Digital Object Identifier 10.1109/ACCESS.2021.3060612

EDITORIAL

IEEE ACCESS SPECIAL SECTION EDITORIAL: FEATURE REPRESENTATION AND LEARNING METHODS WITH APPLICATIONS IN LARGE-SCALE BIOLOGICAL SEQUENCE ANALYSIS

Machine learning has been widely applied in the fields of biomedicine, computational biology, bioinformatics, image processing, and so on. The performance of machine learning methods mainly relies on feature representation that is the mapping from various types of raw data (i.e., image and genomic data) to a discriminant high-dimensional data space, bridging the raw data with the input of learning/inference algorithms. A good representation is often one that captures the discriminative information from the data and supports effective machine learning. However, over the last few decades, most representation learning approaches are labor-intensive and heavily dependent on the professional knowledge of researchers (dependent on handcrafted feature engineering). To conduct more novel applications in bioinformatics and biomedicine, the weakness of current learning algorithms should be overcome by developing novel feature representation learning algorithms, including supervised representation learning algorithms that are learning features from labeled data, unsupervised feature representation strategies that are learning feature representatives from unlabeled data, and deep feature representation learning algorithms that are learning representative features from data using deep learning architectures.

This Special Section of IEEE Access on feature representation and learning methods with applications in bioinformatics and biomedicine aims at bringing together researchers to disseminate their new feature representation and learning algorithms in biomedical and bioinformatics applications while expanding the scope and ease of applicability of machine learning and making progress toward artificial intelligence (AI).

The Call for Papers aroused great enthusiasm in the scientific community and received 122 submissions. Out of these, 67 articles were accepted for inclusion in the Special Section after a thorough review process by at least two independent referees. A brief description of each accepted article follows.

The article "Identifying essential signature genes and expression rules associated with distinctive development stages of early embryonic cells," by Chen *et al.*, provides a group of effective gene signatures and classification rules for embryo cell subtyping, based on a feature list produced by analyzing the single-cell expression profiles of embryo cells using the Monte Carlo feature selection (MCFS) method. More specifically, a group of key gene characters is extracted from the feature list using the incremental feature selection (IFS) method, incorporating support vector machine (SVM); a group of classification rules are produced by applying the Repeated Incremental Pruning to Produce Error Reduction (RIPPER) method on the feature list. The study demonstrates that these extracted features and produced rules can clearly uncover different expression patterns on cells in different stages.

The article "Gene expression difference between primary and metastatic renal cell carcinoma using patient-derived xenografts," by Jiang *et al.*, justifies the applications of patient-derived xenografts (PDX) in metastatic tumor studies by analyzing the single-cell gene expression profiles of cells from PDX of metastatic renal cell carcinoma (mRCC), from PDX of primary RCC (pRCC), and from parental mRCC (pt-mRCC), including the comparison of the gene expression patterns of PDX-mRCC and PDX-pRCC, and the test on pt-mRCC whether they can be correctly classified into the PDX-mRCC class rather than PDX-pRCC. The results show that generated PDX-mRCC features during the comparison can reflect the difference between metastatic and primary tumors, and pt-mRCC are very similar with PDX-mRCC, which can prove that the PDX is a useful model for metastatic tumors since it preserves the essence for tumor metastasis.

The article "VFM: Identification of bacteriophages from metagenomic bins and contigs based on features related to gene and genome composition," by Liu *et al.*, presents a new tool named VFM (Virus Finding & Mining) for identifying bacteriophages from metagenomes. Eighteen new features that base on the information of the codon usage bias, the proportion of hits of clusters of orthologous groups of proteins (COG), and 1-mer and 2-mer frequency are introduced to improve the performance of the classification tool.

The article "Recognizing novel tumor suppressor genes using a network machine learning strategy," by Zhao *et al.*, proposes a novel computational method to identify latent tumor suppressor genes (TSGs), based on a learning scheme that can extract essential properties of validated TSGs. Features that are derived from protein–protein interaction networks via a network embedding method are used for representing the data (validated TSGs together with other genes). A number of random forest (RF) models are built to learn essential differences between validated TSGs and other genes.

The article ''iEsGene-ZCPseKNC: Identify essential genes based on Z curve pseudo K-tuple nucleotide composition,'' by Chen *et al.*, develops a computational predictor for predicting essential genes in archaea, namely iEsGene-ZCPseKNC, based on a novel feature called Z curve pseudo *k*-tuple nucleotide composition (ZCPseKNC). Because the advantages of both Z curve and pseudo *k*-tuple nucleotide composition are incorporated in the proposed feature, it can capture sequence patterns of the essential genes, which is proven to be able to improve the performance of the predictor.

The article ''iRBP-Motif-PSSM: Identification of RNA-binding proteins based on collaborative learning,'' by Gao *et al.*, proposes a novel RNA-binding proteins (RBPs) computational predictor, namely iRBP-Motif-PSSM, based on collaborative learning. The motif information (Motif) and the evolutionary information extracted from the Position Specific Scoring Matrices (PSSM) are used as feature representations for identifying the RNA-binding proteins.

The article ''Improved prediction of cell-penetrating peptides via effective orchestrating amino acid composition feature representation,'' by Fu *et al.*, presents a framework for the prediction of cell-penetrating peptides (CPPs) using machine learning algorithm based on a novel feature extraction approach. Four feature vectors, including group amino acid composition (GAAC), composition of k-spaced amino acid group pairs (CKSAAGP), grouped di-peptide composition (GDPC), and the composition-transition-distribution (CTDC), are extracted by various extraction methods to generate high-quality feature representations for amino acid sequences. The feature extraction algorithm is proven to be rational for the CPPs predictor by the jackknife test.

The article ''A framework integrating heterogeneous databases for the completion of gene networks,'' by Luo *et al.*, proposes a network-based framework, termed NIHO, to optimize the uncompleted state of gene networks (GNs). In the framework, high-level features are learned from several genome networks by an end-to-end scheme that contains neural network and matrix completion and then are used for predicting the interactions among genes.

The article ''iPhoPred: A predictor for identifying phosphorylation sites in human protein,'' by Li *et al.*, proposes a novel sequence-based computational approach to identify phosphorylation sites in human proteomics by a series of statistical feature analyses. An optimal feature subset is obtained using the analysis of variance on the conservation scores and position-associated attributes that reflect the correlation of physicochemical characteristics of amino acid residues in order to achieve a high accuracy for the predictor. An online tool called iPhoPred is available.

The article ''Multi-source medical image fusion based on Wasserstein generative adversarial networks,'' by Yang *et al.*, presents an end-to-end model for the fusion of magnetic resonance imaging (MRI) and positron emission tomography (PET) images based on Wasserstein generative adversarial networks (MWGANs). The details of soft tissue structures in organs from MRI images, and the functional and metabolic information from PET images are extracted as features for generating the fused image by establishing two adversarial games between a generator and two discriminators.

The article ''White blood cell segmentation via sparsity and geometry constraints,'' by Zhong *et al.*, studies a novel segmentation algorithm for white blood cells (WBCs) based on both sparsity and geometry constraints. Specifically, a sparse image representation is constructed via combining the HSL color space and the RGB color channels, and the useful information from the nuclei features is preserved using a sparsity constraint, both of which are the process of feature representation. The geometry constraint, a model fitting strategy, is introduced for WBCs segmentation.

The article ''RDense: A protein-RNA binding prediction model based on bidirectional recurrent neural network and densely connected convolutional networks,'' by Li *et al.*, constructs a novel deep neural network model, namely RDense, to predict protein-RNA binding. A group of vectors that contains the RNA sequence, RNA secondary structure, and pairwise probability features extracted from the secondary structure as feature representations are input into the prediction model that uses the bidirectional long and short memory neural network (Bi-LSTM) along with densely connected convolutional neural networks (DenseNet). It is proven that because both the sequence and structure information are used for representing data, the deep neural network model can learn useful information for predicting protein-RNA binding.

In the article ''A novel model for predicting essential proteins based on heterogeneous protein-domain network,'' by Chen *et al.*, a novel prediction model for inferring potential essential proteins called NPRI is proposed based on a new heterogeneous Protein-Domain network. In the prediction model, the functional features and topological characteristics of each protein are extracted from the heterogeneous protein-domain network that is constructed by three kinds of networks including the weighted PPI networks, domain–domain network, and the initial protein-domain network. Then, these features of each protein are used for constructing a new distribution rate network to infer essential proteins.

The article ''Screening Dys-methylation genes and rules for cancer diagnosis by using the pan-cancer study,'' by Zhang *et al.*, studies the relationship between Dys-methylation genes and the diagnosis of cancer based on the methylation profiles. First, the methylation profiles of samples of 33 cancers as original features are used for producing relevant features. Then, the final features that may identify major tumor types are yielded by different feature ranking methods using the relevant features.

The article "Parallel MOEA based on consensus and membrane structure for inferring phylogenetic reconstruction," by Zhang *et al.*, proposes a parallelized multiobjective evolutionary algorithm (MOEA) for inferring phylogenies by adding consensus for improving the quality of convergence and using membrane structure for model fitting. Biomolecular sequences (gene sequences or protein sequences) are used for reconstructing the phylogenetic tree (reflecting evolutionary relationship) using an evolutionary reconstruction algorithm. The distance matrix constructed from the difference of the sequence is used for reconstructing the evolution tree, which can be seen as the representation of sequences.

The article "Predicting sub-Golgi apparatus resident protein with primary sequence hybrid features," by Wang *et al.*, presents a computational model for identifying sub-Golgi protein types using a machine learning method. PseKNC, K-separated Bigrams, and PsePSSM are extracted and used for selecting the optimal features by feature classifiers such as AdaBoost classifier. The final features are input into the support vector machine (SVM) model for predicting the sub-Golgi protein types, i.e., cis-Golgi proteins or trans-Golgi proteins.

The article "Escherichia Coli DNA N-4-methycytosine site prediction accuracy improved by light gradient boosting machine feature selection technology," by Lv *et al.*, proposes a novel machine learning method-based predictor for identifying DNA N-4-methycytosine (4mC) sites, namely iEC4mC-SVM. Light gradient boosting machine is used for feature selection. It is proven that this feature selection method can improve the performance of the iEC4mC-SVM predictor.

The article "Identification of protein lysine crotonylation sites by a deep learning framework with convolutional neural networks," by Zhao *et al.*, studies the computational prediction methods for identifying protein lysine crotonylation sites. The predictor based on convolutional neural network (CNN) and word embedding approach, called as pKcr, performs better than other predictors that use different features and algorithms. The digital vectors are obtained to represent the input data using the word embedding algorithm on biological characteristics from peptides.

The article "Craniomaxillofacial deformity correction via sparse representation in coherent space," by Li *et al.*, develops an upgraded prediction method based on the sparse representation (SR) technique for the reference jaw shape to guide the orthognathic surgery for patients with Craniomaxillofacial (CMF) deformities in order to alleviate the problem of the hypothesis in current SR-based prediction methods about the relevance between a patient's midface shape and his/her jaw shape. A coherent space learning and a multilayer mapping and refinement scheme are introduced to learn the representation of input data (the patient computed tomography data, namely CT data).

The article "SMOPredT4SE: An effective prediction of bacterial type IV secreted effectors using SVM training with SMO," by Yan *et al.*, establishes a predictor called SMOPredT4SE to identify type IV secretion system (T4SS) secreted effectors (T4SEs) from protein sequences with the aim of helping understand the pathogenic mechanism of T4SS. Combination features including series correlation pseudo amino acid composition and position-specific scoring matrix (PSSM) are used for representing protein sequences, and then applied to train the prediction model using support vector machine (SVM) and sequential minimal optimization (SMO) algorithms.

In the article "Mining key regulators of cell reprogramming and prediction research based on deep learning neural networks," by Ta *et al.*, the authors study the specific gene clusters for cell reprogramming, and identify stage-specific gene clusters (three stages: reprogramming initiation, maturation, and stabilization phase) using differential expression analysis. The DNA-binding profiles are integrated using the transcription factors Oc4 and three histone modifications (HMs) to represent the ChIP-seq data, and then to construct a quantitative model on different genome regions using deep neural networks.

The article "K-means multi-verse optimizer (KMVO) algorithm to construct DNA storage codes," by Cao *et al.*, proposes an improved K-means multi-verse optimizer (KMVO) algorithm to construct the DNA codes that is the first step of constructing a DNA storage system. The related constraints, including Hamming distance constraints, GC-content constraints, and no-runlength constraints, are used for constructing the code boundary using the KMVO algorithm, the results of which show that it can increase storage utilization.

The article "Classification of cancers based on a comprehensive pathway activity inferred by genes and their interactions," by Xu *et al.*, proposes a novel method for cancer classification based on describing pathway activity that incorporates both the genes' activity and their interactions. A weighted activity based on the differential expression degree and their correlation with the phenotype, and the activity of a gene pair are computed for representing cancer gene expression microarray data, and then are applied for calculating the activity score of a pathway.

The article "A hybrid CNN-LSTM model for forecasting particulate matter (PM2.5)," by Li *et al.*, develops a deep learning model for predicting the surface PM2.5 concentration in the next 24 h based on CNN and LSTM. The features related to air quality are extracted by training CNN, and the features of input series data are reflected by LSTM.

The article "Prioritizing human microbe–disease associations utilizing a node-information-based link propagation method," by Peng *et al.*, proposes a novel approach, called link propagation human microbe–disease association prediction (LPHMDA), for discovering the underlying associations between microbes and diseases based on node information.

Adjacency matrixes are calculated to represent associations, while Gaussian interaction profile kernel similarities are computed as node information for microbes and diseases.

The article "Patch-driven tongue image segmentation using sparse representation," by Liu *et al.*, proposes an image segmentation method for tongue diagnosis based on patch driven with sparse representation. Patches in the spatially varying dictionary that are constructed by local patches of training images are used for representing each patch in the testing image sparsely, which derives sparse coefficients. The tongue probability map is estimated using the sparse coefficients and is then employed for segmentation.

The article "Global dissipativity for stochastic genetic regulatory networks with time-delays," by Wang *et al.*, investigates the global dissipativity and corresponding attractive set for Genetic Regulatory Networks (GRNs) with stochastic disturbances and time-delays. Stochastic disturbances are used for the processes of feedback regulation and translation to reflect the inherent noise. Some corresponding sufficient conditions and the attractive set of the GRNs based on linear matrix inequality form.

The article "TM-ZC: A deep learning-based predictor for the Z-coordinate of residues in α-helical transmembrane proteins," by Lu *et al.*, proposes a predictor for identifying the Z-coordinate of α-helical transmembrane proteins based on deep learning, namely TM-ZA. One-hot code and PSSM are calculated for representing the sequence data and are then employed as input for a CNN regression model.

The article "FunEffector-Pred: Identification of fungi effector by activate learning and genetic algorithm sampling of imbalanced data," by Wang *et al.*, designs a fungal effector predictor, called FunEffector-Pred, to address the problem that most fungal effectors lack consensus motifs and data imbalance. The informative features are extracted from an imbalanced data set using a method that combines a granular support vector-based under-sampling (GSV-US) strategy and a genetic algorithm such as Ile, Gly, Val, Leu, and Thr, as well as the combination of aromatic amino acids with positively charged amino acids.

The article "An improved method for identification of pre-miRNA in drosophila," by Yu *et al.*, develops an improved method for identifying microRNAs in Drosophila. Features are extracted using several methods such as iLearn, PyFeat, and Pse-in-One methods, and then Max-Relevance-Max-Distance (MRMD2.0) and t-Distributed Stochastic Neighbor Embedding (t-SNE) are applied to reduce the dimension of the features. Final features are fed into the random forest classifier to identify miRNAs.

The article "Extracting explicable rules for the identification of compound–protein interactions," by Zhu *et al.*, tries to develop a computational method for the identification of compound-protein interaction (CPI) based on explicable rules. The rules involving gene ontology, KEGG pathway, and molecular ACCess System fingerprint descriptor are obtained to describe the functional enrichment of CPIs, which can be seemed as representation features. Because of these extracted features, the proposed classifier can clarify the identification procedures.

The article "Using evolutionary information and multi-label linear discriminant analysis to predict the subcellular location of multi-site bacterial proteins via Chou's 5-steps rule," by Du *et al.*, proposes a method for identification of subcellular location of bacterial proteins based on the consideration of two problems, multi-sites of proteins, and high-dimensional features. For these two problems, two types of evolutionary features are extracted by absolute entropy correlation analysis (AECA-PSSM) and discrete wavelet transform (PSSM-DWT), and are then combined. Multi-label linear discriminant analysis (MLDA) is employed on the combined features for transforming the high-dimensional features into lower dimensional features. Multi-label k-nearest neighbor algorithm (ML-KNN) is used for predicting the subcellular location of bacterial proteins and is proven to be effective.

The article "ItLnc-BXE: A bagging-XGBoost-ensemble method with comprehensive sequence features for identification of plant lncRNAs," by Zhang *et al.*, presents an identification method called ItLnc-BXE for plant lncRNAs based on comprehensive features and the ensemble learning strategy. A set of sequence features containing four categories such as sequence-based features, ORF-based features, codon-based features, and alignment-based features, are collected and filtered to represent transcripts. The ItLnc-BXE model is constructed using ensemble learning on the filtered features to identify plant lncRNAs.

The article "Identifying G-protein coupled receptors using mixed-feature extraction methods and machine learning methods," by Ao *et al.*, proposes an identification method for G-protein-coupled receptors (GPCRs) based on mixed-feature vectors. Three individual features, such as 400D, N-gram, and parallel correlation pseudo amino acid composition (PC-PseAAC), are combined to represent the sequence data. Then, the max-relevance-max-distance (MRMD) and t-distributed stochastic neighbor embedding (t-SNE) are employed to reduce the dimensionality of mixed-feature.

The article "Mammographic classification based on XGBoost and DCNN with multi-features," by Song *et al.*, develops a new computer-aided diagnosis (CAD) method for the classification of benign and malignant masses based on multi-features. Multi-features include two groups of features: features extracted using deep convolution neural network (DCNN) and the image texture features that are the breast mass information in mammograms extracted from scoring features, gray-level co-occurrence matrix (GLCM) and histogram of oriented gradient (HOT). These multi-features are combined to input to the classifier. The classifier is trained by SVM and Extreme Gradient Boosting (XGBoost) to predict mammographic masses as one of three classes: normal, benign, and cancer (malignant) masses.

The article "Identification of type VI effector proteins using a novel ensemble classifier," by Wang *et al.*, proposes an ensemble predictor for the identification of type VI secretion system effectors, termed T6SEs. The protein sequence data is converted to numerical vector representation by encoding with k-separated-bigrams-PSSM. An integrated model constructed by combining six fine-tuned classifiers using a soft voting strategy.

The article "The frequencies of oppositely charged, uncharged polar, and β-branched amino acids determine proteins' thermostability," by Sun *et al.*, studies how to enhance the thermostability of thermophilic proteins, aiming at building a classifier to discriminate thermostability and non-thermostability proteins based on the amino acid frequencies in protein sequences. The amino acid composition (AAC) is extracted from the frequencies of each amino acid in the protein sequence, and then features are extracted from AAC information using several machine-learning-based feature selection algorithms such as SVM, random forest, and regularized logistic regression. The principal component analysis is used for removing noisy and redundant information from extracted features. Analysis results show that amino acids with oppositely charged, reducing the number of uncharged polar, and increasing β-branched aliphatic may increase thermostability.

The article "Integrated analysis of a noncoding RNA-associated competing endogenous RNA network in non-alcoholic fatty liver disease," by Hao *et al.*, studies the relationship between non-alcoholic fatty liver disease (NAFLD) and dysregulation of coding and long noncoding RNAs (lncRNAs). A competing endogenous RNA network is established by incorporating interactions between different RNAs and is used for illustrating the regulatory roles of noncoding miRNAs and lncRNAs.

The article "EnsemPseU: Identifying pseudouridine sites with an ensemble approach," by Bi *et al.*, proposes a computational method, named EnsemPseU, for identifying Pseudouridine Sites based on ensemble learning. Sequence features are extracted using several sequence representation strategies, such as kmer, binary encoding, enhanced nucleic acid composition (ENAC), nucleotide chemical property (NCP), and nucleotide density (ND). Then, to reduce the feature dimensionality and remove redundant information, chi-square feature selection is used. Finally, a prediction model is built using an ensemble algorithm including several machine learning algorithms such as support vector machine (SVM), extreme gradient boosting (XGBoost), naïve Bayes (NB), k-nearest neighbor (KNN), and random forest (RF).

In the article "The feature compression algorithms for identifying cytokines based on CNT features," by Li and Gao, three feature compression algorithms, called the genetic-based algorithm, the greedy-based algorithm, and the brute-force-based algorithm, are proposed for identifying cytokines from other proteins. Features are extracted and selected, and redundant information is removed from 188-dimensional CNT features using these three feature compression algorithms. It is demonstrated that the brute-force-based algorithm achieves the highest accuracy, while the genetic-based algorithm achieves the least number of compressed features, and the greedy-based algorithm consumes the least time.

The article "Identifying essential methylation patterns and genes associated with stroke," by Yu *et al.*, studies the epigenetic contribution and regulation on stroke pathogenesis using a novel computational method. The genome methylation data of stroke patients are screened out, and a group of functional methylated or demethylation genes are identified from the screened data. It is proven that the abnormal methylation status is found in the identified stroke genes.

The article "iRNA-m5C_NB: A novel predictor to identify RNA 5-methylcytosine sites based on the naive Bayes classifier," by Dou *et al.*, introduces a novel computational method for identifying RNA 5-Methylcytosine (m5C) sites based on naïve Bayes algorithm. Features are combined and selected using several feature extraction methods such as Bi-profile Bayes (BPB), enhanced Nucleic Acid Composition (ENAC), electron-ion interaction pseudopotentials (EIIP), and mMGap_1 technology from the samples obtained by hybrid-sampling strategy.

The article "Identification of secreted proteins from malaria protozoa with few features," by Li *et al.*, introduces a computational method for identifying the secreted proteins from malaria protozoa based on machine learning methods with a series of feature processing. Five types of features including amino acid composition, grouped amino acid composition, pseudo-amino acid composition, conjoint triad, and C/T/D are calculated from the protein sequences, and then feature dimension reduction is employed. Finally, the dimension reduced features are combined to represent the protein sequence, and are input into the classifier model based on SVM and random forest algorithm.

The article "The classification of enzymes by deep learning," by Tao *et al.*, summarizes a variety of computational methods for predicting the classification of enzymes using machine learning. Feature extraction methods such as composition, transition, and distribution (CTD), pseudo-amino acid composition (PseAAC), pseudo position-specific scoring matrix (PsePSSM), topological indices, torsion angles and function domain (FunD)-based method, and the classification model based on LDA (namely linear discriminant analysis), KNN, ANN (artificial neural networks), and SVM, which can be used for enzyme classification prediction, are summarized.

The article "A new TTZ feature extracting algorithm to decipher tobacco related mutation signature genes for the personalized lung adenocarcinoma treatment," by He *et al.*, proposes a novel feature extraction method for identification of tobacco-related mutation signature genes in order to assist lung adenocarcinoma treatment. The tobacco exposure pattern (TEP) model is built and optimized to identify signature genes and to uncover their interaction relationships at different molecular levels. Features used for the model are extracted both from mutation frequencies and sequencing information of insertions and deletions based on the z-curve method.

The article ''Prediction of cyclin protein using two-step feature selection technique,'' by Sun *et al.*, proposes a machine-learning-based method for identification of cyclin proteins from primary protein sequences. Features are extracted from amino acid sequences, and then, some feature descriptors are fed into the prediction model such as Autocorrelation, AAC (amino acid composition), and CTDC (composition, transition, and distribution), which remain after feature dimension reduction is employed.

The article ''Automated ischemic stroke subtyping based on machine learning approach,'' by Fang *et al.*, proposes a machine learning-based computational method for classifying ischemic stroke subtypes. First, the importance of features are ranked and the Person correlations between features are analyzed by the Shapiro–Wilk algorithm. Then, the important features for ischemic stroke subtyping are selected using a set of combined classifiers such as linear SVC, Random Forest Classifier, Extra Trees Classifier, AdaBoost Classifier, and Multinomial Naïve Bayes Classifier.

The article ''Identification of SNARE proteins through a novel hybrid model,'' by Li, proposes a computational method for predicting SNRE proteins using a novel hybrid prediction model. The feature extraction method is combined with the oversampling filter and random forest classification method.

The article ''Predicting the influence of MicroRNAs on drug therapeutic effects by random walking,'' by Xu *et al.*, constructs a heterogeneous network for predicting the potential miRNA-drug effect associations. The data of miRNAs, drugs, and miRNA-drug effect associations are obtained to construct the miRNA similarity network, drug similarity network, and miRNA-drug effect association network, respectively, and these three similarity networks are used for constructing a heterogeneous network that can be employed to predict the potential miRNA-drug effect associations using the Bi-Random walk (BiRW) algorithm.

The article ''Efficient ResNet model to predict protein–protein interactions with GPU computing,'' by Lu *et al.*, presents a computational model called ResPPI for protein-protein interaction (PPI) prediction based on the residual network (ResNet) with GPU computing. The protein sequences are represented by an embedding method, and features are extracted by ResNet with deep layers using GPU computing.

The article ''A KNN model based on Manhattan distance to identify the SNARE proteins,'' by Gao and Li, tries different feature extraction approaches and classifiers to find a suitable model for the identification of SNARE proteins. Several feature extraction methods such as 188D method, K-skip-2-gram method, and CKSAAP method are employed, along with different sampling strategies (resampling, SMOTE, and no sampling) and distance measurements (Chebyshev distance, Euclidean distance, Manhattan distance, and Minkowski distance) to combine and construct the classification model. The combined model based on Manhattan distance with CKSAAP feature extraction method and no sampling is proven to be the most suitable for identifying the SNARE proteins.

The article ''LAK: Lasso and K-means based single-cell RNA-Seq data clustering analysis,'' by Hua *et al.*, proposes a novel method named LAK for single-cell RNA-seq data clustering analysis using Lasso and K-means. The input single-cell RNA-seq data expression matrix is transformed to a linear model by preprocessing, including pre-filtering and normalizing. Then, feature selection clustering is achieved by adding Lasso penalty because of the high sparsity of the preprocessed data.

In the article ''Interpretation of electrocardiogram (ECG) rhythm by combined CNN and BiLSTM,'' by Xu *et al.*, the authors propose a deep learning-based method for predicting the classification of ECG heart signals for diagnostic purposes. CNN and bidirectional LSTM (BiLSTM) are employed to construct the classifier, specially including convolutional layers, residual blocks, BiLSTM layers, and fully connected layers. The feature map of the model is recalibrated in the Squeeze and Excitation Network (SENet) of each residual block.

The article ''Similarity-based machine learning model for predicting the metabolic pathways of compounds,'' by Jia *et al.*, presents a new machine learning model for predicting actual metabolic pathways for given compounds based on the similarity. A pair of compounds and metabolic pathways can be seen as a sample, and are represented by seven features (compound associations that measure compound linkages from different aspects) based on similarity. Random forest is employed as a classification algorithm.

The article ''A genotype-based ensemble classifier system for non-small-cell lung cancer,'' by Ren *et al.*, presents an ensemble classifier system for Non-Small-Cell Lung Cancer (NSCLC) based on genotype, aims to identify the genetic subtypes of tumors, and explains the pathologies of NSCLC. Feature selection is implemented by Spearman's correlation statistical methods, and then, the selected genes are input to the ensemble classifier that uses decision tree, SVM, and logistic regression.

The article ''A Pearson based feature compressing model for SNARE protein classification,'' by Li, proposes a feature compressing model for classifying the SNARE proteins based on Pearson. Features are extracted from the protein sequences (SNARE and non-SNARE proteins) using 188D, CKSAAP, CTDD, and CTRIAD feature extraction methods, and then feature filtering is employed. The final features are evaluated using Chi-Square, Information Gain, and Pearson Correlation Coefficient feature selection methods, and then are used for training the random forest classifier.

The article ''Disease cluster detection and functional characterization,'' by Guo *et al.*, studies the detection and functional characterization of disease clusters based on the interaction between genes or function and diseases. The disease data is classified into 15 relatively separated disease clusters using disease clustering based on the separate score between diseases. The separated disease clusters are learned based on information such as disease-associated genes, their GO terms, and KEGG pathways annotations.

The article "Prediction of therapeutic peptides using machine-learning: Computational models, data sets, and feature encodings," by Attique *et al.*, summarizes the feature encoding methods and machine-learning-based classifiers for identifying therapeutic peptides. Feature encoding approaches for protein sequences such as structure-based (chaos game representation, secondary structure-based encoding), sequence-based (k-spaced amino acid pairs, pseudo-amino acid composition, amino acid composition, sequence alignment, discrete Fourier transformation, and protein relatedness measure), and property-based (motif-based features, composition-transition-distribution, and physico-chemical proper) methods are summarized.

The article "A cytokine protein identification model based on the compressed PseKRAAC features," by Gao and Li, proposes a computational model called MRMD-cosine for identifying the cytokine protein based on the compressed PseKRAAC features. Four kinds of feature sets named type1 g-gap, type1 lambda, type2 g-gap, and type2 lambda feature set are extracted from the cytokine proteins by PseKRAAC feature extraction method, and then redundant features are removed using MRMD (Max Relevant Max Distance) algorithm with Euclidean distance, Cosine similarity, and Tanimoto coefficient. Finally, bagging and random forest algorithms are used for constructing the classification model.

The article "Survival marker miRNA-mediated subpathways of breast invasive carcinoma derived from activity profile," by Ning *et al.*, infers miRNA-mediate subpathway activity profiles using integrated information based on the global directed pathway network (GDPN). In the network, genes are as nodes and gene expression profiles, prior gene interaction information and target relations between miRNAs and genes, and topological information are integrated into miRNA-mediated subway activity profile, which proves that the noises from sequencing errors and samples heterogeneity are reduced by integration.

The article "ACP-GCN: The identification of anticancer peptides based on graph convolution networks," by Rao *et al.*, proposes a novel computational anticancer peptide (ACP) prediction model called ACP-GCN based on graph convolution networks. In contrast to the most computational prediction methods that use hand-crafted features, the ACP-GCN model employs the graph convolution network to automatically extract features from peptide data and predict ACPs.

The article "Adaptive unsupervised feature learning for gene signature identification in non-small-cell lung cancer," by Ye *et al.*, proposes a novel method for identifying gene signatures in non-small-cell lung cancer (NSCLC) subtype classification based on adaptive unsupervised feature learning. The informative genes are selected by a joint learning framework for unsupervised feature selection that incorporates linear discriminant analysis, adaptive structure preservation and $l_{2,1}$-norm sparse regression after module-based gene filtering.

The article "GANs-based data augmentation for citrus disease severity detection using deep learning," by Zeng *et al.*, conducts a Deep Convolutional Generative Adversarial Networks (DCGANs) for predicting citrus disease severity using citrus leaf images infected by HLB (Huanglongbing). The data set of citrus leaf images infected by HLB for the detection model was constructed from the HLB-infected leaf images obtained from *PlanVillage* and *crowdAI*. In order to be used for training the DCGANs model, the constructed data set is up to two times itself. Feature extraction of image data is automatically carried out in the DCGANs.

The article "Supply–demand matching in non-cooperative social networks," by Zhang *et al.*, develops a general framework for solving the complex supply–demand matching problems in non-cooperative social networks. A supply network covering task demand is constructed based on a distributed negotiation mechanism, and the social relations of supply network are quantified to assess the cost of communication between supply–need nodes. The process of supply network construction is completed by supply network construction algorithm, preference algorithm, and coordination algorithm.

In the article "Decentralized telemedicine framework for a smart healthcare ecosystem," by Abugabah *et al.*, the authors propose a decentralized telemedicine blockchain-based framework using Ethereum smart contracts. The smart contract contains details such as patient name, patient ID, IPFShashEHR, Ethereum address of the medresearch_org, number of Requests by Telmedcenter, number of approvals by MedInsurer, and contState. It is demonstrated that the secure transfer of sensitive medical data is ensured because the smart contracts govern all of the transactions.

The article "N-GlycoGo: Predicting protein N-glycosylation sites on imbalanced data sets using heterogeneous and comprehensive strategy," by Chien *et al.*, constructs a prediction model called N-GlycoGo for identifying protein N-linked glycosylation sites based on XGBoost using imbalanced Data sets. Features of protein sequences and amino acids in sequences that contain sequence-based, structure-based, and function-based feature information are extracted to encode 11 heterogeneous features. The encoded features are used for training the XGBoost prediction model to find the protein N-linked glycosylation sites.

Finally, the invited article "MF-EFP: Predicting multi-functional enzymes function using improved hybrid multi-label classifier," by Xiao *et al.*, deals with the prediction of multi-functional enzymes function. The authors designed a hybrid model, called MF-EFP, by integrating the feature and neighbor label information. A hybrid model combining SAAC (split amino acid composition) discrete model that can avoid completely losing the sequence order information and PseAAC (pseudo amino acid composition) is constructed to represent enzyme sequences. Several functional enzyme data are input to the constructed feature extraction model, then extracted features are used for training the novel multi-label classifier hML-KNN to predict the function of input enzymes.

In conclusion, we would like to thank all the authors who submitted their research articles to our Special Section.

We highly appreciate the contributions of the reviewers for their constructive comments and suggestions. We also would like to acknowledge the guidance from IEEE ACCESS Editor-in-Chief and staff members.

FEIFEI CUI, *Lead Editor*
Institute of Fundamental and Frontier Sciences
University of Electronic Science and Technology of China
Chengdu 610054, China

QUAN ZOU, *Guest Editor*
Institute of Fundamental and Frontier Sciences
University of Electronic Science and Technology of China
Chengdu 610054, China

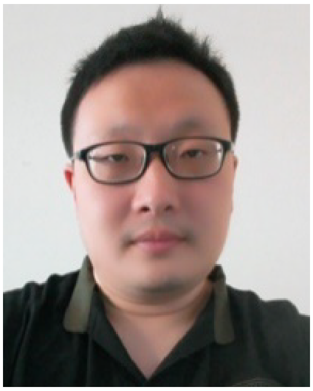
QIN MA, *Guest Editor*
Bioinformatics and Mathematical
Biosciences Laboratory (BMBL)
South Dakota State University
Brookings, SD 57007, USA

LEYI WEI, *Guest Editor*
School of Software
Shandong University
Jinan 250100, China

JIJUN TANG, *Guest Editor*
Department of Computer Science and Engineering
University of South Carolina
Columbia, SC 29208, USA

DARIUSZ MROZEK, *Guest Editor*
Department of Applied Informatics
Silesian University of Technology
44-100 Gliwice, Poland

FEIFEI CUI received the M.S. degree in computer application technology from Shandong University, Jinan, China, in 2012, and the Ph.D. degree in bioinformatics from The University of Tokyo, Japan, in 2020. She is currently working as a Postdoctoral Researcher with the University of Electronic Science and Technology of China. Her research interests include bioinformatics, deep learning, and biological data mining. She was/has been a winner of the Japanese Government (Monbukagakusho: MEXT) Scholarship Program from 2016 to 2019, and the Postdoctoral International Exchange Program since 2020 until 2022.

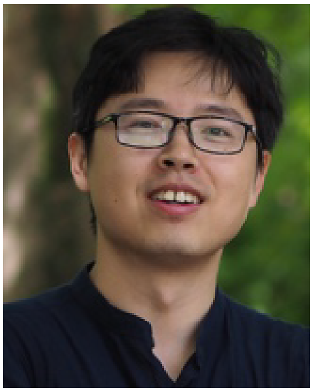


QUAN ZOU (Senior Member, IEEE) received the B.Sc., M.Sc., and Ph.D. degrees in computer science from the Harbin Institute of Technology, China, in 2004, 2007, and 2009, respectively.

He worked with Xiamen University and Tianjin University from 2009 to 2018 as an Assistant Professor, an Associate Professor, and a Professor. He is currently a Professor with the Institute of Fundamental and Frontier Sciences, University of Electronic Science and Technology of China, Chengdu, China. Several related works have been published by *Science*, *Briefings in Bioinformatics*, *Bioinformatics*, and IEEE/ACM TRANSACTIONS ON COMPUTATIONAL BIOLOGY AND BIOINFORMATICS. Google scholar showed that his more than 100 articles have been cited more than 5000 times. He was selected as one of the Clarivate Analytics Highly Cited Researchers in 2018 and 2019. He is a reviewer for many impacted journals and the National Natural Science Foundation of China (NSFC). His research interests include bioinformatics, machine learning, and parallel computing. He is also the Editor-in-Chief of *Current Bioinformatics*, an Associate Editor of IEEE ACCESS, and an Editorial Board Member of *Computers in Biology and Medicine*, *Genes*, and *Scientific Reports*.

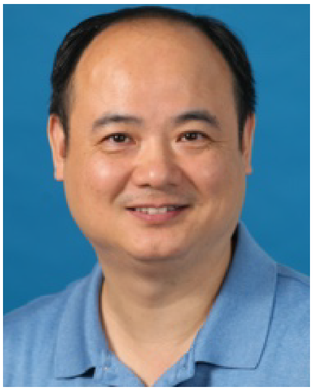
QIN MA received the B.S. degree in applied mathematics and the Ph.D. degree in operational research from Shandong University, China.

He worked in bioinformatics with the University of Georgia for six years. He is currently an Assistant Professor and leads the Bioinformatics and Mathematical Biosciences Laboratory (BMBL), South Dakota State University. He has published in over 30 peer-reviewed journals, including *Nucleic Acids Research*, *Bioinformatics*, *Scientific Reports-Nature*, *BioEnergy Research*, and *BMC Genomics*. His strong and long-term interests include: 1) elucidation of metabolic networks and associated regulatory systems, 2) characterization of functionally genomic organization in bacteria and plants, and 3) development of enabling computational techniques in support of the above scientific studies. He is also a member of the International Society for Computational Biology (ISCB), the Honor Society of Agriculture (SDSU), and the Bioenergy Research Center (BESC), Department of Energy (DOE). He is also an Associate Editor of *BMC Genomics* and an Editor of *Mathematical Biosciences*.



LEYI WEI received the B.Sc. degree in computing mathematics and the M.E. and Ph.D. degrees in computer science and technology from Xiamen University, in 2010, 2013, and 2016, respectively.

He worked as an Assistant Professor with the College of Intelligence and Computing, Tianjin University. He also worked as a Project Researcher with the Laboratory of Functional Analysis in silico, Institute of Medical Science, The University of Tokyo, from 2018 to 2019. He is currently a Professor with the School of Software, Shandong University, Jinan, China. He has published more than 60 peer-reviewed articles (more than 30 articles as first or corresponding author) in top-tier journals, such as *Bioinformatics*, *Briefings in Bioinformatics*, and *BMC Genomics*. He has around 800 citations in Google Scholar, and his H-index is 26. His research interests include bioinformatics, machine learning, deep learning, and computational biology. He has served as an Associate Editor for *Frontiers in Genetics*, the Section Editor for *Current Bioinformatics*, and an Associate Editorial Board Member for *Combinatorial Chemistry & High Throughput Screening*. He is also a Guest Editor of three high-impact journals, *Current Bioinformatics*, *Current Protein & Peptide Science*, and *Computational and Structural Biotechnology Journal*.



JIJUN TANG received the Ph.D. degree from the Department of Computer Science, The University of New Mexico, in 2004. He is currently a Professor and a Graduate Director of the Department of Computer Science and Engineering, University of South Carolina. His research interests include computational biology, bioinformatics, and serious games. He has published more than 120 articles, and his research has been supported by NIH, NSF, NEH, and ONR.

DARIUSZ MROZEK (Senior Member, IEEE) received the Ph.D. degree from the Silesian University of Technology (SUT), Gliwice, Poland, in 2006.

He is currently an Associate Professor and the Head of the Department of Applied Informatics, SUT. He is the author of more than 100 papers published in conference proceedings and international journals, the author of two books on the use of Big Data analytics and high-performance computations in protein bioinformatics published by Springer, the co-editor of 15 other books devoted to databases and data processing, and an editor of many special issues in reputable scientific journals. His research interests include the IoT, parallel and cloud computing, databases, big data, and bioinformatics. He is also focused on developing IoT solutions for healthcare on the use of novel computation techniques to get insights from biological data, including NGS and proteomics data. He has collaborated with qualified institutions by working in different research projects, such as the Imperial College of London or Microsoft Research, USA. He is also a member of the IEEE Engineering in Medicine and Biology Society (EMBS), the IEEE Systems, Man, and Cybernetics Society (SMCS), and IEEE Cloud Computing Community.