IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received February 24, 2021, accepted March 23, 2021, date of publication March 31, 2021, date of current version April 13, 2021.

Digital Object Identifier 10.1109/ACCESS.2021.3070083

Prediction of Piwi-Interacting RNAs and Their Functions via Convolutional Neural Network

MUHAMMAD TAHIR[1], MAQSOOD HAYAT[1], SHAHZAD KHAN[1], AND KIL TO CHONG[2,3]

[1]Department of Computer Science, Abdul Wali Khan University Mardan, Mardan 23200, Pakistan
[2]Advanced Electronics and Information Research Center, Jeonbuk National University, Jeonju 54896, South Korea
[3]Department Electronics and Information Engineering, Jeonbuk National University, Jeonju 54896, South Korea

Corresponding authors: Maqsood Hayat (m.hayat@awkum.edu.pk) and Kil To Chong (kitchong@jbnu.ac.kr)

This work was supported in part by the National Research Foundation of Korea (NRF) Grant by the Korean Government through MSIT under Grant 2020R1A2C2005612, in part by the Brain Research Program of the National Research Foundation (NRF) by the Korean Government through MSIT under Grant NRF-2017M3C7A1044816, and in part by the Research Funds for newly appointed professors of Jeonbuk National University, South Korea, in 2020.

ABSTRACT In eukaryotic cells, Piwi-interacting RNAs (piRNAs) are the type of short chain non-coding RNA molecules, which interconnect with PIWI proteins. It performs various cellular and genetic functions such as gene-specific protein translation, expression regulation, maintenance, and formulation of germ cells. Seeing the prominent contribution of piRNA in eukaryotic organism cells, many attempts were made to identify it computationally, however, unsatisfactory results were obtained. So, it is requisite to extend the concept of a computational tool in such a way that accurately represents piRNA. In this regard, intelligent and high discriminative deep learning i.e., the convolutional neural network based sequential-computational model known as ''piRNA-CNN'' is carried out for the prediction of piRNA. RNA sequences are mathematically expressed using the natural language processing method namely: word2vec in order to get prominent, relevant, and high variated numerical descriptors. The proposed ''piRNA-CNN'' model yields an accuracy of **93.83%** for the first-layer in which the provided query RNA molecule is predicted as non-piRNA or piRNA. In case of the piRNA, the proposed model identified the query as mRNA deadenylation or without deadenylation in the second layer, and achieved **91.19%** of accuracy. The obtained outcomes authenticated that the piRNA-CNN model exposed substantial results matched to the current tools stated in the literature, so far. It is further expected that the suggested predictive tool will assist scientists and researchers to design improved computational tools.

INDEX TERMS RNA, ensemble learning, genetic algorithm, word2vec, CNN.

I. INTRODUCTION

In eukaryotic cells, Piwi-interacting RNAs (piRNAs) are the leading group of short chain non-coding RNA molecules with a length of 24–31 nucleotides long polymer [1]. Various genomic and cellular functions including transposon silencing, gene expression regulation, maintenance and formulation of germ cells, and specific protein translation are performed by piRNAs. Numerous attempts were carried out and finally revealed that piRNAs are involved in various kinds of cancer; so, the study and knowledge regarding such type of RNAs are very imperative in certain areas such as RNA biology and drug development [2]–[4]. In a sequel, Lee *et al.*, and Nishibu *et al.*, performed several experimental methods in order to categorize whether an RNA molecule is piRNAs or not [5], [6]. However, only relying on laboratory experimental methods for sequence analysis are inadequate, inefficient; expensive as well as insensitive in some situations. Viewing the importance of piRNAs, computational approaches are essential to make possible the analysis of piRNAs in a more precise and efficient way. In the real world, there are two types of piRNA are reported, one is carried out deadenylation to target mRNA while the other one is without deadenylation [7]. However, the experimental methods are failed to explicitly explain the difference between these two types. Researchers have only concentrated on classifying piRNAs and non-piRNAs and introduced various computational models. Zhang *et al.* employed a support vector machine (SVM) and k-mer approach for proposing an automated model known as piRNAs predictor [8]. Later on, Wang *et al.* utilized SVM and transposon interaction for discrimination of piRNAs [9]. Likewise, Luo *et al.*, used physicochemical properties of RNA [10]. In a sequel, Li *et al.*, adopted the notion of ensemble learning for the prediction of

The associate editor coordinating the review of this manuscript and approving it for publication was Hao Ji.

piRNAs [11]. Recently, Liu *et al.*, proposed a two-layer automatic model for the prediction of piRNAs and their functional types [1]. Similarly, li *et al.*, introduced a predictor known as ''piRNAPred'' to identify piRNA and their function types by support vector machine [12]. Khan *et al.*, suggested a deep neural network (DNN) based computational model called ''2L-piRNADNN'' utilizing physicochemical behavior of RNA and di-nucleotide auto covariance as feature extraction techniques [13]. Here, an effort was made to propose a computational intelligent model for discrimination of piRNA and its types, adopting contemporary machine learning and deep learning approaches. In the case of Machine learning, two distinct nature of RNA sequences formulation methods were applied to extract numerical features. The extracted feature spaces are then provided to individual learning algorithms. Further, the predictions of individual learners are merged using a bio-inspired evolutionary genetic algorithm in order to correctly identify the desired class. In Deep Learning, word2vec method based feature space is used in combination with the CNN model. The analysis of the developed model is carried out on two benchmark datasets to demonstrate the stability and generalization strength of the model. As shown in Table 2 and 3, the Deep learning approach obtained successful results than the Machine learning approaches and existing methods.

- Natural language processing method ''word2vec'' is used for expressing RNA sequences.
- Compared machine learning algorithms with deep learning algorithm.
- 7-fold is applied for assessment.
- Various performace metrics are used for examining the algorithms performance.
- High throughput intelligent computational predictor is developed for piRNAs.

II. MATERIALS AND METHODS

A. BENCHMARK DATASET

In a biological system, Chou's 5-step rules become a benchmark for introducing a sequence-based statistical predictor [14]. The first and main step is the selection or construction of a valid dataset according to the problem that represents the motif of the target class. Here, the Liu et al dataset is selected as a benchmark dataset S Liu *et al.*, [1]. It can be mathematically expressed as:

$$\begin{cases} S = S^{+} + S^{-} \\ S^{+} = S^{+}_{inst} + S^{+}_{non-inst} \end{cases} \tag{1}$$

where the negative subset consists of 1,418 non-piRNAs segments; the positive subset contains 1,418 piRNAs segments; the subset is composed of 709 samples of piRNAs segments, which have the function of instructing target mRNA deadenylation [7]; while the remaining 709 samples subset belongs to without such function.

In this study, efforts were made to analyze various machine learning and deep learning algorithms in order to develope an intelligent computational model for the identification of piwi-interaction RNA and their types. RNA sequences are also formulated with via discrete and natural language processing methods. Details are mentioned below:

B. MACHINE LEARNING APPROACH

Here, we used various feature extraction methods and classifiers to work as baselines for comparison with the proposed deep-learning approach.

1) **Feature Extraction Methods**

The second step of Chou's rules is how to mathematically express DNA/RNA instances with an operative numerical formulation, which can correctly return the correlation with the desired class to be predicted. However, machine learning algorithms are designed in such a way that merely uses the vector. In order to collect only numeric features in the form of a vector from biological sequences, the discrete feature extraction method pseudo amino acid composition (PseAAC) was used [15]–[19]. The PseAAC concept has been broadly and rapidly exploited in the area of proteomics. Later on, this concept was extended to RNA/DNA sequences in the form of pseudo K-tuple nucleotide composition (PseKNC) [20]–[28]. It is also used for genome analysis. Accordingly, the idea of PseKNC has been implemented for expressing RNA sequences using discrete methods di-nucleotide composition (DNC) and tri-nucleotide composition (TNC).

- **Di-nucleotide Composition (DNC)** is a feature-encoding scheme, which expresses an RNA sequence with the help of two consecutive nucleotides pair. The occurrence frequency of each pair, such as N1N2 represents the 1st pair, N2N3 denotes the 2nd pair, and so forth, is computed. Finally, $4 \times 4 = 16$D resultant features space is generated [22], [25], [29]. DNC can be mathematically formulated as below:

$$S = [f(AU)f(AG)f(AC)\ldots f(UU)]^{T} \tag{2}$$

$$S = [f_1^{di}, f_2^{di}, f_3^{di}, \ldots . f_{16}^{di}]^{T} \tag{3}$$

where $f_1^{di} = f(AU)$ is the frequency of AU, $f_2^{di} = f(AG)$ is the frequency of AG; and so forth, T is a transpose.

- **Tri-nucleotide Composition (TNC)** is another feature-encoding scheme, which represents RNA sequence with the help of three consecutive nucleotides pair. The frequency of each pair is calculated. For example, in RNA sequence, the first pair is N1 N2 N3, the second pair is N2 N3 N4, and so forth, consequently, $4 \times 4 \times 4 = 64$D corresponding features vector is produced [20], [25]. The TNC can be numerically expressed as:

$$S = [f(AAA)f(AAU), \ldots ., f(UUU)]^{T} \tag{4}$$

$$S = [f_1^{3-tuple}, f_2^{3-tuple}, \ldots . f_{64}^{3-tuple}]^{T} \tag{5}$$

where $f_1^{3-tuple} = f(AAA)$ is the frequency of AAA, $f_2^{3-tuple} = f(AAC)$ is the *AAC* in RNA sequences; and so forth.

2) **Classification Algorithms** The next step of Chou's rules is what type of classification hypotheses implement in order to execute the training and predicting process effectively. Here, various supervised learning hypotheses are adopted as an operational engine. These learning hypotheses were implemented in numerous fields of pattern recognition, computational biology, data mining, and bioinformatics [15], [26], [27], [30]–[39]. In this study, we applied various powerful learning algorithms namely: K-nearest neighbor (KNN), Support Vector Machine (SVM), Probabilistic neural network (PNN), Random forest (RF), and Generalize regression neural network (GRNN). The basic idea of these algorithms has been explained and cited in the previous works [16], [25], [40]–[50]
3) **Ensemble Learning** Ensemble classification is a learning technique that is using for enhancing the prediction rate of individual learners as well as reducing generalization errors. Mostly, ensemble classification has obtained efficient performance compared to individual learner based systems due to its discrimination power, because it compensates the weakness of individual learners by each other [51]–[53]. However, there is no predefined rule that how to combine the number of learners in an efficient way. A number of different approaches are formulated to combine the learners. The simplest one is to fuse a large number of learners and then choose the optimal combination. Boosting is another ensemble technique in which, the single learner is re-trained iteratively in order to reduce classification error. Ensemble learning is mostly performed in two different ways, namely: majority voting and weighted voting. Majority voting is a simple approach in which a decision is made on the basis of the majority in a pool of input. In weighted voting, learners are not treated uniformly. Each learner is associated with a weight that is proportional to its performance. High weight leaner has more influence on the learning process. In addition, optimization techniques are also utilized in ensemble learning to minimize classification errors. Optimization techniques are employed in two different ways such as coverage optimization and decision optimization. Coverage optimization is the selection of optimum learners' subset from the utilized learners. On the other hand, decision optimization is the selection of optimal output by combining the predictions of multiple learners. Learner selection is the process to select the subset of k optimum learners from the pool of N learners, which have an advanced prediction rate. In this case, the possible combinations in solution space are NLk (N $\geq$ k), which shows that the solution space is exponential. This issue was resolved by applying an effective bio-inspired tool genetic algorithm (GA) widely used for solving the problem of local search. GA avoids local minima by utilizing crossover and mutation operators and tries to seek an optimum or near optimum solution employing probabilistic search techniques in massive and intricate search space. Few researchers have utilized GA in ensemble learning for learners' selection in order to obtain promising results [54]–[56]. In this research, five diverse nature of learning hypotheses; KNN, PNN, RF, GRNN, and, SVM is operated [18], [57]–[60]. KNN is an example-based learner who operates on the theory of proximity in the value of the attributes [61]. SVM is a powerful operational engine based on the statistical learning theory while PNN is established on Bayes theory [62]. First, the individual learners are trained and their outcomes are saved. Then these outcomes are forwarded to GA for ensemble learning. The process of GA is presented as follows:

- **Chromosome encoding**
  The first step of GA is to encode the solution into a chromosome. The size of the chromosome is limited to the number of learners in the pool and weight is assigned to each learner either 1 or 0 where 1 shows the learner is included in the learning process while 0 denotes the learner does not take part in learning. For instance, chromosome S = 10110110 illustrates that L1, L3, L4, L6, and L7 learners are taken place in ensemble learning. In this work, 100 population and 200 generations were used.
- **Initial population**
  The first step in the function of a GA is to randomly generate an initial population. Every member of this population encodes a conceivable answer for a problem. After making the initial population every member is calculated and allocated fitness value according to the fitness function.
- **Fitness function**
  The assessment of each individual is performed by the fitness function. The fitness value is computed by fusing the predicted outcomes of selected learners in the ensemble and finally, the decision is made on the basis of majority voting. In this study, accuracy is utilized as a fitness function.
- **Selection**
  A fitness-based methodology is used to select individual solutions in the selection process, where the fitter individuals measured by a fitness function, are more likely to be selected. In this study, two high fitness value chromosomes are selected as parents using roulette wheel-based selection.
- **Reproduction**
  In GA, a new generation (offspring) is reproduced by using genetic operators like crossover and/or mutation.

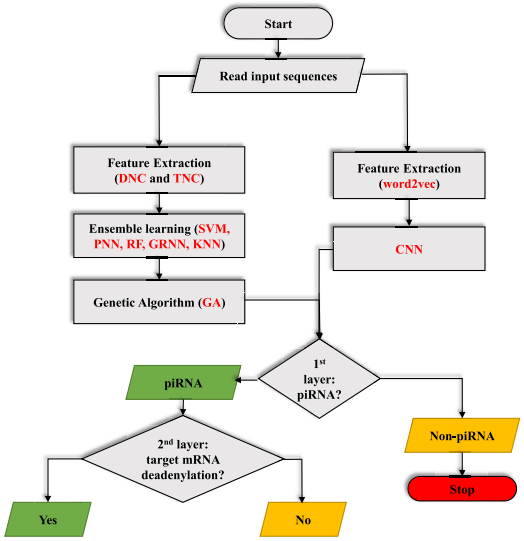


FIGURE 1. Framework of the proposed "piRNA-CNN" computational model.

Crossover is the exchange of information between the parents and offsprings; consequently, the generated offsprings may be better than parents. Here, the m-points crossover operator is used. The mutation operator is used to change the value of one or more genes in the selected chromosome.

- **Termination criteria**

The GA proceeds the next generation till the maximum number of generations and finally, the best solution is returned to the problem.

$$EL = KNN \oplus PNN \oplus SVM \oplus RF \oplus GRNN \quad (6)$$

where the symbol $\oplus$ represents the merging operator and EL represents Ensemble Learning. The procedure of how the ensemble learning functions by merging the five base learners are as per the following:

Assume the anticipated outcomes of a single learner for the genomic query R are $\{C_1, C_2, C_3, C_4, C_5\} \in \{S_1, S_2, S_4, S_5\}$ where C_1, C_2, C_3, C_4, C_5 are single learners and S_1, S_2, S_3, S_4, S_5 are piRNA.

$$Z_j = \sum_{i=1}^{5} \delta(C_i, S_j), \quad \{j = 1, 2, 3 \ldots..\} \quad (7)$$

$$\delta(C_i, S_j) = \begin{cases} 1 & if\ C_i \in S_j \\ 0 & otherwise \end{cases} \quad (8)$$

Lastly, the outcome of the ensemble learner merged through majority voting using GA is produced as:

$$GA_{EL} = Max\{x_1Y_1, x_2Y_2, x_3Y_3, x_4Y_4, x_5Y_5\} \quad (9)$$

where GA_{EL} is the anticipated outcome of the ensemble learner, the Max denotes selecting the maximum one, and x_1, x_2, x_3, x_4, x_5 is the weight of learners. The framework of the proposed system is illustrated in Figure 1.

TABLE 1. List of training parameters.

Parameters	World2vec model
Training Method	CBOW
Corpus	Human Genome
Window size	5
Epochs	20
Negative Sampling	5
Context words	3-mers
Vector size	100
Minimum Count	5

C. DEEP LEARNING APPROACHES

1) **Distributed Feature Representation:** The concept of natural language processing (NLP) was adopted by scientists in order to develop computational models for various biological applications, such as i.e., iN6-Methyl (5-step) [63], and alternative splicing site prediction [64]. Therefore, keeping the significance of NLP models in existing predictors, a distributed feature representation of natural languages processing technique i.e., word2vec method is applied in order to obtain interpretable representations for piwi-interacting RNAs and their functions. In this work, the genomes are collected from the Genbank of http:// hgdownload.soe.ucsc.edu, which are split into '21' chromosomes "Chr1", "Chr2", "Chr3", "Chr4", "Chr5", …, "X", and "Y". Moreover, the chromosome having a sequence length of 100nt is divided into sentences. The words are created by splitting each sentence using 3mers. The continuous bag-of-words (CBOW) approach is utilized in order to train the word2vec model. Whereas, CBOW is used to predict the current word "w(t)" according to the contiguous words in a predefined window. The training parameters of the proposed word2vec model are illustrated in Table 1. At last, after extracted features using the word2vec model, each extracted feature space has the dimension size of n*100, where "n" denotes the number of samples and 100 is the number of information/features against each sample. Moreover, each sample with length "L" is denoted ((L - 2) × 100).

2) **Convolutional Neural Network (CNN):** A CNN is a deep learning algorithm applied for the prediction of image processing as well as sequential based bioinformatics data [47], [65]–[67]. In this context, a one-dimensional (1D) CNN model is very effective for the prediction of the bioinformatics dataset. The architecture of CNN consists of convolution layer, ReLU layer, max-pooling layer, normalization layer, loss layer, dropout layer, fully connected layer. The CNN model is trained by several optimal hyper-parameters i.e., the size of the filters is [3], [5], [7], [9], number of filters are [10], [12], [14], [16], [18], number of convolution layers are [1]–[3], the padding values are same, the stride value is 1, the number of the neurons of the dense layer and dropout probability after dense and convolution layers. The range of dropout probability is [0.25, 0.3, 0.35]. The selection of hyper-parameters

is based on the higher prediction outcomes in terms of sensitivity, specificity, accuracy, MCC, and AUC. Moreover; the normalized class probability of the input data can be displayed using the sigmoid( ) function. These operators can be mathematically expressed as follows:

$$\text{Conv(R)}_{jk} = \text{ReLU}\left(\sum_{fs=0}^{FS-1}\sum_{f=0}^{F-1} W_{fsf}^{k} R_{j+fs,f}\right) \quad (10)$$

ReLU represents the rectified linear function and mathematically can be defined as $\text{ReLU}(y) = max(0, y)$

$$\text{Sigmoid}(y) = \frac{1}{1+e^{-y}} \quad (11)$$

In this work, the proposed "piRNA-CNN" model was implemented using Keras library in python [68]. On the other hand, the number of batch size = 64 and epochs = 100. To train the model, a minimum learning rate of 0.0004 is kept and Adam optimizer is utilized.

D. PREDICTION QUALITY MEASUREMENT

Various performance assessment measures are utilized to examine the success rates of learning algorithms [33], [48], [69]–[72]. Here, Accuracy (Acc), sensitivity (Sn), specificity (Sp) and Mathew's correlation coefficient (MCC) are employed.

$$Accuracy = \frac{T^{+}+T^{-}}{T^{+}+T^{-}+F^{+}+F^{-}} \times 100 \quad (12)$$

$$Sensitivity = \frac{T^{+}}{T^{+}+F^{-}} \times 100 \quad (13)$$

$$Specificity = \frac{T^{-}}{T^{-}+F^{+}} \times 100 \quad (14)$$

$$MCC = \frac{(T^{+}\times T^{-})-(F^{+}\times F^{-})}{\sqrt{(T^{+}+F^{+})(T^{+}+F^{-})(T^{-}+F^{-})(T^{-}+F^{+})}} \quad (15)$$

where T^{+}, F^{-}, T^{-}, and F^{+} indicate true positive, false negative, true negative, and false positive respectively.

E. CROSS-VALIDATION

In literature, there are three popular CV methods used for analysis and prediction: i.e., jackknife test, K-fold cross-validation (or subsampling) test, and independent dataset test. Though, the jackknife test yields unique results for a examine benchmark dataset with high time complexity. In contrast, the K-fold cross-validation test overcomes the complexity issue of the jackknife test along with performing the same characteristics of the former. Therefore, in this study, we have adopted the seven-fold CV test to assess the error rates of the proposed "piRNA-CNN" model. The feature spaces are split into seven subsets at each layer i.e., First-layer and Second-layer, where one subset is used for testing and the rest are used for training to measure the performance. This procedure was repeated seven times until each subset was used as a test set once [1], [73]. Therefore, every seven subsets were single out one by one to test the model and their average outcome is considered the final result.

TABLE 2. Success rates of classification algorithms on DNC and TNC and word2vec feature spaces.

Approaches	Classification Algorithms	Feature Extraction	Accuracy	Sensitivity	Specificity	MCC
	Frist-Layer					
Machine Learning	GRNN	DNN	61.27	75.88	48.07	0.248
		TNN	64.98	79.28	46.09	0.270
	RF	DNN	71.25	78.94	64.28	0.435
		TNN	75.00	52.11	86.84	0.521
	KNN	DNN	73.41	74.25	72.25	0.458
		TNN	77.75	84.13	71.36	0.559
	PNN	DNN	74.18	73.69	74.68	0.483
		TNN	79.44	84.55	74.33	0.592
	SVM	DNN	76.16	79.47	72.84	0.524
		TNN	83.18	84.69	81.66	0.663
	GA-Ensemble	DNN	84.13	84.73	83.53	0.682
		TNN	88.13	91.00	86.70	0.778
Deep Learing	CNN	Word2vec	**93.83**	**94.28**	**93.40**	**0.876**
	Second-Layer					
Machine Learning	GRNN	DNN	64.98	76.66	53.66	0.306
		TNN	67.22	66.41	67.87	0.341
	RF	DNN	67.50	63.15	71.42	0.347
		TNN	73.75	76.31	71.42	0.476
	KNN	DNN	72.14	72.91	71.36	0.442
		TNN	75.45	83.07	67.84	0.515
	PNN	DNN	73.34	74.33	72.35	0.466
		TNN	76.02	81.94	70.09	0.524
	SVM	DNN	74.11	74.75	73.48	0.482
		TNN	78.77	80.25	77.29	0.575
	GA-Ensemble	DNN	82.15	83.03	81.27	0.643
		TNN	86.03	88.28	83.78	0.722
Deep Learing	CNN	Word2vec	**91.19**	**89.86**	**92.64**	**0.824**

TABLE 3. Comparison of the proposed piRNA-CNN model with existing predication models.

Layers	Model	Accuracy	Sensitivity	Specificity	MCC
First-Layer	piRNA-CNN	93.83	94.28	93.40	0.876
	2L-piRNADNN	91.81	90.97	90.97	0.821
	piRNAPred	89.00	90.40	87.50	0.779
	2L-piRNA	86.45	87.60	85.30	0.732
Second-Layer	piRNA-CNN	91.19	89.86	92.64	0.824
	2L-piRNADNN	84.52	81.20	90.27	0.65
	piRNAPred	84.00	84.30	83.60	0.68
	2L-piRNA	80.40	81.50	79.30	0.590

III. RESULTS AND DISCUSSION

Table 2 demonstrates the prediction performance of machine learning and deep learning classification algorithms on various benchmark datasets. In machine learning classification algorithms SVM, KNN, RF, GRNN, and PNN along with ensemble models are utilized in combination with discrete feature spaces. On the other hand, in the Deep learning classification algorithm, CCN in conjunction with natural language processing technique based feature space is used. The **first-layer** prediction performance of the proposed model in Table 2 shows that the performance of the deep learning based approach is much better than not only individual machine learning approaches but also from the ensemble model. The success rates of the deep learning approach in terms of accuracy, sensitivity, and specificity are 4.98%, 3.28%, and 6.70%, respectively are improved than machine learning approaches. In the **second-layer**, the predictive outcome of the deep learning method is 5.16% higher than the highest result of the machine learning method. Finally, a comparison has been made between the developed model and the current state-of-the-art methods as shown in Table 3 on CV tests such as 7-folds. The pioneer works on these data have been carried out by 2L-piRNA, piRNAPred, and 2L-piRNADNN. After empirically examining the outcomes of the developed model and already existing models, it is observed that the accuracy of our developed computational model for the first-layer is 2.04% higher than

existing methods. Similarly, for the second layer, the developed computational model obtained 6.67% higher accuracy than existing methods. Establishing a user-friendly and open access web-predictor provides a practical platform for researchers in the design of pharmaceutical drugs and also expedient for academia as established in a series of recent publications [1], [74]–[81].

IV. CONCLUSION

An attempt was made to develop an intelligent and high-throughput computational model namely "piRNA-CNN" for the identification of piRNA and non-piRNA, in this study. Here, analysis has been drawn between machine learning algorithms and deep learning algorithms. First, two discrete feature encoding methods such as DNC and TNC are applied to excerpt numerical values from RNA sequences. Then these feature spaces are provided to five machine learning algorithms and noted their outcomes. Furthermore, the concept of ensemble learning is adapted to merge the prediction of individual learners in order to minimize variance instigated by the peculiarities of a single training. It is shown that ensemble learning with TNC feature space achieved efficient outcomes compared to individual learners. The ensemble was carried out via GA. In contrast, RNA sequences are expressed by the natural language processing technique word2vec. Then the obtained feature space is provided to deep learning algorithm CNN for prediction of piRNAs. The results demonstrate that the success rate of the CNN base model is much better than machine learning algorithms. In conclusion, the obtained outcomes authenticated that the piRNA-CNN model exposed substantial results matched to the current tools stated in the literature, so far. It is further expected that the suggested predictive tool will assist scientists and researchers to design improved computational tools.

REFERENCES

[1] B. Liu, F. Yang, and K.-C. Chou, "2L-piRNA: A two-layer ensemble classifier for identifying PIWI-interacting RNAs and their function," *Mol. Therapy-Nucleic Acids*, vol. 7, pp. 267–277, Jun. 2017.

[2] J. Cheng, H. Deng, B. Xiao, H. Zhou, F. Zhou, Z. Shen, and J. Guo, "PiR-823, a novel non-coding small RNA, demonstrates *in vitro* and *in vivo* tumor suppressive activity in human gastric cancer cells," *Cancer Lett.*, vol. 315, no. 1, pp. 12–17, Feb. 2012.

[3] M. Moyano and G. Stefani, "PiRNA involvement in genome stability and human cancer," *J. Hematol. Oncol.*, vol. 8, no. 1, p. 38, Dec. 2015.

[4] A. Hashim, F. Rizzo, G. Marchese, M. Ravo, R. Tarallo, G. Nassa, G. Giurato, G. Santamaria, A. Cordella, C. Cantarella, and A. Weisz, "RNA sequencing identifies specific PIWI-interacting small non-coding RNA expression patterns in breast cancer," *Oncotarget*, vol. 5, no. 20, p. 9901, 2014.

[5] E. J. Lee, S. Banerjee, H. Zhou, A. Jammalamadaka, M. Arcila, B. S. Manjunath, and K. S. Kosik, "Identification of piRNAs in the central nervous system," *RNA*, vol. 17, no. 6, pp. 1090–1099, Jun. 2011.

[6] T. Nishibu, Y. Hayashida, S. Tani, S. Kurono, K. Kojima-Kita, R. Ukekawa, T. Kurokawa, S. Kuramochi-Miyagawa, T. Nakano, K. Inoue, and S. Honda, "Identification of MIWI-associated poly(A) RNAs by immunoprecipitation with an anti-MIWI monoclonal antibody," *BioScience Trends*, vol. 6, no. 5, pp. 248–261, Nov. 2012.

[7] L.-T. Gou, P. Dai, J.-H. Yang, Y. Xue, Y.-P. Hu, Y. Zhou, J.-Y. Kang, X. Wang, H. Li, M.-M. Hua, S. Zhao, S.-D. Hu, L.-G. Wu, H.-J. Shi, Y. Li, X.-D. Fu, L.-H. Qu, E.-D. Wang, and M.-F. Liu, "Pachytene piRNAs instruct massive mRNA elimination during late spermiogenesis," *Cell Res.*, vol. 24, no. 6, pp. 680–700, Jun. 2014.

[8] Y. Zhang, X. Wang, and L. Kang, "A k-mer scheme to predict piRNAs and characterize locust piRNAs," *Bioinformatics*, vol. 27, no. 6, pp. 771–776, Mar. 2011.

[9] K. Wang, C. Liang, J. Liu, H. Xiao, S. Huang, J. Xu, and F. Li, "Prediction of piRNAs using transposon interaction and a support vector machine," *BMC Bioinf.*, vol. 15, no. 1, p. 419, Dec. 2014.

[10] L. Luo, D. Li, W. Zhang, S. Tu, X. Zhu, and G. Tian, "Accurate prediction of transposon-derived piRNAs by integrating various sequential and physicochemical features," *PLoS ONE*, vol. 11, no. 4, Apr. 2016, Art. no. e0153268.

[11] D. Li, L. Luo, W. Zhang, F. Liu, and F. Luo, "A genetic algorithm-based weighted ensemble method for predicting transposon-derived piRNAs," *BMC Bioinf.*, vol. 17, no. 1, p. 329, Dec. 2016.

[12] T. Li, M. Gao, R. Song, Q. Yin, and Y. Chen, "Support vector machine classifier for accurate identification of piRNA," *Appl. Sci.*, vol. 8, no. 11, p. 2204, Nov. 2018.

[13] S. Khan, M. Khan, N. Iqbal, T. Hussain, S. A. Khan, and K.-C. Chou, "A two-level computation model based on deep learning algorithm for identification of piRNA and their functions via Chou's 5-steps rule," *Int. J. Peptide Res. Therapeutics*, vol. 26, pp. 795–809, 2020.

[14] K.-C. Chou, "Some remarks on protein attribute prediction and pseudo amino acid composition," *J. Theor. Biol.*, vol. 273, no. 1, pp. 236–247, Mar. 2011.

[15] M. Hayat and A. Khan, "Predicting membrane protein types by fusing composite protein sequence features into pseudo amino acid composition," *J. Theor. Biol.*, vol. 271, no. 1, pp. 10–17, Feb. 2011.

[16] M. Hayat and A. Khan, "MemHyb: Predicting membrane protein types by hybridizing SAAC and PSSM," *J. Theor. Biol.*, vol. 292, pp. 93–102, Jan. 2012.

[17] H. Mohabatkar, "Prediction of cyclin proteins using Chou's pseudo amino acid composition," *Protein Peptide Lett.*, vol. 17, no. 10, pp. 1207–1214, Oct. 2010.

[18] M. Hayat, A. Khan, and M. Yeasin, "Prediction of membrane proteins using split amino acid and ensemble classification," *Amino Acids*, vol. 42, no. 6, pp. 2447–2460, Jun. 2012.

[19] H. Lin and H. Ding, "Predicting ion channels and their types by the dipeptide mode of pseudo amino acid composition," *J. Theor. Biol.*, vol. 269, no. 1, pp. 64–69, Jan. 2011.

[20] S.-H. Guo, E.-Z. Deng, L.-Q. Xu, H. Ding, H. Lin, W. Chen, and K.-C. Chou, "INuc-PseKNC: A sequence-based predictor for predicting nucleosome positioning in genomes with pseudo K-tuple nucleotide composition," *Bioinformatics*, vol. 30, no. 11, pp. 1522–1529, Jun. 2014.

[21] H. Lin, E.-Z. Deng, H. Ding, W. Chen, and K.-C. Chou, "IPro54-PseKNC: A sequence-based predictor for identifying sigma-54 promoters in prokaryote with pseudo K-tuple nucleotide composition," *Nucleic Acids Res.*, vol. 42, no. 21, pp. 12961–12972, Dec. 2014.

[22] W. Chen, T.-Y. Lei, D.-C. Jin, H. Lin, and K.-C. Chou, "PseKNC: A flexible Web server for generating pseudo K-tuple nucleotide composition," *Anal. Biochem.*, vol. 456, pp. 53–60, Jul. 2014.

[23] W. Chen, X. Zhang, J. Brooker, H. Lin, L. Zhang, and K.-C. Chou, "PseKNC-general: A cross-platform package for generating various modes of pseudo nucleotide compositions," *Bioinformatics*, vol. 31, no. 1, pp. 119–120, Jan. 2015.

[24] W.-R. Qiu, S.-Y. Jiang, B.-Q. Sun, X. Xiao, X. Cheng, and K.-C. Chou, "IRNA-2methyl: Identify RNA 2'-O-methylation sites by incorporating sequence-coupled effects into general PseKNC and ensemble classifier," *Medicinal Chem.*, vol. 13, no. 8, pp. 734–743, Nov. 2017.

[25] M. Kabir, M. Iqbal, S. Ahmad, and M. Hayat, "ITIS-PseKNC: Identification of translation initiation site in human genes using pseudo K-tuple nucleotides composition," *Comput. Biol. Med.*, vol. 66, pp. 252–257, Nov. 2015.

[26] M. Tahir and M. Hayat, "INuc-STNC: A sequence-based predictor for identification of nucleosome positioning in genomes by extending the concept of SAAC and Chou's PseAAC," *Mol. BioSystems*, vol. 12, no. 8, pp. 2587–2593, 2016.

[27] M. Tahir, M. Hayat, and M. Kabir, "Sequence based predictor for discrimination of enhancer and their types by applying general form of Chou's trinucleotide composition," *Comput. Methods Programs Biomed.*, vol. 146, pp. 69–75, Jul. 2017.

[28] M. Tahir, M. Hayat, and S. A. Khan, "A two-layer computational model for discrimination of enhancer and their types using hybrid features pace of pseudo K-tuple nucleotide composition," *Arabian J. Sci. Eng.*, vol. 43, no. 12, pp. 6719–6727, Dec. 2018.

[29] W. Chen, P.-M. Feng, H. Lin, and K.-C. Chou, "IRSpot-PseDNC: Identify recombination spots with pseudo dinucleotide composition," *Nucleic Acids Res.*, vol. 41, no. 6, p. e68, Apr. 2013.

[30] P.-M. Feng, W. Chen, H. Lin, and K.-C. Chou, "IHSP-PseRAAAC: Identifying the heat shock protein families using pseudo reduced amino acid alphabet composition," *Anal. Biochem.*, vol. 442, no. 1, pp. 118–125, Nov. 2013.

[31] G.-S. Han, Z.-G. Yu, and V. Anh, "A two-stage SVM method to predict membrane protein types by incorporating amino acid classifications and physicochemical properties into a general form of Chou's PseAAC," *J. Theor. Biol.*, vol. 344, pp. 31–39, Mar. 2014.

[32] B. Liu, D. Zhang, R. Xu, J. Xu, X. Wang, Q. Chen, Q. Dong, and K.-C. Chou, "Combining evolutionary information extracted from frequency profiles with sequence-based kernels for protein remote homology detection," *Bioinformatics*, vol. 30, no. 4, pp. 472–479, Feb. 2014.

[33] W.-R. Qiu, X. Xiao, W.-Z. Lin, and K.-C. Chou, "IUbiq-Lys: Prediction of lysine ubiquitination sites in proteins by extracting sequence evolution information via a gray system model," *J. Biomolecular Struct. Dyn.*, vol. 33, no. 8, pp. 1731–1742, Aug. 2015.

[34] B. Liu, L. Fang, S. Wang, X. Wang, H. Li, and K.-C. Chou, "Identification of microRNA precursor with the degenerate K-tuple or Kmer strategy," *J. Theor. Biol.*, vol. 385, pp. 153–159, Nov. 2015.

[35] B. Liu, L. Fang, R. Long, X. Lan, and K.-C. Chou, "IEnhancer-2L: A two-layer predictor for identifying enhancers and their strength by pseudo K-tuple nucleotide composition," *Bioinformatics*, vol. 32, no. 3, pp. 362–369, Feb. 2016.

[36] M. Tahir and M. Hayat, "Machine learning based identification of protein–protein interactions using derived features of physiochemical properties and evolutionary profiles," *Artif. Intell. Med.*, vol. 78, pp. 61–71, May 2017.

[37] M. Iqbal and M. Hayat, "'iSS-Hyb-mRMR': Identification of splicing sites using hybrid space of pseudo trinucleotide and pseudo tetranucleotide composition," *Comput. Methods Programs Biomed.*, vol. 128, pp. 1–11, May 2016.

[38] A. Khan, A. Sohail, U. Zahoora, and A. S. Qureshi, "A survey of the recent architectures of deep convolutional neural networks," *Artif. Intell. Rev.*, vol. 53, no. 8, pp. 5455–5516, Dec. 2020.

[39] A. Khan, A. S. Qureshi, N. Wahab, M. Hussain, and M. Y. Hamza, "A recent survey on the applications of genetic programming in image processing," 2019, *arXiv:1901.07387*. [Online]. Available: http://arxiv.org/abs/1901.07387

[40] P. Feng, H. Yang, H. Ding, H. Lin, W. Chen, and K.-C. Chou, "IDNA6 mA-PseKNC: Identifying DNA N6-methyladenosine sites by incorporating nucleotide physicochemical properties into PseKNC," *Genomics*, vol. 111, no. 1, pp. 96–102, Jan. 2019.

[41] B. Liu, S. Wang, R. Long, and K.-C. Chou, "IRSpot-EL: Identify recombination spots with an ensemble learning approach," *Bioinformatics*, vol. 33, no. 1, pp. 35–41, Jan. 2017.

[42] M. Tahir, H. Tayara, and K. T. Chong, "IRNA-PseKNC(2methyl): Identify RNA 2'-O-methylation sites by convolution neural network and Chou's pseudo components," *J. Theor. Biol.*, vol. 465, pp. 1–6, Mar. 2019.

[43] M. Tahir, B. Jan, M. Hayat, S. U. Shah, and M. Amin, "Efficient computational model for classification of protein localization images using extended threshold adjacency statistics and support vector machines," *Comput. Methods Programs Biomed.*, vol. 157, pp. 205–215, Apr. 2018.

[44] M. Tahir, M. Hayat, and S. A. Khan, "INuc-ext-PseTNC: An efficient ensemble model for identification of nucleosome positioning by extending the concept of Chou's PseAAC to pseudo-tri-nucleotide composition," *Mol. Genet. Genomics*, vol. 294, no. 1, pp. 199–210, Feb. 2019.

[45] S. Akbar and M. Hayat, "IMethyl-STTNC: Identification of N6-methyladenosine sites by extending the idea of SAAC into Chou's PseAAC to formulate RNA sequences," *J. Theor. Biol.*, vol. 455, pp. 205–211, Oct. 2018.

[46] M. Kabir and M. Hayat, "IRSpot-GAEnsC: Identifing recombination spots via ensemble classifier and extending the concept of Chou's PseAAC to formulate DNA samples," *Mol. Genet. Genomics*, vol. 291, no. 1, pp. 285–296, Feb. 2016.

[47] M. Cherian and S. P. Sathiyan, "Neural network based ACC for optimized safety and comfort," *Int. J. Comput. Appl.*, vol. 42, no. 14, pp. 1–4, Mar. 2012.

[48] L. Breiman, "Random forests," *Mach. Learn.*, vol. 45, no. 1, pp. 5–32, 2001.

[49] W. Lou, X. Wang, F. Chen, Y. Chen, B. Jiang, and H. Zhang, "Sequence based prediction of DNA-binding proteins based on hybrid feature selection using random forest and Gaussian Naïve Bayes," *PLoS ONE*, vol. 9, no. 1, Jan. 2014, Art. no. e86703.

[50] P. Jiang, H. Wu, W. Wang, W. Ma, X. Sun, and Z. Lu, "MiPred: Classification of real and pseudo microRNA precursors using random forest prediction model with combined features," *Nucleic Acids Res.*, vol. 35, pp. W339–W344, May 2007.

[51] H.-B. Shen and K.-C. Chou, "Using ensemble classifier to identify membrane protein types," *Amino Acids*, vol. 32, no. 4, pp. 483–488, May 2007.

[52] L. Nanni and A. Lumini, "Ensemblator: An ensemble of classifiers for reliable classification of biological data," *Pattern Recognit. Lett.*, vol. 28, no. 5, pp. 622–630, Apr. 2007.

[53] L. Nanni and A. Lumini, "Genetic programming for creating Chou's pseudo amino acid based features for submitochondria localization," *Amino Acids*, vol. 34, no. 4, pp. 653–660, May 2008.

[54] Z.-H. Zhou, J. Wu, and W. Tang, "Ensembling neural networks: Many could be better than all," *Artif. Intell.*, vol. 137, nos. 1–2, pp. 239–263, May 2002.

[55] R. Moghdani and K. Salimifard, "Volleyball premier league algorithm," *Appl. Soft Comput.*, vol. 64, pp. 161–185, Mar. 2018.

[56] L. S. Oliveira, R. Sabourin, F. Bortolozzi, and C. Y. Suen, "A methodology for feature selection using multiobjective genetic algorithms for handwritten digit string recognition," *Int. J. Pattern Recognit. Artif. Intell.*, vol. 17, no. 06, pp. 903–929, Sep. 2003.

[57] K.-C. Chou and Y.-D. Cai, "Using functional domain composition and support vector machines for prediction of protein subcellular location," *J. Biol. Chem.*, vol. 277, no. 48, pp. 45765–45769, Nov. 2002.

[58] Y.-D. Cai, G.-P. Zhou, and K.-C. Chou, "Support vector machines for predicting membrane protein types by using functional domain composition," *Biophys. J.*, vol. 84, no. 5, pp. 3257–3263, May 2003.

[59] A. Khan, S. F. Tahir, A. Majid, and T.-S. Choi, "Machine learning based adaptive watermark decoding in view of anticipated attack," *Pattern Recognit.*, vol. 41, no. 8, pp. 2594–2610, Aug. 2008.

[60] A. Khan, A. Majid, and T.-S. Choi, "Predicting protein subcellular location: Exploiting amino acid based sequence of feature spaces and fusion of diverse classifiers," *Amino Acids*, vol. 38, no. 1, pp. 347–350, Jan. 2010.

[61] J. Ahmad, F. Javed, and M. Hayat, "Intelligent computational model for classification of sub-golgi protein using oversampling and Fisher feature selection methods," *Artif. Intell. Med.*, vol. 78, pp. 14–22, May 2017.

[62] S. Ahmad, M. Kabir, and M. Hayat, "Identification of heat shock protein families and J-protein types by incorporating dipeptide composition into Chou's general PseAAC," *Comput. Methods Programs Biomed.*, vol. 122, no. 2, pp. 165–174, Nov. 2015.

[63] I. Nazari, M. Tahir, H. Tayara, and K. T. Chong, "IN6-methyl (5-step): Identifying RNA N6-methyladenosine sites using deep learning mode via Chou's 5-step rules and Chou's general PseKNC," *Chemometric Intell. Lab. Syst.*, vol. 193, Oct. 2019, Art. no. 103811.

[64] M. Oubounyt, Z. Louadi, H. Tayara, and K. To Chong, "Deep learning models based on distributed feature representations for alternative splicing prediction," *IEEE Access*, vol. 6, pp. 58826–58834, 2018.

[65] S. Bai, J. Z. Kolter, and V. Koltun, "An empirical evaluation of generic convolutional and recurrent networks for sequence modeling," 2018, *arXiv:1803.01271*. [Online]. Available: http://arxiv.org/abs/1803.01271

[66] M. Tahir, H. Tayara, and K. T. Chong, "IPseU-CNN: Identifying RNA pseudouridine sites using convolutional neural networks," *Mol. Therapy-Nucleic Acids*, vol. 16, pp. 463–470, Jun. 2019.

[67] H. Tayara, M. Tahir, and K. T. Chong, "Identification of prokaryotic promoters and their strength by integrating heterogeneous features," *Genomics*, vol. 112, no. 2, pp. 1396–1403, Mar. 2020.

[68] F. Chollet. (2015). *Keras: Deep Learning Library for Theano and Tensorflow*. [Online]. Available: https://keras.io/k

[69] W. Chen, H. Lv, F. Nie, and H. Lin, "I6 mA-pred: Identifying DNA N^6-methyladenine sites in the rice genome," *Bioinformatics*, vol. 35, no. 16, pp. 2796–2800, Aug. 2019.

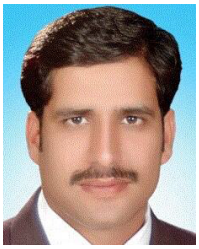
[70] W. Alam, S. D. Ali, H. Tayara, and K. T. Chong, "A CNN-based RNA N6-methyladenosine site predictor for multiple species using heterogeneous features representation," *IEEE Access*, vol. 8, pp. 138203–138209, 2020.

[71] W. Alam, H. Tayara, and K. T. Chong, "XG-ac4C: Identification of N4-acetylcytidine (ac4C) in mRNA using extreme gradient boosting with electron-ion interaction pseudopotentials," *Sci. Rep.*, vol. 10, no. 1, pp. 1–10, Dec. 2020.

[72] S. D. Ali, W. Alam, H. Tayara, and K. Chong, "Identification of functional piRNAs using a convolutional neural network," *IEEE/ACM Trans. Comput. Biol. Bioinf.*, early access, Oct. 29, 2020, doi: 10.1109/TCBB.2020.3034313.
[73] Z.-C. Xu, P. Wang, W.-R. Qiu, and X. Xiao, "ISS-PC: Identifying splicing sites via physical-chemical properties using deep sparse auto-encoder," *Sci. Rep.*, vol. 7, no. 1, pp. 1–12, Dec. 2017.
[74] X. Xiao, H.-X. Ye, Z. Liu, J.-H. Jia, and K.-C. Chou, "iROS-gPseKNC: Predicting replication origin sites in DNA by incorporating dinucleotide position-specific propensity into general pseudo nucleotide composition," *Oncotarget*, vol. 7, no. 23, p. 34180, 2016.
[75] C.-J. Zhang, H. Tang, W.-C. Li, H. Lin, W. Chen, and K.-C. Chou, "iOri-human: Identify human origin of replication by incorporating dinucleotide physicochemical properties into pseudo nucleotide composition," *Oncotarget*, vol. 7, no. 43, p. 69783–69793, 2016.
[76] W. Chen, P. Feng, H. Yang, H. Ding, H. Lin, and K.-C. Chou, "iRNA-AI: Identifying the adenosine to inosine editing sites in RNA sequences," *Oncotarget*, vol. 8, no. 3, p. 4208, 2017.
[77] W.-R. Qiu, X. Xiao, Z.-C. Xu, and K.-C. Chou, "iPhos-PseEn: Identifying phosphorylation sites in proteins by fusing different pseudo components into an ensemble classifier," *Oncotarget*, vol. 7, no. 32, p. 51270, 2016.
[78] X. Cheng, S. G. Zhao, X. Xiao, and K. C. Chou, "iATC-mISF: A multi-label classifier for predicting the classes of anatomical therapeutic chemicals," *Bioinformatics*, vol. 33, no. 3, pp. 341–346, Feb. 2017.
[79] W. Chen, H. Ding, P. Feng, H. Lin, and K.-C. Chou, "iACP: A sequence-based tool for identifying anticancer peptides," *Oncotarget*, vol. 7, no. 13, p. 16895, 2016.
[80] W. Chen, H. Tang, J. Ye, H. Lin, and K.-C. Chou, "iRNA-PSEU: Identifying RNA pseudouridine sites," *Mol. Therapy-Nucleic Acids*, vol. 5, no. 7, p. e332, 2016.
[81] B. Liu, H. Wu, D. Zhang, X. Wang, and K.-C. Chou, "Pse-Analysis: A Python package for DNA/RNA and protein/peptide sequence analysis based on pseudo components and kernel methods," *Oncotarget*, vol. 8, no. 8, p. 13338, 2017.

MUHAMMAD TAHIR received the Master of Information (M.I.T.) degree from Gomal University D. I. Khan, in 2005, the M.S. (CS) degree in multimedia and communication from Mohammad Ali Jinnah University (MAJU), Islamabad, in 2011, and the Ph.D. degree from the Department of Computer Science, Abdul Wali Khan University Mardan (AWKUM), Pakistan. He has completed Postdoctoral Research with the Advanced Electronics and Information Research Center, Jeonbuk National University, Jeonju, South Korea. He has been working as a Lecturer with the Department of Computer Science, AWKUM, since November 2010. His main research interests include bioinformatics, machine learning, and deep learning.

MAQSOOD HAYAT received the M.C.S. degree from Gomal University D. I. Khan, in 2004, the M.S. degree in software and system engineering from Mohammad Ali Jinnah University (MAJU), Islamabad, Pakistan, in 2009, and the Ph.D. degree from the Department of Computer and Information Sciences, Pakistan Institute of Engineering and Applied Sciences, Islamabad. He is currently working as an Associate Professor with the Department of Computer Science, Abdul Wali Khan University Mardan, Pakistan. His main research interests include machine learning, pattern recognition, evolutionary computing, and its application in bioinformatics.



SHAHZAD KHAN received the Master of Computer Science (M.Sc.) degree from Gomal University D. I. Khan, in 2004, and the M.S. degree in computer science from Abdul Wali Khan University Mardan, Pakistan, in 2017. His main research interests include bioinformatics, machine learning, and deep learning.

KIL TO CHONG received the Ph.D. degree in mechanical engineering from Texas A&M University, in 1995. He is currently a Professor with the School of Electronics and Information Engineering, Chonbuk National University, Jeonju, South Korea, and the Head of the Advanced Research Center of Electronics. His research interests include machine learning, signal processing, motor fault detection, network system control, and time-delay systems.