Journal of information and communication convergence engineering
http://jicce.org

J. Inf. Commun. Converg. Eng. 21(4): 294-299, Dec. 2023 Regular paper

MicroRNA–Gene Association Prediction Method using Deep Learning Models

Seung-Won Yoon, In-Woo Hwang, and Kyu-Chul Lee*, *Member, KIICE*

Department of Computer Science, Chungnam National University, Daejeon 34134, Republic of Korea

Abstract

Micro ribonucleic acids (miRNAs) can regulate the protein expression levels of genes in the human body and have recently been reported to be closely related to the cause of disease. Determining the genes related to miRNAs will aid in understanding the mechanisms underlying complex miRNAs. However, the identification of miRNA-related genes through wet experiments (in vivo, traditional methods are time- and cost-consuming). To overcome these problems, recent studies have investigated the prediction of miRNA relevance using deep learning models. This study presents a method for predicting the relationships between miRNAs and genes. First, we reconstruct a negative dataset using the proposed method. We then extracted the feature using an autoencoder, after which the feature vector was concatenated with the original data. Thereafter, the concatenated data were used to train a long short-term memory model. Our model exhibited an area under the curve of 0.9609, outperforming previously reported models trained using the same dataset.

Index Terms: Deep Learning, miRNA, Gene, Association prediction, Genomics

I. INTRODUCTION

The genetic information of deoxy ribonucleic acid (DNA) comprises RNA, and genes are expressed by synthesizing proteins. Most diseases are caused by abnormal production and overexpression of related proteins. MicroRNAs (miRNAs) are crucial in regulating the expression of genes as proteins in the human body. MicroRNAs (miRNAs) are single-stranded RNA (small-RNA) consisting of 21-25 nucleotides. It regulates cell proliferation, differentiation, and death by decomposing messenger RNA (mRNA), a template for synthesizing proteins that drive life phenomena. This implies that it can block the production of disease-causing proteins. Moreover, recent studies have reported a close relationship between miRNAs and the causes of diseases; accordingly, miRNAs have been actively investigated in the fields of biology and pharmaceuticals [1].

Research into miRNA-related genes is vital for understanding the complex roles and mechanisms of miRNAs in gene regulation. Furthermore, identifying miRNA-related genes can significantly contribute to the prediction of miRNA-related diseases and the potential of miRNAs for disease treatment. In addition, it can contribute to research on developing drugs related to miRNAs and control the degree of miRNA activity [2].

The identification of miRNA-related genes using traditional biopharmaceutical experiments is time- and cost-consuming owing to the difficulties associated with obtaining miRNAs and their complex relationships. Therefore, as an alternative to direct experiments, computational methods have been employed to predict the relationship between miRNAs and genes using models.

Accordingly, this study employed a deep learning model to predict the relationships between miRNAs and genes.

Received 21 July 2023, **Revised** 7 September 2023, **Accepted** 16 September 2023
***Corresponding Author** Kyu-Chul Lee (E-mail: kclee@cnu.ac.kr)
Department of Computer Science, Chungnam National University, Daejeon 34134, Republic of Korea

Open Access https://doi.org/10.56977/jicce.2023.21.4.294 print ISSN: 2234-8255 online ISSN: 2234-8883

The contributions of this study are as follows:

- We adjust the negative label with an adjusted threshold by examining the negative data distribution.
- We extract the feature vector of the data using an autoencoder and concatenate the vector with the original data, thereby enhancing the prediction performance of the model.
- We developed an miRNA-gene association prediction LSTM model with outstanding performance using concatenated data as input data.

The remainder of this paper is organized as follows. Section 2 presents the related studies on association prediction methods. Section 3 describes the proposed association prediction method. In Section 4, we present our dataset and results. Finally, the conclusions are presented in Section 5 and directions for future research are outlined.

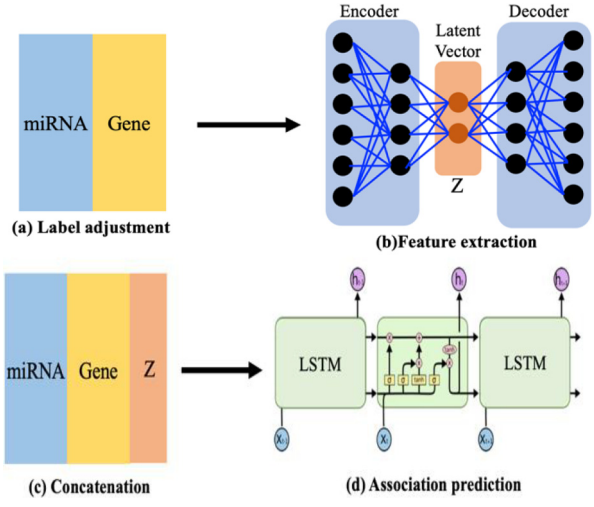

Fig. 1. Overall workflow of our proposed method. (a) Labeling was adjusted by setting the threshold more precisely, and the label-adjusted data was input into an autoencoder. (b) The feature of the data was extracted using the autoencoder, and (c) the feature was concatenated with the original data. (d) Prediction of the association by inputting the concatenated data into the relevance prediction deep learning model.

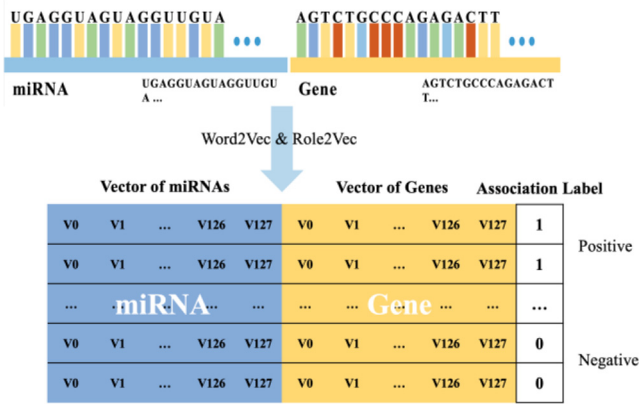

Fig. 2. Example of miRNA and gene sequence composition and input data vectorized by Word2Vec and Role2Vec

II. RELATED WORK

Several studies have been conducted to predict the correlations between miRNAs, genes, and diseases. Most of these studies relied on rule-based and machine learning methods. However, these methods have inherent limitations, particularly in the study of gene clusters. Owing to the complex nature of data, researchers must individually designate features [3].

To address these challenges, the deep learning approach has emerged as a promising alternative for predicting gene-domain relationships. One study used restricted Boltzmann machines (RBMs) to predict miRNA-gene associations [4]. RBMs and hierarchical and energy-based models have been pivotal in the realm of deep learning research. This study incorporated the sequence data of miRNAs and genes with frequency-based weights and made predictions based on functional similarity. DeepMirTar[5] employs an autoencoder deep learning model to predict the association between miRNAs and mRNAs. In this study, miRNA and mRNAs sequences were subjected to one-hot encoding with negative data randomly generated using the miRanda algorithm. Additionally, research has been conducted using a combination of convolutional neural network (CNN) and Bi-recurrent neural network (RNN) to forecast miRNA-gene associations [6]. This involves embedding miRNAs and genes using one-hot encoding and the random generation of negative data. The CNN model was instrumental in extracting features, and the predictions were based on the Bi- long short-term memory (LSTM) model. Furthermore, the application of MiRTDL to a CNN as a classifier led to an impressive association prediction accuracy of 89.88%[7]. Another notable example is SG-LSTM[8], which utilizes LSTM to predict miRNA-gene correlations, boasting an association prediction of area under of curve (AUC) = 0.94.

Deep learning models, which are recognized for their speed and enhanced computational performance compared to traditional mathematical models, are gaining traction in the gene association prediction domain. In line with this trend, our study introduced a novel deep learning model dedicated to predicting the relationships between miRNAs and genes.

III. SYSTEM MODEL AND METHODS

In this section, we describe our proposed miRNA-gene prediction method. Fig. 1 shows the overall workflow of this study.

A. Label Adjustment

To train the association prediction deep learning model, a labeled dataset is essential (1: positive/0: negative). Informa-

tion on positive miRNA-gene association data (label: 1) was obtained through various wet experiments [9]. Information on negative miRNA-gene association data (label: 0) did not exist in any wet experiments. This is because no absolute criteria exist for identifying biologically unrelated miRNA genes. However, in computational research, there is a good method known as vectorization, which can suggest a criterion in which there is no correlation if the distance between the vectors is relatively long.

In this study, the SG-LSTM-core [8] data set. The positive dataset refers to various open experimental results [9], whereas the negative dataset is created based on data at a greater distance (less similar) than the corresponding threshold by designating the average distance of the positive data as a threshold. The miRNA and gene data vectorized using Word2Vec and Role2Vec were separated in the vector space (Fig. 2). We found that the distance in the vector space is related to relevance; therefore, unlike other related studies that set negative datasets as random pairs, we set as negative data sets those data that are more distant than the average value of the distance between pairs of positive datasets. In this process, we used Euclidean distance and cosine similarity as the distance criteria. The Euclidean distance is most commonly used to calculate the distance between two vectors, and the cosine similarity is the degree of similarity between vectors measured by the cosine of the angle between two vectors in inner space. The miRNA-gene distance refers to the measurement result of the Euclidean distance and cosine similarity between the miRNA and gene by vectorizing the miRNA and gene. This implies that accurate negative data can be generated depending on the sophistication of the threshold. Therefore, the threshold is crucial in the performance of deep learning models. Therefore, we vectorized the data to understand the distribution, after which it was adjusted to the threshold that demonstrated the best performance.

B. Autoencoder

An autoencoder is a deep learning model consisting of an encoder that compresses the features of the input data into a z (feature) vector and a decoder that attempts to create the same data as the input data through the z vector. The input and output data of the autoencoder were of the same size, and the loss function was the difference between the input data and the data generated by the decoder. Therefore, the performance of the autoencoder depends on the accuracy of feature extraction from the input data. This indicates that an autoencoder is a deep learning model that specializes in extracting the features of the input data.

In this study, the encoder consists of three stacks of layers and the decoder consists of three stacks of layers. The miRNA-gene features were extracted using an autoencoder (Fig. 3), after which they were concatenated into the original data to improve the performance of the association prediction model.

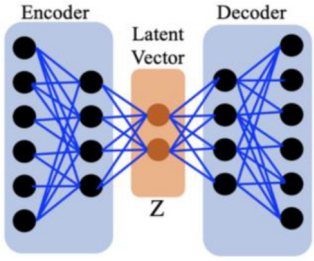

Fig. 3. Structure of the autoencoder model used in this study. The autoencoder consisted of an encoder, latent vector, and a decoder. The latent vector of the autoencoder can be considered as a compressed feature of the input data.

C. LSTM deep learning model

Probabilistic methods, such as the Gaussian maximum likelihood model and hidden Markov model, have been mainly used to predict sequential data classes. However, these probabilistic methods cannot provide high-performance predictions because of their inability to learn large amounts of sequential data. However, deep learning models can handle the problem of nonlinear dependency and learn large amounts of sequential data. As shown in Fig. 2, the miRNAs and genes were composed of the sequence data. The sequences that constitute each feature represent important features. Therefore, deep learning models that specialize in time-series data, such as RNN, are well suited for this study because they can understand the flow of sequences. An RNN, a deep learning model optimized for sequential data learning, can be used by sharing previous information; however, when the distance between the previous data and the point using the data is long, the gradient gradually decreases, and the learning ability is significantly reduced. Consequently, the vanishing gradient problem occurs. This problem can be addressed using LSTM by adding (1) cell state (*Ct*) (Fig. 4) to the hidden state of the RNN. The update process for the cell state is as follows: First, an elementary multiplication operation is performed on the previous cell state value (*Ct*-1) and the output of the forget gate (*ft*), which determines how much to forget. Thereafter, it is updated with the current cell state value and the output of the input gate, which determines the amount to memorize using an elementary multiplication operation. In this study, an LSTM deep learning model was developed to predict the relationship between miRNAs and genes.

The number of LSTM layers (Fig. 2(c)) was set to three. To predict the relationship between miRNAs and genes, previously concatenated data were used as inputs for the LSTM model.

$$C_t = f_t * C_{t-1} + i_t * \tilde{C}_t \quad (1)$$

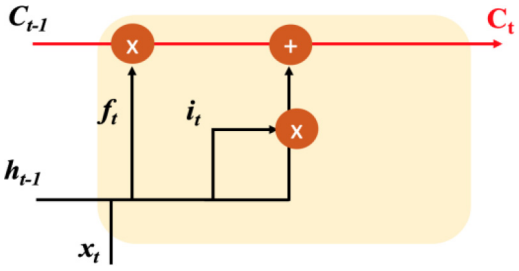

Fig. 4. Cell state equation of the Long Short-Term Model (LSTM). Cell state is the key to LSTM.

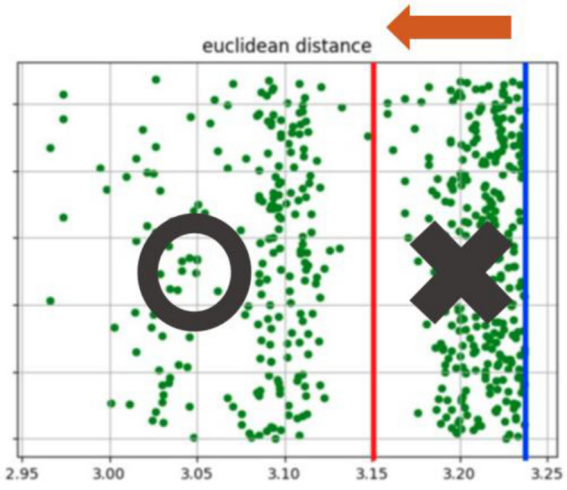

Fig. 5. Green dots are negative samples near the threshold. The blue line (3.23) is the original threshold and the red line (3.15) is our adjusted threshold. The number of data between the blue and red lines is 193. We deleted these data from our negative dataset. X axis indicates the Euclidean distance between miRNA.

IV. Experiments

In this section, the dataset, experimental settings, and performance of the proposed model are described.

A. Dataset and Adjustment

In this study, we utilized the SG-LSTM-core dataset consisting of 31,080 gene-miRNA pairs (383 miRNAs and 318 genes). The dataset consisted of 15,540 pairs of relationships validated using mirTarBase [9] and 15,540 generated negative samples. Every row of the dataset included 128 dimensions of miRNA embedding and 128 dimensions of genes. Each embedding merges the sequence and geometric features. Additionally, a label of 1 or 0 was used to indicate whether a pair of miRNA and gene was related.

We adjusted the threshold created by the average distance (miRNA-gene) of the positive samples to obtain a more sophisticated negative label by observing the distribution of the data. The adjusted threshold is shown in Fig. 4.

Using this adjustment, 193 data points were deleted from the negative set, further improving the prediction performance of our model. We can infer that the deleted data can interfere with the learning of the model because there was no significant difference between the distance of the positive samples and the vector distance. We assessed data distribution differences solely using Euclidean distance. Despite experimenting with varying the threshold for cosine similarity distance, this adjustment did not enhance model performance.

B. Experimental settings

All the experiments were conducted on an Intel(R) Core(TM) i7-7700 CPU @ 3.60 GHz, 32 GB RAM, and GeForce GTX 1080 Ti GPU.

Autoencoder. As previously mentioned, we input label-adjusted data into our autoencoder to extract the data features [10]. Subsequently, batch normalization was applied to each layer of the encoder and decoder, and a Leaky Relu was used as the activation function for each layer. The Adam optimization function was used for optimization and the mean squared error (MSE) was used for the loss function. The epoch size was 200, and the batch size was 64. These were the optimal parameters used in our experiments. The ratio of the training data to the validation data was 8:2, and the training loss and validation loss were 0.0029 and 0.0017, respectively (Fig. 6).

The miRNA-gene features were extracted using a well-trained autoencoder. The size of the extracted feature dimension was 16 and it was concatenated with the 256-dimensional original data. Through several experiments, we confirmed

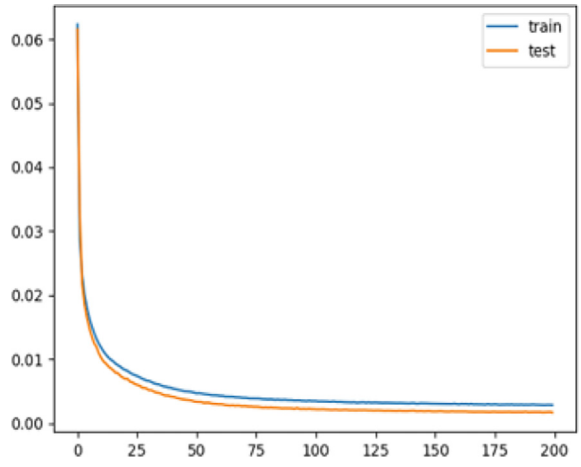

Fig. 6. Train mean squared error (MSE) loss and test MSE loss of the autoencoder (X_axis: epoch/ Y_axis: MSE loss).

that a 16-dimensional feature vector is the best size for improving the performance of our model. Thus, a total of 272 data dimensions were obtained.

LSTM. After concatenation, 30,658 272-dimensional data were used as inputs to the LSTM model, and three LSTM layer were designated as 3 [11]. For the loss function, cross-entropy was used to calculate the degree of loss by converting the predicted value to a value between 0 and 1. Cross-entropy is primarily used for classification problems. The Adam optimizer was used for optimization, which was characterized by a lack of effect of gradient scaling on the step size during training. The epoch size was 200 and the batch size was set to 128. The ratio of the training to test data was 8:2. The performance of the proposed model was tested using a 5-fold cross validation. Additionally, the LSTM model could predict the association score between a gene and an miRNA.

C. Prediction Results

The performance of the model was expressed as AUC, which best represents the performance of the deep learning model. For the model validation test, five model performance results (0.9484, 0.9609, 0.9586, 0.9494, and 0.9565) were obtained through the 5-fold cross validation (Fig. 7). The average value was 0.9548 and the best AUC was 0.9601. Fig. 8 shows the ROC curve of the results of the proposed method. We also present additional performance

AUC – Model 1	AUC – Model 2	AUC – Model 3	AUC – Model 4	AUC – Model 5	Avg. AUC
0.9565	**0.9609**	0.9494	0.9586	0.9484	0.9548

Fig. 7. Five-fold cross validation AUC results of our method.

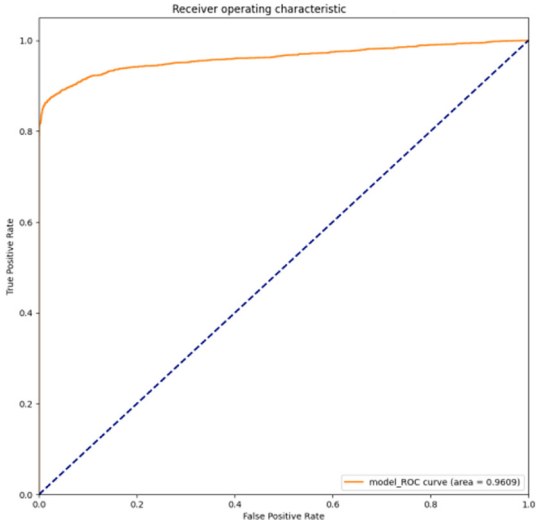

Fig. 8. ROC curve of the best result of our method (AUC = 0.9609).

Precision	Recall	Accuracy	F1-score	AUC
0.938	0.896	0.92	0.92	0.9609

Fig. 9. Classification performance metrics for our model. (Precision, Recall, Accuracy, F1-score, AUC)

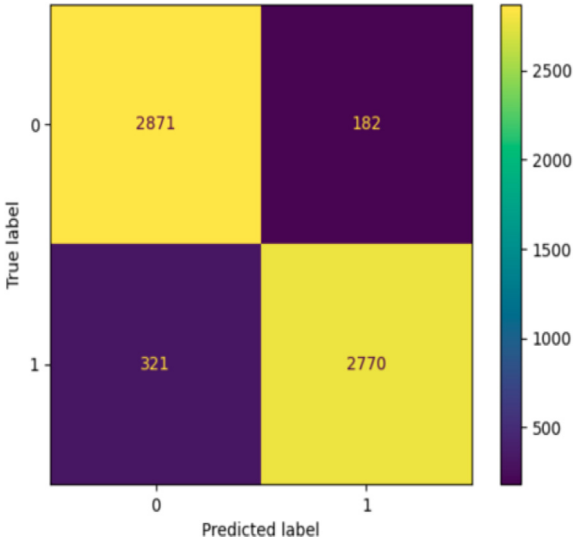

Fig. 10. Confusion matrix of Our model.

metrics, including precision, recall, accuracy, and F1-score, as shown in Fig. 9. We present a confusion matrix to demonstrate the classification performance of our model (Fig. 10).

The performance of a previously reported state-of-the-art (SOTA) model for miRNA-gene association prediction [5] was 0.94 (best AUC) was lower than that of the proposed method.

V. DISCUSSION AND CONCLUSIONS

In this study, we adjusted the negative label through sophisticated threshold adjustments and developed an autoencoder to extract features and an LSTM deep learning model to predict the relationship between miRNAs and genes. The proposed model exhibited better performance (AUC = 0.9609) than the previously reported SOTA model (AUC = 0.94), confirming that our method exhibited the best miRNA-gene association prediction performance. The good performance of the proposed method can be attributed to three main reasons. First, sophisticated negative labeling provides data with accurate labels for deep learning models. Second, the feature vector extracted by the autoencoder aids in the association prediction performance of the proposed model. Third, the LSTM model with high-quality labeling and extracted features as inputs exhibited improved learning performance. In future studies, we will generate a large miRNA-gene dataset to predict a broader range of miRNA-

gene relationships. In addition, we introduced different distance criteria to build a more sophisticated negative dataset.

ACKNOWLEDGEMENTS

This work was supported by research fund of Chungnam National University.

REFERENCES

[1] L. Fu and Q. Peng, "A deep ensemble model to predict miRNA disease association," *Scientific reports*, vol. 7, no. 1, pp. 1-13, Nov. 2017. DOI: 10.1038/s41598-017-15235-6.

[2] K. Deepthi and A. S. Jereesh, "An Ensemble Approach Based on Multi-Source Information to Predict Drug-MiRNA Associations via Convolutional Neural Networks," *IEEE Access*, vol. 9, pp. 38331-38341, 2021. DOI: 10.1109/access.2021.3063885.

[3] M. Lindow and J. Gorodkin, "Principles and limitations of computational microRNA gene and target finding," *DNA and cell biology*, vol. 26, no. 5, pp. 339-351, May 2007. DOI: 10.1089/dna.2006.0551.

[4] Y. Liu, J. Luo, and P. Ding, "Inferring MicroRNA Targets Based on Restricted Boltzmann Machines," *IEEE J Biomed Health Inform*, pp 427-436, Jan. 2019. DOI: 10.1109/JBHI.2018.2814609.

[5] M. Wen, P. Cong, Z. Zhang, H. Lu, and T. Li, "DeepMirTar: a deep learning approach for predicting human miRNA targets," *Bioinformatics*, vol. 34, no. 22, pp 3781-3787, Nov. 2018. DOI: 10.1093/bioinformatics/bty424.

[6] T. Gu, X. Zhao, W. B. Barbazuk, and J.-H. Lee, "miTAR: a hybrid deep learning-based approach for predicting miRNA targets," *BMC bioinformatics*, vol 22, no. 1, pp 1-16, Feb. 2021. DOI: 10.1186/s12859-021-04026-6

[7] S. Cheng, M. Guo, C. Wang, X. Liu, Y. Liu, and X. Wu, "MiRTDL: a deep learning approach for miRNA target prediction," *IEEE/ACM transactions on computational biology and bioinformatics*, vol. 13, no. 6, pp. 1161-1169, Nov. 2016. DOI: 10.1109/TCBB.2015.2510002.

[8] W. Xie, J. Luo, C. Pan, and Y. Liu, "SG-LSTM-FRAME: a computational frame using sequence and geometrical information via LSTM to predict miRNA-gene associations," *Briefings in bioinformatics*, vol. 22, no. 2, pp. 2032-2042, Mar. 2021. DOI: 10.1093/bib/bbaa022.

[9] S. D. Hsu, F-M. Lin, W. Y. Wu, C. Liang, W-C. Huang, W-L. Chan, W-T. Tsai, G-Z. Chen, C. J. Lee, C-M. Chiu, C. H. Chien, M. C. Wu, C. Y. Huang, A. P. Tsou, and H. D. Huang, "miRTarBase: a database curates experimentally validated microRNA-target interactions," *Nucleic acids research*, vol. 39, no. suppl_1, pp. D163-D169, Jan. 2011. DOI: 10.1093/nar/gkq1107.

[10] Q. Meng, D. Catchpoole, D. Skillicom, and P. J. Kennedy "Relational autoencoder for feature extraction," in *2017 International joint conference on neural networks (IJCNN)*, Anchorage, USA, pp. 364-371, 2017. DOI: 10.1109/IJCNN.2017.7965877.

[11] Y. Yu, X. Si, C. Hu, and J. Zhang, "A review of recurrent neural networks: LSTM cells and network architectures," *Neural computation*, vol. 31, no. 7, pp. 1235-1270, Jul. 2019. DOI: 10.1162/neco_a_01199.

Seung-Won Yoon

He received a bachelor's degree from Chungnam National University's Department of Computer Science Engineering in 2018. Since 2018, he has been pursuing an integrated PhD program at Chungnam National University. His current research interests include deep learning, machine learning, and bioinformatics.

In-Woo Hwang

He received a bachelor's degree in Information and Communication Convergence Engineering & Design in 2021 from Mokwon University. After that, I received my master's degree from Chungnam National University in 2023. And he is currently going to enter Chungnam National University as a Ph.D. His recent interests are deep learning, machine learning, and graph-based deep learning.

Kyu-Chul Lee

He received his B.S., M.S., and Ph.D. degrees in computer engineering at Seoul National University in 1984, 1986, and 1996, respectively. In 1994, he worked as a visiting researcher at the IBM Almaden Research Center, San Jose, California. From 1995 to 1996, he worked as a Visiting Professor at the CASE Center at Syracuse University, Syracuse, New York. He is currently a professor in the Department of Computer Engineering at Chungnam National University, Daejeon, Korea. His current areas include database systems, semantic web, big data processing, and artificial Intelligence. He has published over 100 technical articles in various journals and conferences. He is a member of ACM, the IEEE Computer Society, and the Korea Information Science Society.