X; \theta) \]  

where \( y \) is the predicted result of whether a protein is a urine excretory protein, \( f \) is the prediction model based on the parameters \( \theta \), and \( X \) is the input profile of each protein, which is a matrix with \( T \) rows and \( N \) columns. In this study, \( T \) is the sequence length (1000), and \( N \) is size of the amino acid vocabulary (20).

First, to identify the local information of proteins, that is, the protein motifs, a CNN module is applied in the prediction model. In DeepUEP, the first convolutional layer in the CNN module contains 80 convolution kernels of different sizes (10 filters for each of the following sizes: 1, 3, 5, 9, 15, 21, 27, and 33). Each kernel serves as a ‘feature filter’ that can swipe on the input profiles and detect motifs regardless of their position in the sequence. The weights of each convolution kernel are adjusted during model training to find the motifs that can improve urine excretory protein prediction. In addition, the second convolutional layer contains 64 convolution kernels of size 3 \( \times \) 80, and it handles the 1000 \( \times \) 80 feature map obtained from the first convolutional layer. The output of the second convolutional layer is a 1000 \( \times \) 64 feature map. The generated feature map from the CNN module represents a protein sequence in a more abstract way, and it has been indicated to be beneficial for protein classification in combination with other deep-learning architectures, such as RNNs [45].

Therefore, to further extract the long spatial dependencies between the detected motifs from the CNN module, the RNN module is applied in the prediction model. However, the original RNN encounters gradient vanishing and gradient explosion problems. Therefore, we use LSTM, which is normally augmented by recurrent gates (input gates, output gates, and forget gates), to prevent back-propagating errors from gradient vanishing and gradient explosion [46]. Furthermore, to better calculate the relationship between one motif and other motifs in the ends of the sequence, the bidirectional LSTM (BLSTM) is used in the prediction model.
BLSTM combines forward LSTM and backward LSTM to preserve the upstream and downstream information of the protein sequence by combining the two hidden states [47]. In DeepUEP, the RNN module analyses the 1000 × 64 feature map from the CNN module using 256 LSTM units in both directions, extracting the long spatial dependencies between motifs and yielding a 1000 × 512-dimensional output.

However, the original RNN and LSTM are not ideal for long sequences for the following reasons. 1) The encoder of the original RNN and LSTM compresses all the information of the input sequence into a fixed-length hidden vector $c$, ignoring the length of the sequence. When the input sequence is long, the performance of these models is very low [48]–[50]. 2) It is unreasonable to encode the input sequence to a fixed length and give the same weight to each position in the sequence. The attention mechanism, which focuses on the part of the position from a long sequence according to the current state, was proposed to solve these problems [51]. Therefore, we also apply the attention module in the prediction model to identify important regions of protein sequences for classification.

In DeepUEP, the attention module contains an LSTM with 512 units through 10 decoding steps and an attention mechanism feedforward neural network (FFN) with 256 units. We define the hidden state of the LSTM encoder at a sequence position $j$ as $h_j$, and the last hidden state of the encoder as $h_T$, which is the input of the attentive decoder $f$. If we suppose that $s_i$ is the hidden state of the decoder in the $i$ decoding steps, then $s_i$ can be computed by the following formula:

$$s_i = f(h_T, s_{i-1}, c_i)$$

(2)

where $c_i$ is the weighted average of the hidden states of the LSTM encoder, which is calculated as follows:

$$c_i = \sum_{j=1}^{T} \alpha_{ij} h_j$$

(3)

The weight $\alpha_{ij}$ of each $h_j$ is computed by the following formula:

$$\alpha_{ij} = \frac{\exp(e_{ij})}{\sum_{k=1}^{T} \exp(e_{ik})}$$

(4)

where $e_{ij}$, which scores how well the hidden state of the encoder around position $j$ and the hidden state of the decoder at position $i$ match, is defined as follows:

$$e_{ij} = v^T \tanh(W_d s_{i-1} + W_e h_j)$$

(5)

where $W_d$, $W_e$ and the column vector $v$ are the trainable parameters of the attention function. The attentive decoder $f$ is run for $K$ decoding steps, and $c_K$ obtained in the last step is used as the input of the output module for predicting urine excretory proteins. In DeepUEP, the output module consists of a fully connected dense layer with 512 units and a softmax layer for binary classification.

D. MODEL TRAINING

The parameters of our model are optimized using the Adam stochastic optimization method [52] with the following parameters: a learning rate of 0.001, a decay rate for the first-moment estimates of 0.9, and an exponential decay rate for the second-moment estimate of 0.999. Cross-entropy loss between the true and predicted distributions of urine excretory proteins is applied as the loss function. The proposed and compared models are evaluated on a workstation with the Ubuntu 18.04 LTS system and equipped with an Intel Core i7-7800X CPU with 128 GB of RAM and an NVIDIA GeForce RTX 2080Ti GPU.

In the deep-learning framework, deep neural networks with a large number of parameters are very powerful [53]. However, it is difficult to address overfitting in these frameworks, which combine different types of neural networks. In this study, we use three strategies to reduce overfitting in the model. The first strategy involves adding multiple dropout layers in the model. Dropout, which is a technology used to solve overfitting problems in deep-learning frameworks, randomly discards elements in the neural network during training, which can prevent the units from adapting too much. During training, samples are taken from different sparse networks with different index numbers [54], and during testing, an unthinned network with a smaller weight can easily be used to approximate the average of all these thinning network predictions. The second strategy for reducing overfitting is to use regularization [55] in the model. The training optimization algorithm is a function consisting of two terms: a loss term, which is used to measure the degree to which the model fits the data, and a regularization term, which is used to measure the complexity of the model and prevent overfitting. In this study, L2 regularization is adopted to prevent overfitting in the model. The third strategy for reducing overfitting is to use early stopping during the training iteration. Specifically, during model training, if the loss on validation data is not reduced in 50 epochs, then the training procedure will be stopped [56].

III. RESULTS

A. PERFORMANCE MEASUREMENTS

To estimate the performance of the prediction model, we use the following measures: accuracy, sensitivity (recall), specificity, precision, F-score, Matthews correlation coefficient (MCC), and the area under the receiver operating characteristic (ROC) curve (AUC).

$$\text{accuracy} = \frac{TP + TN}{N_{total}}$$

(6)

$$\text{sensitivity} = \text{recall} = \frac{TP}{TP + FN}$$

(7)

$$\text{specificity} = \frac{TN}{TN + FP}$$

(8)

$$\text{precision} = \frac{TP}{TP + FP}$$

(9)
TABLE 1. Comparison of the performances of different models on the training dataset by 10-fold cross-validation.

| Method               | Sensitivity | Specificity | Precision | Accuracy | F-score | MCC     | AUC     |
|----------------------|-------------|-------------|-----------|----------|---------|---------|---------|
| BLSTM                | 81.43%      | 85.52%      | 84.39%    | 83.52%   | 0.829   | 0.670   | 0.869   |
| BLSTM_Attention      | 90.77%      | 87.52%      | 87.88%    | 89.14%   | 0.893   | 0.783   | 0.941   |
| CNN_BLSTM            | 82.77%      | 85.71%      | 84.98%    | 84.26%   | 0.839   | 0.685   | 0.887   |
| CNN_BLSTM_Attention  | 91.76%      | 90.73%      | 90.92%    | 91.25%   | 0.913   | 0.825   | 0.942   |

\[
F\text{-score} = 2 \cdot \frac{\text{precision} \cdot \text{recall}}{\text{precision} + \text{recall}}
\]

\[
\text{MCC} = \frac{(TP \times TN - FP \times FN)}{\sqrt{(TP + FN)(TP+FP)(TN+FP)(TN+FN)}}
\]

where \(TP\) is the number of true positives, \(TN\) is the number of true negatives, \(FP\) is the number of false positives, \(FN\) is the number of false negatives, and \(N_{\text{total}}\) is the number of total samples in the dataset. The F-score is the harmonic mean of precision and recall, and it is maximized at 1 and minimized at 0. The MCC can be used as a measure of the quality of binary classifications [57]. Its value is between \(-1\) and \(+1\), where \(+1\) means perfect classification, \(0\) means no better than random classification, and \(-1\) signifies total disagreement between the prediction and the true condition.

The ROC curve is a graphical plot that illustrates the classification ability of a binary classifier by plotting the true-positive rate (TPR) against the false-positive rate (FPR) with various discrimination thresholds [58]. When using normalized units, the AUC can be between 0 and 1, and it is equal to the probability that a classifier will rank a randomly chosen positive instance higher than a randomly chosen negative one. In brief, the larger the AUC value, the higher the correct rate. The precision-recall curve is a plot of the precision and the recall with various discrimination thresholds.

B. MODEL SELECTION

To select the model, we compare the performances of different models on the training dataset using 10-fold cross-validation. For all models, the same training sets and validation sets are used. The comparison results of the performance using different models are shown in Table 1. The average ROC and precision-recall curves for different models are plotted in Fig. 2.

FIGURE 2. ROC curves and precision-recall curves of different models on the training dataset by 10-fold cross-validation.

CNN_BLSTM_Attention and BLSTM_Attention achieve satisfactory predictions of urine excretory proteins with accuracies of 91.25% and 89.14%, respectively. Comparing these results with those of CNN_BLSTM and BLSTM (accuracies of 84.26% and 83.52%), we can see that the attention mechanism greatly improves the prediction performance. Moreover, by comparing BLSTM_Attention and CNN_BLSTM_Attention and BLSTM and CNN_BLSTM, we can see that the CNN module also improves the prediction performance to some extent. As shown in Table 1 and Fig. 2, the CNN_BLSTM_Attention model outperforms the other models for predicting urine excretory proteins, with average sensitivity, specificity, precision, accuracy, F-score, MCC and AUC values of 91.76%, 90.73%, 90.92%, 91.25%, 0.913, 0.825 and 0.942, respectively. Therefore, we decided to use this model for the rest of the experiments.

C. EVALUATING THE PERFORMANCE OF DEEPUEP

To evaluate DeepUEP against other methods based on conventional machine learning algorithms, 10-fold cross-validation is performed on the training dataset. The same training sets and validation sets are used for all methods. Among these methods, the SVM-radial basis function (SVM-RBF) method was proposed by Hong et al. [3]. Since the authors did not provide the relevant source code or program, we implement the feature selection process and the traditional prediction model of SVM-RBF ourselves based on the description in the reference. Furthermore, for a more comprehensive and systematic comparison, we also build prediction models employing k-nearest neighbours (KNN), decision tree, random forest, adaptive boosting (AdaBoost) and linear SVM based on the selected features. The details of these methods are shown in Supplement 1.

FIGURE 3. ROC curves and precision-recall curves of different methods on the training dataset by 10-fold cross-validation.

The average ROC and precision-recall curves for different methods are plotted in Fig. 3, which shows that DeepUEP
TABLE 2. The performance of DeepUEP and other methods on the training set.

| Method       | Sensitivity | Specificity | Precision | Accuracy | F-score | MCC  | AUC   |
|--------------|-------------|-------------|-----------|----------|---------|------|-------|
| KNN          | 84.49%      | 78.49%      | 79.71%    | 81.49%   | 0.820   | 0.631| 0.881 |
| Decision Tree| 75.20%      | 79.77%      | 78.81%    | 77.49%   | 0.770   | 0.550| 0.824 |
| Random Forest| 76.10%      | 67.75%      | 70.24%    | 71.93%   | 0.731   | 0.440| 0.792 |
| AdaBoost     | 79.05%      | 83.15%      | 82.43%    | 81.10%   | 0.807   | 0.622| 0.897 |
| SVM-Linear   | 85.52%      | 83.84%      | 84.11%    | 84.68%   | 0.851   | 0.694| 0.918 |
| SVM-RBF      | 89.66%      | 81.94%      | 83.24%    | 85.80%   | 0.863   | 0.718| 0.928 |
| DeepUEP      | 91.76%      | 90.73%      | 90.92%    | 91.25%   | 0.913   | 0.825| 0.942 |

TABLE 3. The performance of DeepUEP and other methods on the independent testing set.

| Method       | Sensitivity | Specificity | Precision | Accuracy | F-score | MCC  | AUC   |
|--------------|-------------|-------------|-----------|----------|---------|------|-------|
| KNN          | 84.57%      | 77.90%      | 78.72%    | 81.18%   | 0.815   | 0.626| 0.867 |
| Decision Tree| 79.71%      | 71.55%      | 73.04%    | 75.56%   | 0.762   | 0.514| 0.799 |
| Random Forest| 80.57%      | 64.64%      | 68.78%    | 72.47%   | 0.742   | 0.457| 0.790 |
| AdaBoost     | 85.43%      | 73.76%      | 75.89%    | 79.49%   | 0.804   | 0.595| 0.876 |
| SVM-Linear   | 84.29%      | 85.08%      | 84.53%    | 84.69%   | 0.851   | 0.694| 0.915 |
| SVM-RBF      | 82.28%      | 88.12%      | 87.00%    | 85.25%   | 0.835   | 0.706| 0.918 |
| DeepUEP      | 89.05%      | 88.92%      | 88.27%    | 88.98%   | 0.887   | 0.780| 0.932 |

outperformed the methods based on conventional machine learning algorithms. The performance metrics of DeepUEP and other methods are shown in Table 2. The average sensitivity, specificity, precision, accuracy, F-score, MCC and AUC values by 10-fold cross-validation are 91.76%, 90.73%, 90.92%, 91.25%, 0.913, 0.825 and 0.942, respectively. As shown in Table 2, DeepUEP performs better than other methods. The performance distribution of DeepUEP on each fold of cross-validation is generally desirable, ranging from 87.79% to 96.83% for sensitivity and 85.38% to 96.00% for specificity.

Then, we use the entire training dataset, which contains 1000 instances of positive data and 1000 of negative data, to train our model and the other methods based on conventional machine learning, and we use the independent testing set containing 350 positive and 362 negative data instances for predictions. The validation set is selected from training data and is completely separate from the test set. The ROC curves and the precision-recall curves are plotted in Fig. 4, which shows that DeepUEP performs much better than other methods in terms of both ROC and precision-recall estimators. The performance metrics of DeepUEP and other methods are shown in Table 3. The sensitivity, specificity, precision, accuracy, F-score, MCC and AUC values of DeepUEP on the independent testing set are 89.05%, 88.92%, 88.27%, 88.98%, 0.887, 0.780 and 0.932, respectively. Thus, on the independent testing sets, we find that DeepUEP outperformed the methods based on conventional machine learning algorithms.

D. PREDICTING AND RANKING THE KNOWN URINE EXCRETORY PROTEINS

We rank 20,186 human proteins that have been reviewed in the Universal Protein Resource (UniProt) database [59] using the S-value, which is defined as follows:

\[ S = 2 \times (\arg \max(p) - 0.5) \times \max(p) \quad (12) \]

where \( FC_{ij} = \begin{cases} fc_{ij} - 1 & (fc_{ij} \geq 1) \\ 1 - \frac{1}{fc_{ij}} & (fc_{ij} < 1) \end{cases} \) is the result of the softmax output from the proposed model, and the values of \( FC_{ij} = \begin{cases} fc_{ij} - 1 & (fc_{ij} \geq 1) \\ 1 - \frac{1}{fc_{ij}} & (fc_{ij} < 1) \end{cases} \) are 0 and 1, which are the indices of the maximum values. In total, 908 (90.80%) of the positive training data are ranked among the top 6000 proteins. Among these 6000 proteins, 312 (89.14%) are in the positive test data (350).

We also rank the urinary protein biomarkers that have been associated with human diseases in the published literature and that do not overlap with the training dataset. In this study, we collect these proteins from the Urinary Protein Biomarker Database [60]. This database was built by manually collecting existing studies of urinary protein biomarkers from...
the published literature. There are 261 human biomarkers of excretory proteins in urine that have been reviewed in UniProt. These proteins are relevant to different diseases, such as Dent’s disease, Kawasaki disease, nephropathy, and several cancers, including bladder cancer, prostate cancer and gastric cancer. To remove the effects of false-positive biomarkers obtained by proteomics [61], we also count the biomarkers that have been detected in more than one study. There are 135 and 78 biomarkers that have been detected in two or more and three or more studies, respectively. All these biomarkers are shown in Table S1. The true-positive rates of these biomarkers from different numbers of studies for different top numbers are shown in Fig. 5, which shows that our approach yields better results for biomarkers that have been detected in more than one study. The details of the results are shown in Table S2, which contains the number of biomarkers and the true-positive rate included in the top proteins of different numbers.

Moreover, 111 (82.22%) and 70 (89.74%) of these 135 and 78 biomarkers are ranked among the top 6000, respectively. The p-values for the hypergeometric probability of the corresponding rankings are 9.34E-37 and 1.41E-28, respectively. Table 4 also shows the results obtained by using the SVM-RBF method [3]. In the table, the top number of 6000 is approximately equal to the threshold that achieved the performance values in Table 3. The value is also approximately equal to the number of proteins in healthy urine identified in the existing literature [62].

### Table 4. Comparison of the ranking results of human urinary protein biomarkers.

| Top number | SVM-RBF | DeepUEP |
|------------|---------|---------|
|            | N | P-value       | N | P-value |
| 500        | 0 | -              | 12 | 4.03E-07 |
| 1000       | 2 | 0.1549        | 22 | 1.36E-11 |
| 1500       | 6 | 0.1669        | 31 | 1.14E-15 |
| 2000       | 9 | 0.1256        | 40 | 2.62E-20 |
| 2500       | 11 | 0.1165      | 47 | 1.50E-23 |
| 3000       | 13 | 0.1094       | 50 | 4.08E-23 |
| 3500       | 17 | 0.0655       | 56 | 3.83E-26 |
| 4000       | 17 | 0.0993       | 59 | 2.52E-26 |
| 4500       | 35 | 5.20E-6      | 62 | 8.85E-27 |
| 5000       | 51 | 3.32E-14     | 64 | 2.38E-26 |
| 5500       | 53 | 6.26E-14     | 66 | 4.01E-26 |
| 6000       | 54 | 5.26E-13     | 70 | 1.41E-28 |

|                                     | SVM-RBF | DeepUEP |
|-------------------------------------|---------|---------|
| Top number                          | N | P-value       | N | P-value |
| 500                                 | 0 | -              | 12 | 4.03E-07 |
| 1000                                | 2 | 0.1549        | 22 | 1.36E-11 |
| 1500                                | 6 | 0.1669        | 31 | 1.14E-15 |
| 2000                                | 9 | 0.1256        | 40 | 2.62E-20 |
| 2500                                | 11 | 0.1165      | 47 | 1.50E-23 |
| 3000                                | 13 | 0.1094       | 50 | 4.08E-23 |
| 3500                                | 17 | 0.0655       | 56 | 3.83E-26 |
| 4000                                | 17 | 0.0993       | 59 | 2.52E-26 |
| 4500                                | 35 | 5.20E-6      | 62 | 8.85E-27 |
| 5000                                | 51 | 3.32E-14     | 64 | 2.38E-26 |
| 5500                                | 53 | 6.26E-14     | 66 | 4.01E-26 |
| 6000                                | 54 | 5.26E-13     | 70 | 1.41E-28 |

1 The number of urinary protein biomarkers in selected top number

### E. Application to Lung Cancer for Identification of Urinary Protein Biomarkers

Several researchers have predicted urinary protein biomarkers indicating lung cancer through experiments. Nolen et al. analysed the biomarkers in the urine of non-small-cell lung carcinoma (NSCLC) patients and obtained 36 biomarkers for diagnosing lung cancer, which are given in Table S4 [64]. Using our model, 32 of the 36 (88.88%) biomarkers are identified as urine excretory proteins.

To further validate the feasibility of DeepUEP for identifying the urinary protein biomarkers of cancer, we perform differential expression analyses based on proteomic data and transcriptomic data from lung cancer. For the proteomic data and transcriptomic data, we apply the Wilcoxon signed-rank test and the fold change to detect differentially expressed proteins and genes between lung cancer samples and adjacent normal samples. The Wilcoxon signed-rank test is used to assess the statistical significance of the observed differential expression in cancer vs. normal samples, and the statistical significance cut-off value is set to 0.05. We calculate the expression fold change using the following formula [65]:

\[
FC_{ij} = \begin{cases} 
\frac{f_{cij}}{f_{cij} - 1} & (f_{cij} \geq 1) \\
1 - \frac{1}{f_{cij}} & (f_{cij} < 1)
\end{cases}
\]

where \(f_{cij} = C_{ij}/N_{ij}\) (13)

\[
FC_i = \frac{1}{m} \sum_{j=1}^{m} FC_{ij}
\]

(14)

where \(m\) is the sample number, \(C_{ij}\) and \(N_{ij}\) indicate the expression of the cancer sample and adjacent normal sample, respectively, for patient \(j\) on protein (gene) \(i\). The values of \(FC_i\) are greater than zero for upregulated proteins (genes) and less than zero for downregulated proteins (genes). In this study, we applied 0.5 as the threshold on the absolute value of 

### Figure 5. The true-positive rates for different top numbers.

Pathways are the carbon metabolism, biosynthesis of antibiotics and metabolic pathways (see Table S3).
be used as biomarkers to not only detect urological diseases but also diagnose systemic diseases. At present, urinary protein biomarkers of diseases are identified through comparative proteomic analyses of urinary samples from patient and control groups. However, urine could have many proteins, and the dynamic range in the abundance of these proteins could span a few orders of magnitude. Thus, comparative analyses of proteomic data of urine samples can be very challenging. Therefore, if a computational method for predicting urine excretory proteins by combining transcriptome or proteomics data could be developed, then the urinary protein biomarkers of different diseases could be identified.

At present, few studies have focused on building computational models for predicting urine excretory proteins. In this research, we present a computational model based on a deep-learning framework to accurately predict urine excretory proteins. The deep neural network-based, end-to-end prediction model, which consists of a convolutional neural network module, a recurrent neural network module with long short-term memory cells and an attention module, can recognize urine excretory proteins using only amino acid sequence information. Specifically, this paper proposes a predictive model based on a deep-learning framework for predicting urine excretory proteins using only amino acid sequence information, and the performance of this model is better than that of the method proposed by Hong et al [3].

We analyse the most important positions in the sequence that the attention mechanism focuses on, and the results are shown in Fig. 6; the x-axis is the sequence position. For proteins longer than 1000 amino acids, the middle part of the protein is removed, and for proteins less than 1000 amino acids in length, the sequences are padded from the middle to align the N-terminus and C-terminus.

F. WEB SERVER OF DEEPUEP
A web server is developed for the prediction of urine excretory proteins, and it can be accessed at the following URL: http://www.csbg-jlu.info/DeepUEP/. The web server provides the following two functional modules: (1) urine excretory protein prediction based on protein sequences; and (2) prediction result browsing for all human proteins. The web server of DeepUEP supports sequences in FASTA format as input. Users can input sequences in the text area or upload a FASTA file. The original results can be downloaded conveniently. Each prediction task will be assigned a Job ID, and users can use the Job ID to download previous results.

IV. DISCUSSION AND CONCLUSION
The identification of disease-related biomarkers is an important, effective method for the early diagnosis of diseases, which plays an important role in the prevention and control of diseases. Diagnostic biomarkers are biological molecules that are produced by various diseases and can be used to distinguish normal samples from diseased samples. With the rapid development of genomics technology, the use of bioinformatics technology to easily and systematically detect clinical biomarkers with high sensitivity and specificity from big data has become a popular research topic. In recent years, as proteomic analysis has been developed, various clinical disease biomarkers have been discovered in body fluids. However, most of these biomarkers are found in blood, and only a few are urinary biomarkers.

Urinary biomarkers are superior to blood biomarkers because urine is relatively simple in composition, readily available, and non-invasively collected. Urine excretory proteins mainly originate from the urinary system or blood by glomerular filtration. Therefore, urine excretory proteins can be used as biomarkers to not only detect urological diseases
on feature selection and considered most excreted proteins to have specific signal peptides similar to those of proteins secreted through the endoplasmic reticulum (ER). From Fig. 6, we can see that there is also some attention at the C-terminus, which could indicate the presence of retention signals, such as the ER retention signal encoded by the amino acid sequence KDEK (Lys-Asp-Glu-Leu) or HDEL (His-Asp-Glu-Leu) [42], [43]. Hong et al. also indicated that the secondary structure types of proteins, such as the percentage of alpha helices in a protein sequence, are also prominent features of urinary excretory proteins [3] and the charge of a protein is also a key feature of urine excretory proteins. Charge is a factor for determining which proteins can be filtered through the glomerular membrane. Wang et al. found that transmembrane domains and the radius of gyration are also prominent features of urine excretory proteins [24]. Published studies have observed that proteins with a radius smaller than 1.8 nm can pass through the GBM slit diaphragm barrier, whereas proteins with a radius larger than 4.0 nm are retained.

**ACKNOWLEDGEMENT**

The authors thank the support from their affiliations.

**REFERENCES**

[1] R. Liu, X. Wang, K. Aihara, and L. Chen, “Early diagnosis of complex diseases by molecular biomarkers, network biomarkers, and dynamical network biomarkers,” *Med. Res. Rev.*, vol. 34, no. 3, pp. 455–478, May 2014.

[2] R. Mayeux, “Biomarkers: Potential uses and limitations,” *NeuroRx*, vol. 1, no. 2, pp. 182–188, Apr. 2004.

[3] C. S. Hong, J. Cui, Z. Ni, Y. Su, D. Puett, F. Li, and Y. Xu, “A computational method for prediction of excretory proteins and application to identification of gastric cancer markers in urine,” *PLoS ONE*, vol. 6, no. 2, Feb. 2011, Art. no. e16875.

[4] Y. H. Lee and D. T. Wang, “Saliva: An emerging biofluid for early detection of diseases,” *Amer. J. Dentistry*, vol. 22, no. 4, pp. 241–248, Aug. 2009.

[5] M. Nagpal, S. Singh, P. Singh, P. Chauhan, and M. A. Zaidi, “Tumor markers: A diagnostic tool,” *Nat. J. Maxillofacial Surg.*, vol. 7, no. 1, pp. 17–20, Jan./Jun. 2016.

[6] S. Nair, K. D. Tang, L. Kenny, and C. Punyadeera, “Salivary exosomes as a source of urinary biomarkers,” *Nephrology-lined Proteomics*, vol. 5, no. 10, pp. 1760–1771, Aug. 2009.

[7] S.-J. Li, M. Peng, H. Li, B.-S. Liu, C. Wang, J.-R. Wu, Y.-X. Li, P. Legrain, and N. Nonomura, “Discovery of novel urinary biomarker of bladder cancer by proteomic analysis of urinary and tissue-exudative exosomes,” *Int. Urol.*, vol. 25, no. 328, Oct. 2018.

[8] D. V. Thongboonkerd, “Practical points in urinary proteomics,” *J. Proteome Res.*, vol. 6, no. 10, pp. 3881–3890, Oct. 2007.

[9] D. Zhou, P. S. T. Yuen, T. Pisitkun, P. A. Gonzales, H. Yasuda, J. W. Dear, P. Gross, M. A. Knepper, and R. A. Star, “Collection, storage, preservation, and normalization of human urinary exomes for biomarker discovery,” *Kidney Int.*, vol. 69, no. 8, pp. 1471–1476, Apr. 2006.

[10] C. Shao, M. L. Li, X. D. Li, L. L. Wei, L. S. Zhu, F. Yang, L. L. Jia, Y. Mu, J. N. Wang, Z. G. Guo, D. Zhang, J. R. Yin, Z. G. Wang, W. Sun, Z. G. Zhang, and Y. H. Gao, “A tool for biomarker discovery in the urinary proteome: A manually-curated human and animal urine protein biomarker database,” *Febbs J.*, vol. 279, p. 231, Sep. 2012.

[11] M. A. Rogers, P. Clarke, J. Noble, N. P. Munro, A. Paul, P. J. Selby, and R. E. Banks, “Proteomic profiling of urinary proteins in renal cancer by surface enhanced laser desorption ionization and neural-network analysis: Identification of key issues affecting potential clinical utility,” *Cancer Res.*, vol. 63, no. 20, pp. 6971–6983, Oct. 2003.

[12] D. Theodorescu, E. Schiffer, H. W. Bauer, F. Douwes, F. Eichhorn, R. Polley, T. Schmidt, W. Schöfer, P. Zürbig, D. M. Good, J. J. Coon, and H. Mischak, “Discovery and validation of urinary biomarkers for prostate cancer,” *Proteomics- Clin. Appl.*, vol. 2, no. 4, pp. 556–570, Apr. 2008.

[13] K. Matsuzaki, K. Fujita, T. Shiomizu, R. Narumi, Y. Hayashi, K. Jingushi, T. Kato, A. Kawashima, T. Ujike, A. Nagahara, M. Uemura, T. Tomonaga, and N. Nonomura, “Discovery of novel urinary biomarker of bladder cancer by proteomic analysis of urinary and tissue-exudative exosomes,” *Int. Urol.*, vol. 25, no. 328, Oct. 2018.

[14] J. B. Liu, Y. Y. Yip, E. M. Swisher, K. Agnew, K. E. Hellstrom, and I. Hellstrom, “Detection of the HE4 protein in urine as a biomarker for ovarian neoplasms: Clinical correlates,” *Gynecol. Oncol.*, vol. 137, no. 3, pp. 430–435, Jun. 2015.

[15] T. Gonzales, C. J. Kuo, S.-W. Chi, E. Hsi, C.-C. Chen, K.-T. Lee, and S.-H. Chiou, “Onco-proteogenomics identifies urinary S100A9 and GRN as potential combinatorial biomarkers for early diagnosis of hepatocellular carcinoma,” *BBA Clin.*, vol. 3, pp. 205–213, Jun. 2015.

[16] Y. Wang, W. Du, Y. Liang, X. Chen, C. Zhang, W. Pang, and Y. Xu, “PUBPro: A computational pipeline for prediction of urine excretory proteins,” in *Proc. Int. Conf. Adv. Data Mining Appl. Cham*, Switzerland: Springer, 2016, pp. 714–725.

[17] D. Wang, S. Zeng, C. Xu, W. Qiu, Y. Liang, T. Joshi, and D. Xu, “MusiteDeep: A deep-learning framework for general and kinase-specific phosphorylation site prediction,” *Bioinformatics*, vol. 33, no. 24, pp. 3909–3916, Dec. 2017.

[18] H. Liang, X. Sun, Y. Sun, and Y. Gao, “Text feature extraction based on deep learning: A review,” *EURASIP J. Wireless Commun. Netw.*, vol. 2017, no. 1, pp. 1–12, Dec. 2017.

[19] L. Wei, Y. Ding, R. Su, J. Tang, and Q. Zou, “Prediction of human protein subcellular localization using deep learning,” *J. Parallel Distrib. Comput.*, vol. 117, pp. 212–217, Jul. 2018.

[20] Z. Zhang, Y. Zhao, X. Liao, W. Shi, K. Li, Q. Zou, and S. Peng, “Deep learning in omics: A survey and guideline,” *Briefings Funct. Genomics*, vol. 16, no. 1, pp. 41–57, Feb. 2019.

[21] Q. Zou, P. Xing, L. Wei, and B. Liu, “Gene2vec: Gene subsequence embedding for prediction of mammalian N6-methyladenosine sites from mRNA,” *RNA*, vol. 25, no. 2, pp. 205–218, Feb. 2019.

[22] S. F. Altschul, T. L. Madden, A. A. Schäffer, J. Zhang, Z. Zhang, W. Miller, and D. J. Lipman, “Gapped BLAST and PSI-BLAST: A new generation of protein database search programs,” *Nucleic Acids Res.*, vol. 25, no. 17, pp. 3389–3402, Sep. 1997.

[23] S.-J. Li, M. Peng, H. Li, B.-S. Liu, C. Wang, J.-R. Wu, Y.-X. Li, and R. Zeng, “Sys-BodyFluid: A systematical database for human body fluid proteome research,” *Nucleic Acids Res.*, vol. 37, pp. D907–D912, Jan. 2009.

[24] P. Legrain et al., “The human proteome project: Current state and future direction,” *Mol. Cellular Proteomics*, vol. 10, no. 7, Jul. 2011, Art. no. M111.009993.
GAOYANG LI received the Ph.D. degree in bioinformatics from Jilin University, Changchun, China, in 2018. From 2015 to 2017, he was a Visiting Scholar with the University of Georgia, USA. He is currently a Postdoctoral Fellow with Tongji University. His major research interests include bioinformatics and computational biology.

HUANSHENG CAO received the Ph.D. degree in cell and molecular biology from Fordham University, USA, in 2012. From 2014 to 2017, he was a Postdoctoral Fellow with the University of Georgia, USA. He is currently an Assistant Research Professor with the Biodesign Institute, Arizona State University, USA. His major research interests include bioinformatics and system biology.

YING LI received the Ph.D. degree in computational mathematics from Jilin University, Changchun, China, in 2004. From 2005 to 2007, she was a Postdoctoral Fellow with Tsinghua University. From 2011 to 2012, she was a Visiting Scholar with the University of Georgia, USA. She is currently an Associate Professor with the College of Computer Science and Technology. She has published more than 30 journal articles and conference papers. Her current research interests include machine learning and bioinformatics.

YANCHUN LIANG received the Ph.D. degree in applied mathematics from Jilin University, Changchun, China, in 1997. From 1990 to 1991, he was a Visiting Scholar with The University of Manchester, U.K. From 2000 to 2001, he was a Visiting Professor with the National University of Singapore. From 2002 to 2004, he was a Guest Professor with the Institute of High Performance Computing of Singapore. From 2006 to 2008, he was a Visiting Professor with Trento University, Italy. From 2011 to 2019, he was a Visiting Professor with the University of Missouri. He is currently a Professor with the College of Computer Science and Technology, Jilin University. He is also with the School of Computer Science, Zhuhai College of Jilin University, China. He has published more than 400 articles. His research was featured in the IEEE Transactions on Knowledge and Data Engineering, the IEEE Transactions on Systems, Man, and Cybernetics Part A: Systems and Humans, Briefings in Bioinformatics, Bioinformatics, the Journal of Micromechanics and Microengineering, Physical Review E, Applied Artificial Intelligence, and so on. His research interests include computational intelligence, machine learning methods, text mining, and bioinformatics. He was a recipient of several grants from NSFC, EU, and so on.