Research *Barut&Altuntaş/Kırklareli University Journal of Engineering and Science 9-1 (2023) 63-82*

DOI: 10.34186/klujes.1248062 *Geliş Tarihi:05.02.2023* *Kabul Tarihi:29.06.2023*

COMPARISON OF PERFORMANCE OF DIFFERENT N VALUES WITH N-FOLD CROSS-VALIDATION IN A GRAPH-BASED LEARNING MODEL FOR lncRNA-DISEASE PREDICTION

Zeynep BARUT[1*], Volkan ALTUNTAŞ[2]

[1,2]Bursa Teknik Üniversitesi Mühendislik ve Doğa Bilimleri Fakültesi Bilgisayar Mühendisliği Bölümü, Bursa*

Abstract

In machine learning, the value of n in the n-fold cross validation method significantly affects the performance of the created model. In some cases, increasing n increases the accuracy, while in some cases it only increases the computational cost. That is, the n value represents the amount of data used to increase the accuracy of the model. However, the accuracy of the model may not increase at the same rate as the value of n increases. In this case, the correct selection of the n value is of great importance. In the studies that have been done, the value of n is usually taken as five or ten because these two values are thought to produce average estimates. However, there is no official rule. It has been observed that few studies have been carried out to use different n values in the training of different models. In this study, various n values (2, 3, 4, 5, 6, 7, 8, 9 and 10) and four data sets were used with the VGAELDA model, a model that combines variational inference and graphic autoencoders to determine the relationships between lncRNA and disease. A performance evaluation was performed on the IncRNA-disease model using The obtained results were compared and the most suitable n value for the model was determined. In addition, the missing functions were examined for four data sets and the results were interpreted. In future studies, it is aimed to carry out a more comprehensive study by increasing the number of data sets.

Keywords: Graph Autoencoder, Variational Inference, Representation Learning

Sorumlu Yazar: Zeynep BARUT, zebarut@gmail.com

Research *Barut&Altuntaş/Kırklareli University Journal of Engineering and Science 9-1 (2023) 63-82*

DOI: 10.34186/klujes.1248062 *Geliş Tarihi:05.02.2023* *Kabul Tarihi:29.06.2023*

IncRNA-HASTALIK TAHMİNİ İÇİN GRAPH TABANLI BİR ÖĞRENME MODELİNDE N-FOLD CROSS-VALIDATION İLE FARKLI N DEĞERLERİNİN PERFORMANSININ KARŞILAŞTIRILMASI

Öz

Makine öğrenmesinde, n-katlı çapraz doğrulama yöntemindeki n değeri, oluşturulan modelin performansını önemli ölçüde etkilemektedir. Bazı durumlarda n'nin artması doğruluğu artırırken bazı durumlarda sadece hesaplama maliyetini arttırmaktadır. Yani n değeri, modelin doğruluğunu artırmak için kullanılan veri miktarını temsil eder. Ancak, n değeri arttıkça modelin doğruluğu aynı oranda artmayabilir. Bu durumda n değerinin doğru seçilmesi büyük önem taşır. Yapılmış olan çalışmalarda genellikle n değeri beş veya on alınmaktadır çünkü bu iki değerin ortalama tahminler ürettiği düşünülmektedir. Ancak resmi bir kural yoktur. Farklı modellerin eğitiminde farklı n değerlerinin kullanılması için az sayıda çalışma yapıldığı görülmüştür. Bu çalışmada, lncRNA ve hastalık arasındaki ilişkileri belirlemek için varyasyonel çıkarım ve grafik autoencoder'ları birleştiren bir model olan VGAELDA modeli ile çeşitli n değerleri (2, 3, 4, 5, 6, 7, 8, 9 ve 10) ve dört veri seti kullanılarak IncRNA-hastalık modeli üzerinde bir performans değerlendirilmesi yapılmıştır. Elde edilen sonuçlar karşılaştırılmış ve model için en uygun n değeri belirtilmiştir. Ayrıca kayıp fonksiyonlar dört veri seti için incelenerek sonuçlar yorumlanmıştır. Gelecekte yapılacak olan çalışmalarda veri seti sayısının arttırılması ile daha geniş kapsamlı bir çalışma yapılması hedeflenmektedir.

Anahtar Kelimeler: Graf Otomatik Kodlayıcı, Varyasyonel Çıkarım, Temsil Öğrenimi

1. INTRODUCTION

lncRNA (Long non-coding Ribonucleic acid) is RNA consisting of many nucleotides and non-coding, functioning through biochemical mechanisms. It is linked to many human diseases, as it has various biological tasks, such as the regulation of gene expressions. For example, because it has a tumor suppressive function, it causes the onset of cancer in humans [1]. MiRNA (Micro RNA) is a class of short IncRNA molecules. Although miRNAs are small, they are important regulators of gene expression associated with a variety of cellular processes. For this reason, changes in miRNAs have been associated with a number of diseases such as cancer, epidemics, and immune system-related diseases [2]. Sun et al. [3] found in their study that MEG3(Maternally Expressed Gene 3), a gene that encodes an lncRNA associated with many cancer types, forms gastric cancer cells. According to Faghihi et al. [4] found in their study that lncRNA BACE1 (Beta-site Amyloid precursor protein Cleaving Enzyme 1) causes Alzheimer's. Therefore, the

DOI: 10.34186/klujes.1248062 Geliş Tarihi:05.02.2023 Kabul Tarihi:29.06.2023

relationships between lncRNA and disease should be examined in more detail in order to find solutions to diseases more easily.

For this purpose, machine learning methods can be examined in three approaches. In the first approach, matrix analysis is used. Matrix analysis is divided into manifold editing and matrix completion. In their study, Chen and Yan [5] proposed the LRLSLDA (Laplacian Regularized Least Squares for LncRNA–Disease Association) tool, which applies LRLS (Laplacian Regularized Least Square) by creating graphs to determine the relationships between lncRNA and disease. Lu et al. [6] proposed a matrix completion based method, SIMCLDA (Speedup Inductive Matrix Completion LncRNA–Disease Association), to determine the relationships between lncRNA and disease. In the second approach, features of different nature are combined. Lan et al. [7] in their study, an application was created to determine the disease by combining the characteristics of lncRNA and diseases. In the third approach, the graph autoencoder model was used for representative learning of lncRNA and disease characteristics. Xuan et al. [8] used convolutional and graph neural networks together to determine the relationships between lncRNA and disease. Wu et al. [9] used graph autoencoder to determine the relationships between lncRNA and disease. Tamilarasi and Rani [10] tried to obtain the best n value in the cross validation method with different machine learning methods on crime data. As a result of the study, it was seen that KNN had better performance than other methods trained with the same n value. Jung et al. [11] used artificial neural network-based models to accurately predict nitrate loads in river basins. The accuracy of various n values for the training of artificial neural networks has been investigated. As a result of the study, it was seen that the use of n = 10 had better performance when looking at the overall data sets. In the literature, it seems that n = 10 is more common than other values, but there is no official rule. However, several studies have extensively investigated how different n-fold values affect validation results in various machine learning methods tested with a dataset with available numerical properties [12,13,14]. In this study, the VGAELDA (Variational Graph Autoencoders LncRNA–Disease Association) model, which uses variable inference and graph autoencoder together, was used to determine the relationships between IncRNA and disease. This model is divided into two as Variational Graph Autoencoder and Graph Autoencoder. These autoencoder types are trained with a variable maximization algorithm. These methods increase the

DOI: 10.34186/klujes.1248062 *Geliş Tarihi:05.02.2023* *Kabul Tarihi:29.06.2023*

predictive ability of the created model. In the study, four different data sets were used to find the appropriate n value [15,16, 17]. Other studies in the literature using graph neural network-based computational methods to predict relationships between IncRNA-disease are summarized in Table 1. When the studies were examined, it was seen that the studies in which different n-fold cross validation values were compared for different data sets were not sufficient in the studies performed to predict the relationships between the unknown IncRNA-disease. In this study, n values between 2 and 10 were examined for 4 different data sets, contributing to the literature.

Table 1. Graph neural network-based computational methods [18].

Method	Description	N Value	Source Code
GCNLDA [8]	A novel method based on graph convolution and convolutional neural network	5	Unavailable
GAERF [19]	A computational method based on graph autoencoder and random forest	5	Unavailable
MLGCNET [20]	A framework using multi-layer aggregation graph convolutional network and extra trees	5	https://github.com/ QingwWu/MLGCNET
MGATE [21]	A method using multi channel graph attention autoencoder	5	https://github.com/ sheng-n/MGATE
GANLDA [22]	An end-to-end computational model based on graph attention network	10	Unavailable
GTAN [23]	A novel method based on graph neural network with attribute level attention mechanisms and multilayer convolutional neural networks	5, 10, 20	Unavailable
GAMCLDA [9]	A computational framework based on graph autoencoder matrix completion	10	Unavailable
GCRFLDA [24]	A method using graph convolutional matrix completion with conditional random field and attention mechanism	5	https://github.com/ jademyC1221/ GCRFLDA
HGATLDA [25]	A heterogeneous graph attention network framework based on meta-paths	5	Unavailable

DOI: 10.34186/klujes.1248062 *Geliş Tarihi:05.02.2023* *Kabul Tarihi:29.06.2023*

2. MATERIAL AND METHOD

2.1. Graph Semi Supervised Learning

This method is a type of semi-supervised learning using labeled and unlabeled data. This method tries to extract the label information of the unlabeled data from the data in the graph structure. It is used in many applications where a large number of unlabeled samples are obtained, although there are few labeled samples. The manifold assumption is used for the method. In the manifold assumption, samples located close together on a low-dimensional manifold share similar label. This assumption is used to construct the graph structure. The graph structure creates a graph where the nodes specify the samples and the weighted edges indicate the similarity between the nodes. This method of constructing graphs shows that nodes associated with weighted edges tend to have similar labels, in accordance with the manifold assumption. It tries to find the tags of unlabeled samples by making use of tagged samples. For example, the labels of the nodes associated with the labels of the labeled samples can be similarly predicted. These methods are important because of the abundance of unlabeled data and often reduce the need for data labeling and aid in a better understanding of datasets [26].

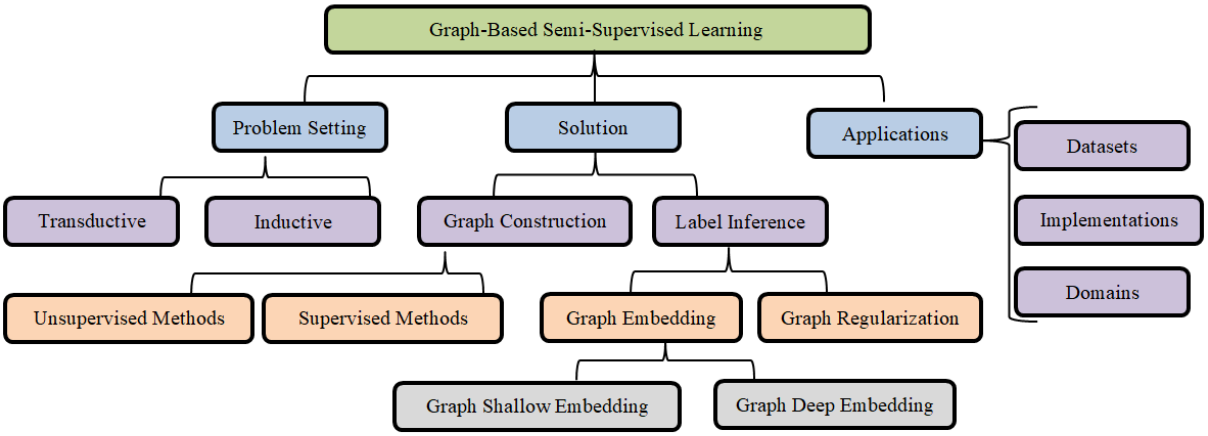

Figure 1. General structure of Graph semi-supervised learning

This method is classified as in Figure 1. A similarity graph is created, and label extraction is done using this graph. Label extraction is difficult to do, it is divided into two as graphic editing methods

DOI: 10.34186/klujes.1248062 Geliş Tarihi:05.02.2023 Kabul Tarihi:29.06.2023

and graphic embedding methods. Graphical editing methods create a framework with the loss function, while graphic embedding methods create a new unified representation with the encoder-decoder framework [26].

2.2. Graph Markov Neural Networks

This method models the distribution of object labels with a field that can be trained using the variational EM (Expectation Maximization) algorithm. It combines statistical relational learning methods and graph neural networks. Useful representations for predicting object tags are learned and dependencies between object tags are modeled. In stage E, the graph neural network learns object representations to approximate distributions of object labels, while in stage M, a different graph neural network is available to model the label dependence. Studies on object classification, connection classification and unsupervised node representation learning have shown that the use of neural network method is advantageous. The general structure of the neural network is given in Figure 2. Yellow and gray squares indicate labeled and unlabeled objects. Graph markov neural networks are trained by switching between stages E and M [27].

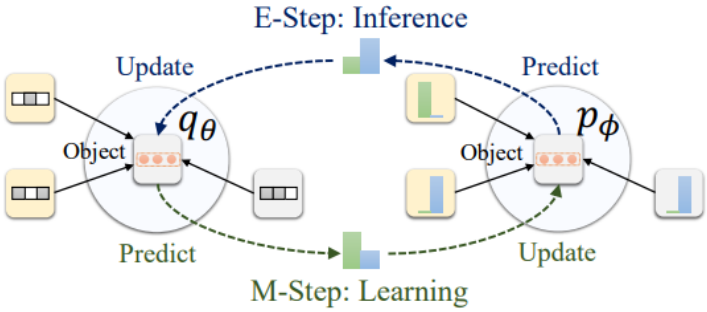

Figure 2. General structure of graph markov neural networks [27].

2.3. Geometric Matrix Completion

Models in this method are commonly used in recommendation systems and have the advantage of storing relationships between users and items with the help of graphs. However, the number of parameters to be learned in these models varies according to the number of users and items. The geometric matrix completion method proposes using geometric deep learning on graphs to

DOI: 10.34186/klujes.1248062 Geliş Tarihi:05.02.2023 Kabul Tarihi:29.06.2023

overcome this limitation. This method uses a multi-graph convolutional neural network that learns the graph structures of the elements, and a recurrent neural network that implements a learnable spread in the matrix. This structure always requires the same number of parameters regardless of the matrix size, so it is convenient in terms of the number of operations that will occur [28].

2.4. Graph Autoencoder

This method is an embedding method that maps graph data to a low-dimensional space and reduces computational cost. It is a neural network that transforms the input data into a representation and reconstructs it from the encoder's output. This neural network structure uses graph neural network as input data. This method is widely used as it shows great potential in size reduction. It consists of two parts, encoder and decoder. Figure 3 shows the model of a graph autoencoder and a graph autoencoder built with a graph neural network. Input data is generated with the encoder. The decoder can regenerate the initial input data. The decoder in the given model is a graph neural network [29].

2.5. N-Fold Cross-Validation

N-fold cross-validation is a method used to evaluate machine learning models. This method divides the dataset into n-folds and in each iteration, one is used as the test set and the remainder as the training set. These steps are repeated until the data set is fully evaluated. This method is used to evaluate the accuracy and general validity of machine learning models. In this way, it can be understood whether the predictions of the models are specific to the data set and more reliable results can be obtained. The general structure of the method is shown in Figure 4 [30].

DOI: 10.34186/klujes.1248062 *Geliş Tarihi:05.02.2023* *Kabul Tarihi:29.06.2023*

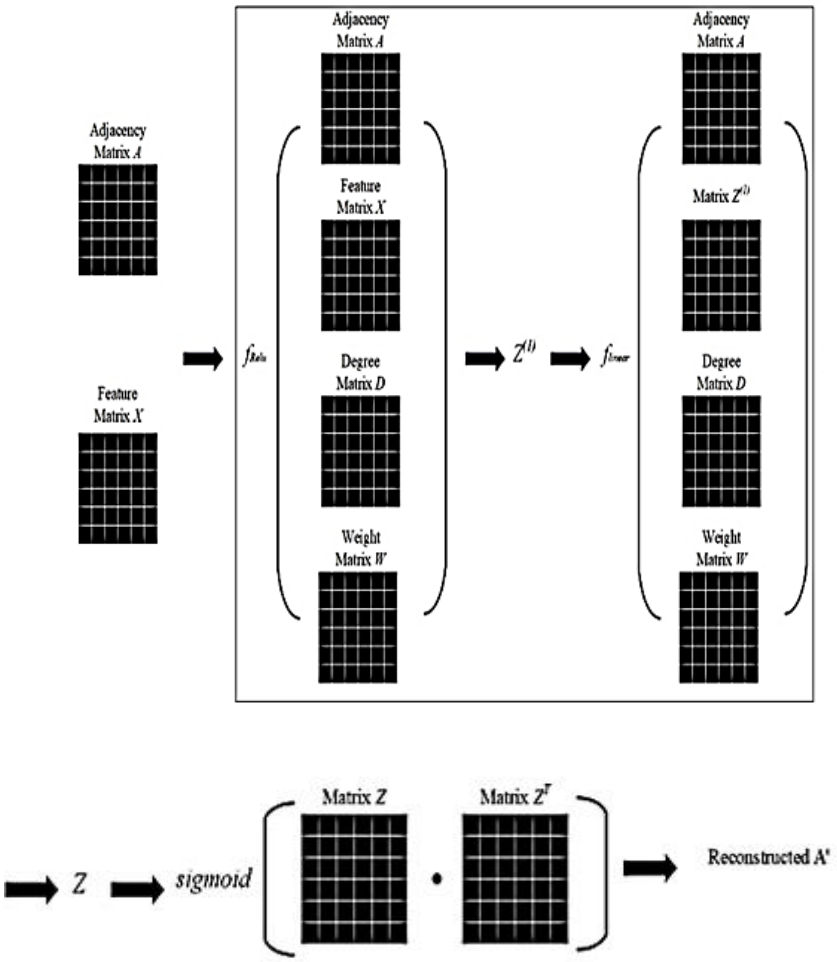

Figure 3. An autoencoder model based on Graph neural network [29].

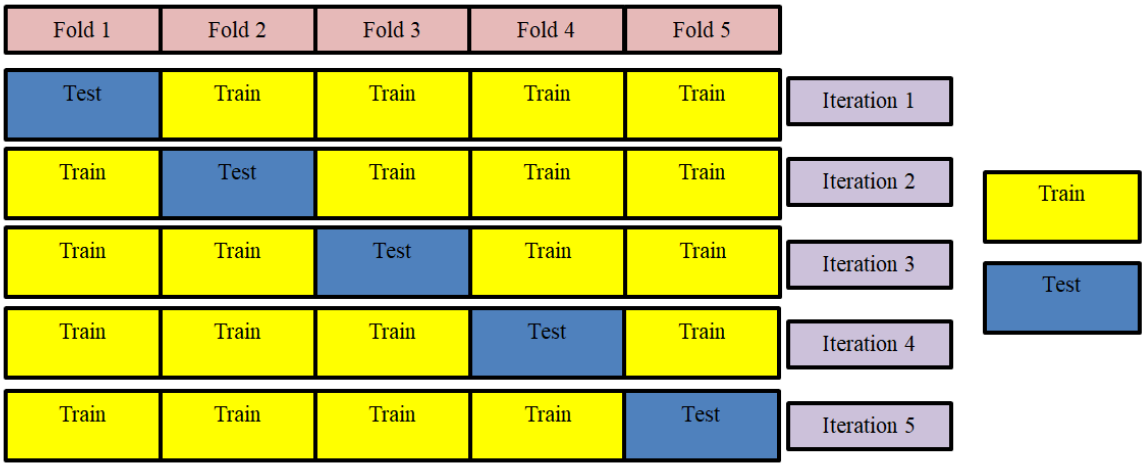

Figure 4. N-fold cross-validation general structure

DOI: 10.34186/klujes.1248062 *Geliş Tarihi:05.02.2023* *Kabul Tarihi:29.06.2023*

2.6. VGAELDA

Graph semi-supervised learning is a learning method that aims to learn the relationships between features in the dataset. This method represents a dataset as a graph and aims to learn the relationships between nodes and edges on the graph. This method may yield better results than other methods for detecting similarities between features because graphs can more directly represent relationships between features in a dataset. VGAELDA is designed to solve the graph semi-supervised learning problem. With this model, a representation learning model was created by combining feature extraction and tag propagation networks and trained with a variational EM algorithm using variable inference and graph Markov neural networks. GNNq (Graph Neural Networks q), is a neural network model that utilizes the structure of a graph to extract features from it, it is called a variational graph autoencoder. On the other hand, GNNp (Graph Neural Networks p), is another neural network model that uses the graph's structure to propagate tags through it, it is referred to as a graph autoencoder. Both GNNq and GNNp are designed to operate on graphs, but they have different functions and purposes. While generating low-dimensional representations in the feature extraction stage, GNNp is determined and GNNq is trained with high-dimensional features. In the tag propagation stage, GNNq is determined, and the IncRNA-disease matrix is given as input to GNNp for training. These stages are performed continuously. The structure of the model used in the study is given in Figure 5 and the algorithm of the model is given in Figure 6. The hidden vector size of the model was determined as 256, the size of the lncRNA embedding vectors was 300, and the epoch value was 500 [16].

The EM algorithm is applied continuously until the GNNq/GNNp losses are minimized. As given in Equation 1, the GNNq loss function (Lossq) is calculated by the reconstruction error Lqr and the KL deviation LKL. IncRNA features have a Gaussian distribution, and the reconstruction error is calculated as given in Equation 2. Disease characteristics have a Bernoulli distribution and are calculated as given in Equation 3. KL deviation loss is calculated as given in Equation 4. GNNp loss function (Lossp) is calculated with reconstruction error and manifold loss as given in Equation 5. The reconstruction error given in Equation 6 is calculated by the cross-entropy of the estimated and actual labels [16].

DOI: 10.34186/klujes.1248062 *Geliş Tarihi:05.02.2023* *Kabul Tarihi:29.06.2023*

$$L_q = L_{qr} + L_{KL} \tag{1}$$

$$L_{qr} = \frac{1}{2}\left\|X - \acute{X}\right\|_F^2 \tag{2}$$

$$L_{qr} = -\sum_{i,j} X_{i,j} \log \acute{X}_{ij} \tag{3}$$

$$L_{KL} = -\sum_{i,j} \frac{1}{2}\left(1 + 2\log\sigma_{ij} - \mu_{ij}^2 - \sigma_{ij}^2\right) \tag{4}$$

$$L_p = L_{pr} + \gamma L_m \tag{5}$$

$$L_{pr} = -\sum_{i,j} Y_{ij} \log F_{ij} \tag{6}$$

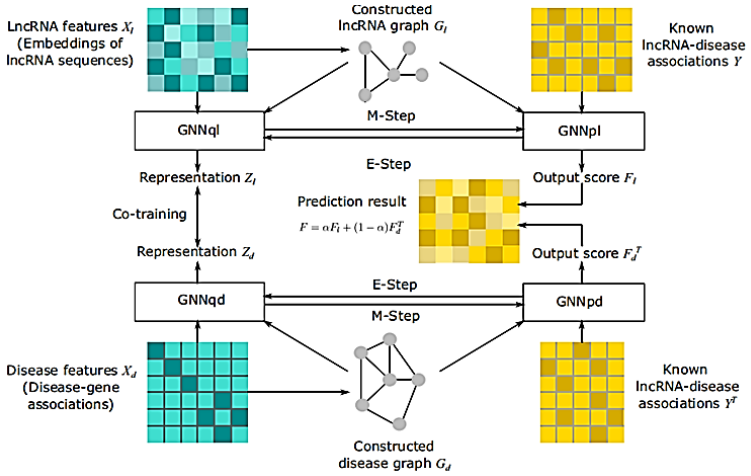

Figure 5. Structure of the VGAELDA model [16].

Algorithm 1 VGAELDA Algorithm

Input: lncRNA features X_l, disease features X_d, initial association matrix Y, parameter α, β, γ
Output: score matrix F

Construct graph G_l and G_d through Eq. (16), from lncRNA features X_l and disease features X_d respectively
repeat
  $X_l', Z_l \leftarrow \mathrm{GNNql}(G_l, X_l)$
  $F_l, Z_l' \leftarrow \mathrm{GNNpl}(G_l, Y)$
  $X_d', Z_d \leftarrow \mathrm{GNNqd}(G_d, X_d)$
  $F_d, Z_d' \leftarrow \mathrm{GNNpd}(G_d, Y^T)$
  Compute L_{ql} and L_{qd} through Eq. (17) respectively
  Compute L_{pl} and L_{pd} through Eq. (21) respectively
  Compute co-training loss L_c through Eq. (24) // train GNNql and GNNqd collaboratively
  $\mathcal{L}_q \leftarrow \alpha L_{ql} + (1-\alpha)L_{qd} + \beta L_c$ // Eq. (25)
  $\mathcal{L}_p \leftarrow \alpha L_{pl} + (1-\alpha)L_{pd}$ // Eq. (26)
  Update the weights of GNNql, GNNpl, GNNqd and GNNpd, by optimizing $\mathcal{L}_q$ and $\mathcal{L}_p$ alternately // train GNNq and GNNp alternately via variational EM algorithm
until Convergence
$F \leftarrow \alpha F_l + (1-\alpha)F_d^T$ // Eq. (28)
return F

Figure 6. The algorithm of the VGAELDA model [16].

DOI: 10.34186/klujes.1248062 Geliş Tarihi:05.02.2023 Kabul Tarihi:29.06.2023

Four different data sets were used in the study. Dataset 1 is an lncRNA-disease association dataset containing 540 associations between 115 lncRNAs and 178 diseases. This dataset was collected from the LncRNADisease Database [31]. Dataset 2 is an lncRNA-disease association dataset containing 2697 associations between 240 lncRNAs and 412 diseases. This data set was also collected from the LncRNADisease Database [31]. Dataset 3 includes 240 lncRNAs, 495 miRNAs, and 412 diseases. This dataset comes from Fu et al.'s study of LncRNA–Disease Association prediction [32]. Dataset 4 is a miRNA-disease association dataset containing 4264 associations between 348 miRNAs and 210 diseases. This dataset was downloaded from the HMDD v3.0 database [33]. The methods described in the study were developed using the Python programming language and implemented in the PyCharm integrated development environment. The choice of n value for cross validation significantly affects the performance of the determined model. Studies suggest choosing values of 5, 10 or 20 because it is generally preferable to use more models for training purposes [34]. For this reason, the value of n was chosen between 2 and 10 for the study.

3. RESULTS AND DISCUSSION

In this study, various n values in cross validation were compared and the results were examined. An attempt has been made to find the appropriate n value that provides better prediction accuracy and AUROC/AUPR. For the study, calculations were made with 2, 3, 4, 5, 6, 7, 8, 9 and 10 n values. While the receiver operating characteristic (ROC) curve is created with TPR (True Positive Rate) and FPR (False Positive Rate), the area under the ROC curve (AUROC) and the area under the Precision-recall curve (AUPR) are the methods used to show the success of the model. A n value with the highest cross validation values for AUROC and AUPR and not too computational complexity is accepted as the most appropriate n value. The results for different n values are given in Table 2. Looking at the majority of datasets, n=10 outperformed all other n-values. In data set 1, it was observed that there was an increase in AUROC from n=2 to n=7, a decrease in n=8 and an increase in n=9,10 values again. A sustained increase was observed for AUPR. In data set 2, it was observed that there was an increase from n=2 to n=5 for AUROC and AUPR, a decrease in n=6 and an increase again in values from n=7 to 10. In data set 3, it was observed that there was an increase from n=2 to n=7 for AUROC and AUPR, a decrease in n= 8 and an increase in n=9,10

DOI: 10.34186/klujes.1248062 Geliş Tarihi:05.02.2023 Kabul Tarihi:29.06.2023

values again. In data set 4, it was observed that there was an increase from n=2 to n=7 for AUROC and AUPR, decreases in n= 8 and 10 values, but increased again in n=9 values. The values of Lossp and Lossq loss functions are given in Table 3. Looking at the missing functions for the four data sets, n=5, 6 and 9 values for Lossp in Dataset 1, n=2, 4 and 10 values for Lossq, n=7, 8 and 9 values for Lossp in Dataset 2 , n=2, 3 and 10 for Lossq, n=4, 6 and 9 for Loss in Dataset 3, n=2, 4 and 10 for Lossq, n=6 for Lossp in Dataset 4, It was seen that it gave low results at 8 and 9 values, and n=2, 3 and 10 values for Lossq. Execution time values are given in Table 4. Looking at the execution time for the four datasets, it is seen that the seconds value increases as the n value increases, except for the n=6 and n=10 values in Dataset 1. Except for the n=8 value in Dataset 2, it was observed that the seconds value increased as the n value increased. Except for the n=10 value in Dataset 3, it was observed that the seconds value increased as the n value increased. In Dataset 4, it was observed that the seconds value increased as the n value increased. ROC and PR curves of Dataset 1, Dataset 2, Dataset 3 and Dataset 4 from n=2 to n=10 are given in Figure 7, Figure 8, Figure 9 and Figure 10, respectively. The blue line represents the ROC curve and the orange line the PR curve. Performance results of datasets with various n values show that n is not generalizable for the VGAELDA model. Therefore, it shows that the performance of a model with different n-fold cross-validation values is determined by many components related to model structure and complexity, and the grade and number of dataset used. Studies show that a better prediction model can be created by increasing the prediction accuracy of the cross-validation algorithm.

DOI: 10.34186/klujes.1248062 *Geliş Tarihi:05.02.2023* *Kabul Tarihi:29.06.2023*

Table 2. AUROC and AUPR results for different n values

	Dataset 1		Dataset 2		Dataset 3		Dataset 4	
N values for Cross Validation	**AUROC**	**AUPR**	**AUROC**	**AUROC**	**AUPR**	**AUPR**	**AUROC**	**AUPR**
N=2	0.7593	0.4494	0.7403	0.4042	0.7138	0.3585	0.7608	0.5770
N=3	0.8890	0.6882	0.8506	0.6523	0.8066	0.5253	0.8483	0.7123
N=4	0.9122	0.7360	0.9211	0.7661	0.8564	0.6021	0.9069	0.8146
N=5	0.9631	0.8449	0,9584	0,8580	0.9229	0.6959	0.9300	0.8487
N=6	0.9664	0.8689	0.9336	0.8283	0.9367	0.7216	0.9350	0.8693
N=7	0.9665	0.8707	0.9729	0.8894	0.9455	0.7511	0.9607	0.9015
N=8	0.9648	0.8747	0.9748	0.8954	0.9326	0.7505	0.9570	0.8961
N=9	0.9758	0.9019	0.9785	0.9021	0.9512	0.7581	**0.9696**	**0.9180**
N=10	**0.9770**	**0.9046**	**0.9898**	**0.9343**	**0.9694**	**0.8111**	0.9634	0.9171

Table 3. Lossp and Lossq results for different n values

	Dataset 1		Dataset 2		Dataset 3		Dataset 4	
N values for Cross Validation	**Lossp**	**Lossq**	**Lossp**	**Lossq**	**Lossp**	**Lossq**	**Lossp**	**Lossq**
N=2	0.2332	**0.0769**	0.2672	**0.0779**	0.2770	**0.0745**	0.2266	0.0158
N=3	0.1706	0.0919	0.2469	0.0903	0.2648	0.0904	0.2178	**0.0134**
N=4	0.1802	0.0873	0.2483	0.1033	**0.2349**	0.0850	0.2107	0.0313
N=5	0.1690	0.1015	0.2296	0.1097	0.2642	0.1015	0.2110	0.0288
N=6	**0.1586**	0.0970	0.2088	0.1242	0.2370	0.0925	0.2074	0.0365
N=7	0.1755	0.1058	**0.1951**	0.1311	0.2588	0.1014	0.2078	0.0427
N=8	0.1703	0.1084	0.1953	0.1413	0.2591	0.1043	**0.1926**	0.0614
N=9	0.1637	0.0999	0.1957	0.1502	0.2543	0.0971	0.2009	0.0556
N=10	0.1883	0.0785	0.2211	0.0809	0.2906	0.0793	0.2414	0.0213

Research *Barut&Altuntaş/Kırklareli University Journal of Engineering and Science 9-1 (2023) 63-82*
DOI: 10.34186/klujes.1248062 *Geliş Tarihi:05.02.2023* *Kabul Tarihi:29.06.2023*

Table 4. Execution time results for different n values

	Dataset 1	Dataset 2	Dataset 3	Dataset 4
N values for Cross Validation	**Execution Time (seconds)**	**Execution Time (seconds)**	**Execution Time (seconds)**	**Execution Time (seconds)**
N=2	55.6821	346.2108	61.0048	60.6895
N=3	84.6763	533.7133	102.6874	90.3744
N=4	117.1502	754.4565	124.5229	127.4893
N=5	179.1467	965.4385	146.4717	160.9431
N=6	176.5914	1395.0183	205.6096	207.4925
N=7	208.4421	1488.4724	224.7853	213.0182
N=8	238.8682	1448.2642	269.3764	376.0928
N=9	583.8292	1564.2915	296.0036	420.0812
N=10	420.2618	1875.5056	282.5071	500.3574

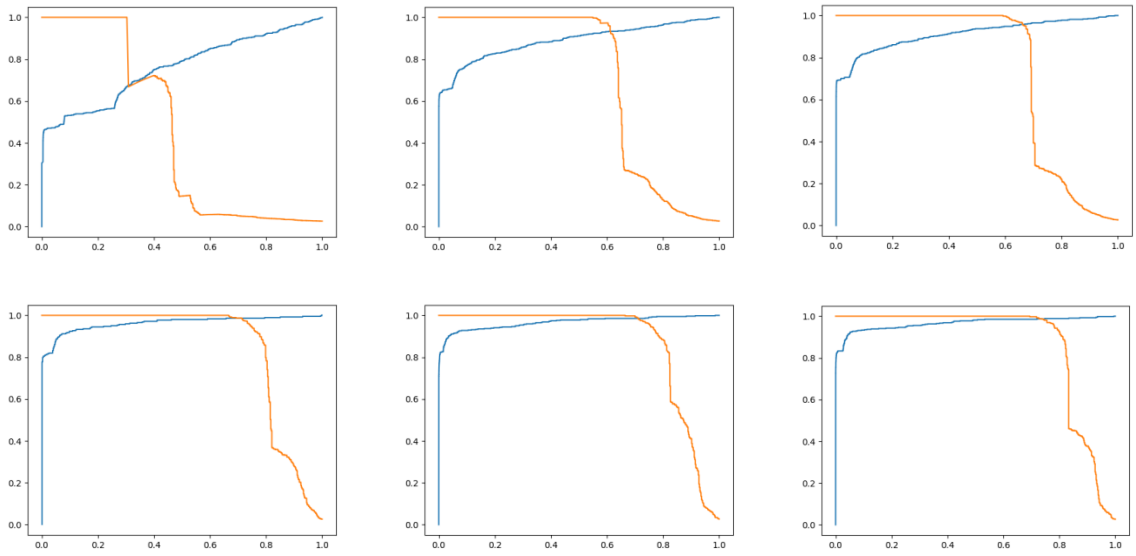

Figure 7. Plot graph results of different n values for Dataset 1 *(continuing)*

Research *Barut&Altuntaş/Kırklareli University Journal of Engineering and Science 9-1 (2023) 63-82*
DOI: 10.34186/klujes.1248062 *Geliş Tarihi:05.02.2023* *Kabul Tarihi:29.06.2023*

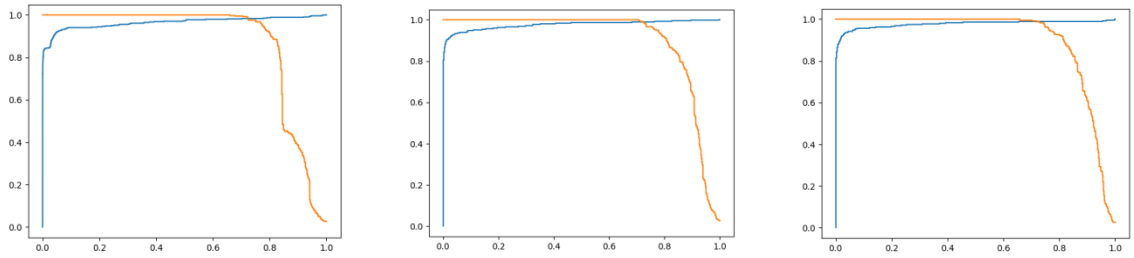

Figure 7. Plot graph results of different n values for Dataset 1

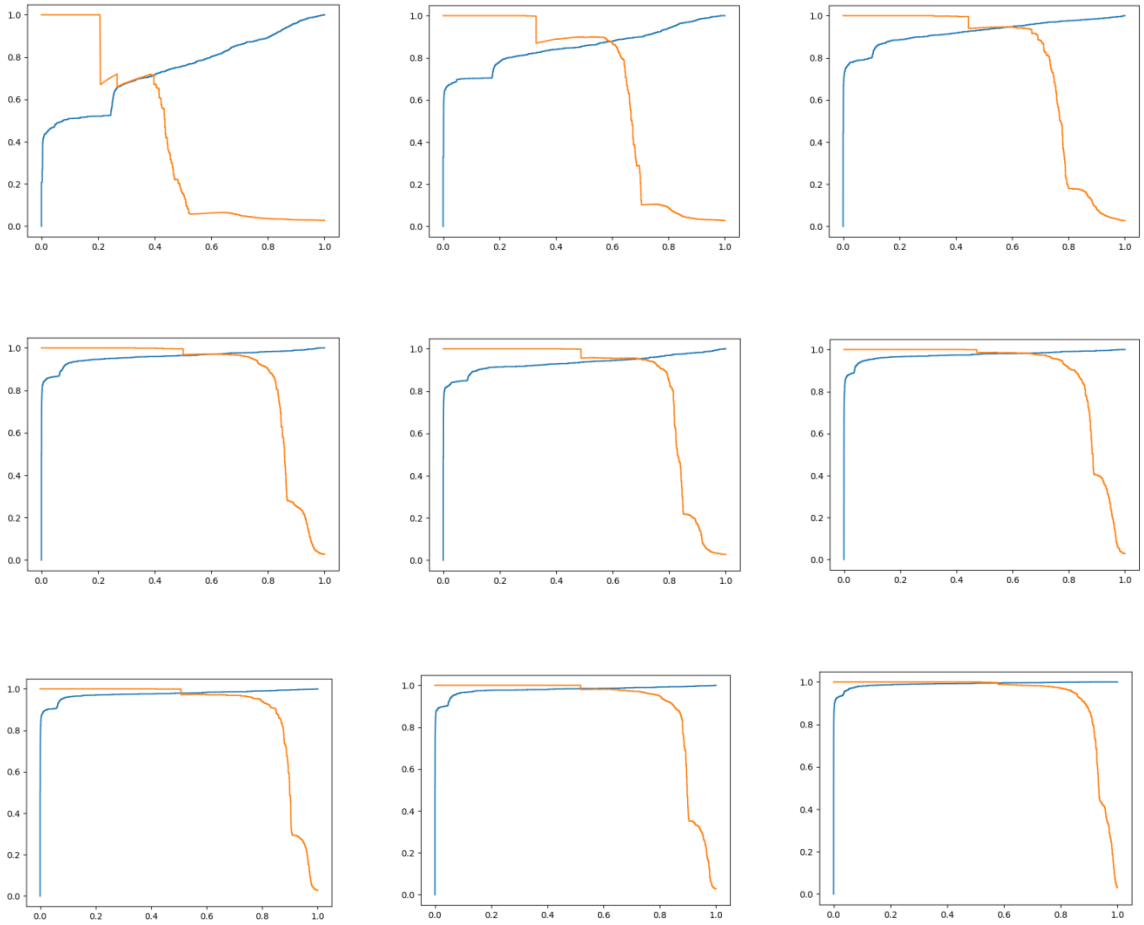

Figure 8. Plot graph results of different n values for Dataset 2

DOI: 10.34186/klujes.1248062 *Geliş Tarihi:05.02.2023* *Kabul Tarihi:29.06.2023*

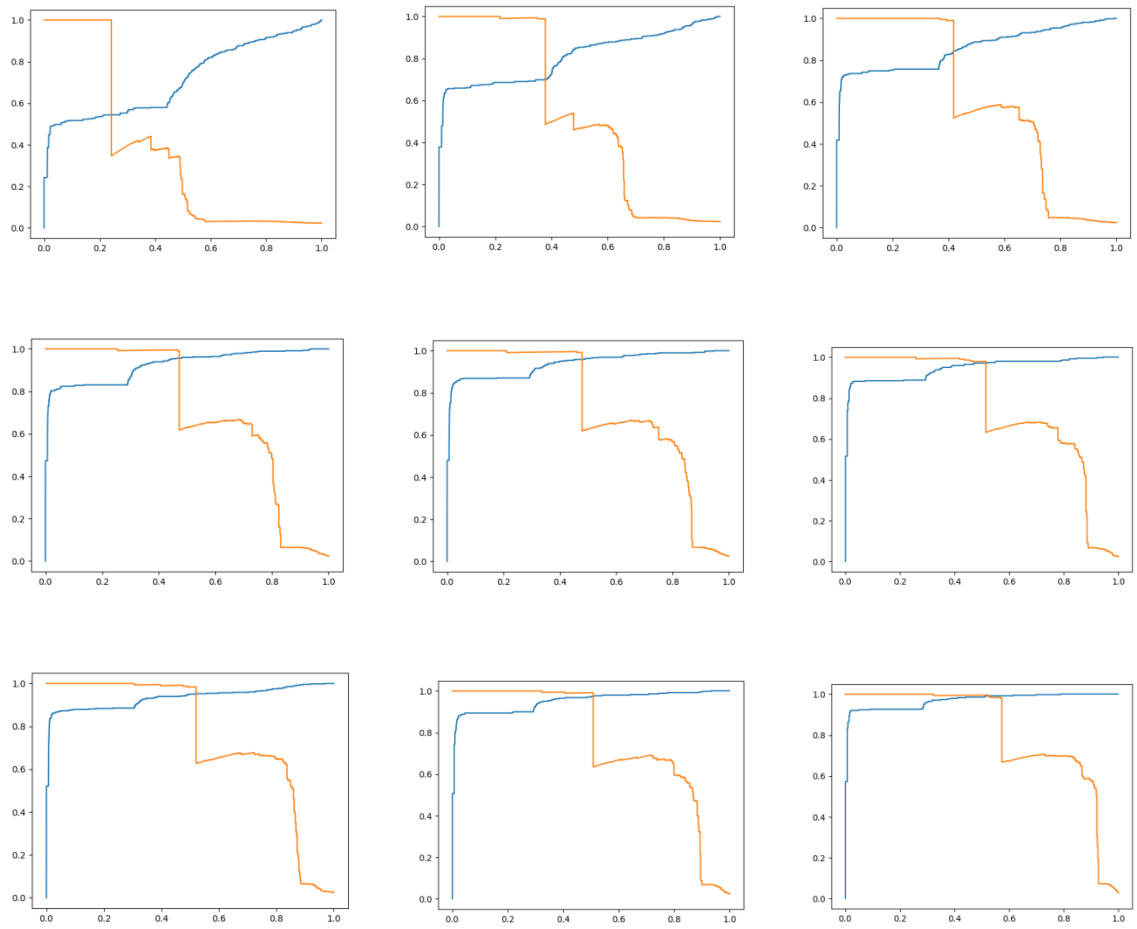

Figure 9. Plot graph results of different n values for Dataset 3

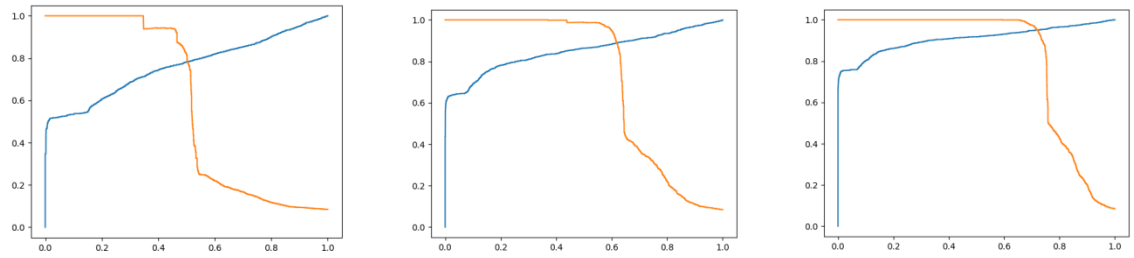

Figure 10. Plot graph results of different n values for Dataset 4 *(continuing)*

DOI: 10.34186/klujes.1248062 Geliş Tarihi:05.02.2023 Kabul Tarihi:29.06.2023

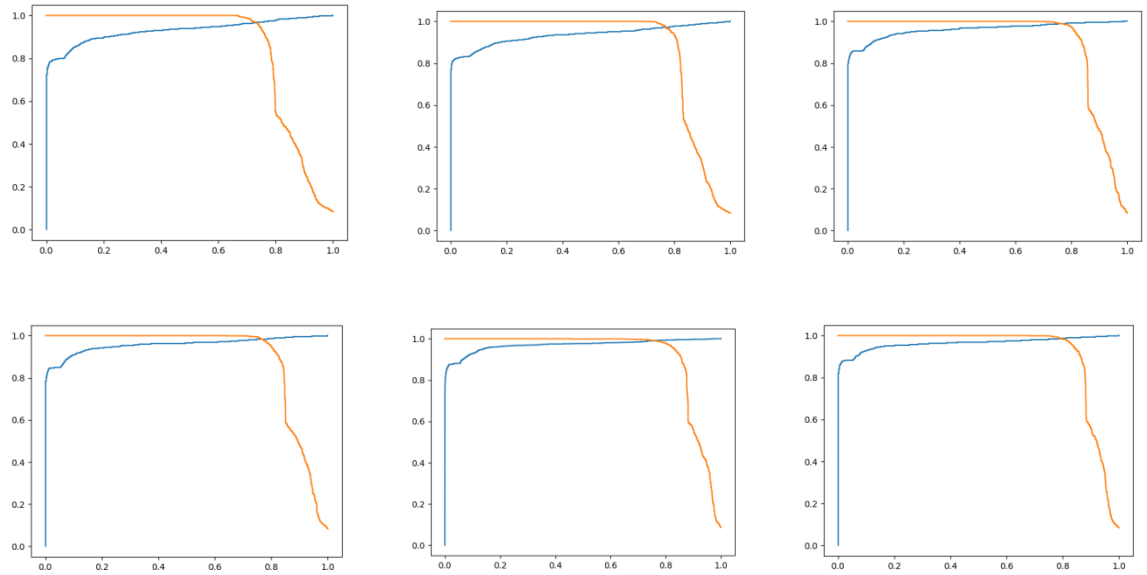

Figure 10. Plot graph results of different n values for Dataset 4

4. CONCLUSION

In this study, the VGAELDA model, a model that combines variational inference and graphic autoencoders, was used to determine the relationships between lncRNA and disease. In the study, the performance of various n values (2, 3, 4, 5, 6, 7, 8, 9 and 10) in n-fold cross-validation on different data sets was investigated. The performance values of the model for the same classification task differ from one dataset to another. In the study, when the majority of the data sets are examined, the n=10 value outperformed all other n values. When the missing functions are examined for the four data sets, it is seen that they generally give low results for the values of n=2, 8 and 10. In addition, when the data sets are examined in general, it is seen that the execution time increases as the n value increases. In some cases, an increase in n increases the accuracy, while in some cases it only increases the computational cost. When the n value is increased, the accuracy of the model does not increase at the same rate. Because of these situations, choosing the n value is important because a small n value has little variance, is easy to calculate, and has high bias. But a large n value is difficult in terms of complexity, has high variance, low bias. Therefore, the size of each data set must be appropriate for the n-value to provide an accurate estimate of the model's

DOI: 10.34186/klujes.1248062 Geliş Tarihi:05.02.2023 Kabul Tarihi:29.06.2023

performance. In order to find the most appropriate n value in improving the accuracy, it would be appropriate to work with various n values on a specific data set and model.

CONFLICT OF INTEREST

The authors declare no conflict of interest.

REFERENCES

[1] Coşan, D.T., Yağcı, E., Kurt, H., Epigenetikten Kansere Uzanan Çizgiler: Uzun Kodlamayan RNA'lar. Osmangazi Journal of Medicine, 40(3), S 114-121, 2018.

[2] Karaarslan, Z. Ö., Serin, M. S., Hastalıkların tanı ve tedavi stratejilerinde miRNA ve diğer non-protein-coding RNA'lar. Mersin Üniversitesi Sağlık Bilimleri Dergisi, 9(3), S 159-172, 2016.

[3] Sun, M., Xia, R., Jin, F., Xu, T., Liu, Z., De, W., Liu, X., Downregulated long noncoding RNA MEG3 is associated with poor prognosis and promotes cell proliferation in gastric cancer. Tumor Biology, 35(2), S 1065-1073, 2014.

[4] Faghihi, M.A., Modarresi, F., Khalil, A.M., Wood, D.E., ahagan, B.G., Morgan, T.E., Finch, C.E., Laurent, G., Kenny, P.J., Wahlestedt, C., Expression of a noncoding RNA is elevated in Alzheimer's disease and drives rapid feed-forward regulation of β-secretase. Nature medicine, 14(7), S 723-730, 2008.

[5] Chen, X., Yan, G. Y., Novel human lncRNA–disease association inference based on lncRNA expression profiles. Bioinformatics, 29(20), S 2617-2624, 2013.

[6] Lu, C., Yang, M., Luo, F., Wu, F.X., Li, M., Pan, Y., Li, Y., Wang, J., Prediction of lncRNA–disease associations based on inductive matrix completion. Bioinformatics, 34(19), S 3357-3364, 2018.

[7] Lan, W., Li, M., Zhao, K., Liu, J., Wu, F. X., Pan, Y., Wang, J., LDAP: a web server for lncRNA-disease association prediction. Bioinformatics, 33(3), S 458-460, 2017.

[8] Xuan, P., Pan, S., Zhang, T., Liu, Y., Sun, H., Graph convolutional network and convolutional neural network based method for predicting lncRNA-disease associations. Cells, 8(9), 1012, 2019.

[9] Wu, X., Lan, W., Chen, Q., Dong, Y., Liu, J., & Peng, W., Inferring LncRNA-disease associations based on graph autoencoder matrix completion. Computational Biology and Chemistry, 87, 107282, 2020.

[10] Tamilarasi, P., Rani, R., Diagnosis of crime rate against women using k-fold cross validation through machine learning. In 2020 Fourth International Conference on Computing Methodologies and Communication (ICCMC), S 1034-1038, 2020.

DOI: 10.34186/klujes.1248062 Geliş Tarihi:05.02.2023 Kabul Tarihi:29.06.2023

[11] Jung, K., Bae, D. H., Um, M. J., Kim, S., Jeon, S., Park, D., Evaluation of nitrate load estimations using neural networks and canonical correlation analysis with k-fold cross-validation. Sustainability, 12(1), 400, 2020.

[12] Fang, L., Liu, S., Huang, Z., Uncertain Johnson–Schumacher growth model with imprecise observations and k-fold cross-validation test. Soft Computing, 24(4), S 2715-2720, 2020.

[13] Wayahdi, M. R., Syahputra, D., Ginting, S. H. N., Evaluation of the K-Nearest Neighbor Model With K-Fold Cross Validation on Image Classification. INFOKUM, 9(1), S 1-6, 2020.

[14] Marcot, B. G., Hanea, A. M., What is an optimal value of k in k-fold cross-validation in discrete Bayesian network analysis?, Computational Statistics, 36(3), S 2009-2031, 2021.

[15] Yao, D., Zhan, X., Zhan, X., Kwoh, C. K., Li, P., Wang, J., A random forest based computational model for predicting novel lncRNA-disease associations. BMC bioinformatics, 21(1), S 1-18, 2020.

[16] Shi, Z., Zhang, H., Jin, C., Quan, X., & Yin, Y., A representation learning model based on variational inference and graph autoencoder for predicting lncRNA-disease associations. BMC bioinformatics, 22(1), S 1-20, 2021.

[17] Zhang, G., Li, M., Deng, H., Xu, X., Liu, X., Zhang, W., SGNNMD: signed graph neural network for predicting deregulation types of miRNA-disease associations. Briefings in Bioinformatics, 23(1), 2022.

[18] Sheng, N., Huang, L., Lu, Y., Wang, H., Yang, L., Gao, L., Xie, X., Fu, Y., Wang, Y., Data resources and computational methods for lncRNA-disease association prediction. Computers in Biology and Medicine, 2023.

[19] Wu, Q.-W., Xia, J.-F., Ni, J.-C., Zheng, C-H., GAERF: predicting lncRNA-disease associations by graph auto-encoder and random forest. Briefings Bioinf, 22(5), 2021.

[20] Wu, Q. W., Cao, R. F., Xia, J. F., Ni, J. C., Zheng, C. H., Su, Y. S., Extra Trees Method for Predicting LncRNA-Disease Association Based On Multi-Layer Graph Embedding Aggregation. IEEE/ACM transactions on computational biology and bioinformatics, 19(6), S 3171–3178, 2022.

[21] Sheng, N., Huang, L., Wang, Y., Zhao, J., Xuan, P., Gao, L., Cao, Y., Multi-channel graph attention autoencoders for disease-related lncRNAs prediction. Briefings in bioinformatics, 23(2), 2022.

[22] Lan, W., Wu, X., Chen, Q., Peng, W., Wang, J., Chen, Y.-P., GANLDA: graph attention network for lncRNAdisease associations prediction. Neurocomputing, 469, S 384–393, 2022.

[23] Xuan, P., Zhan, L., Cui, H., Zhang, T., Nakaguchi, T., Zhang, W., Graph triple-attention network for disease-related LncRNA prediction. IEEE journal of biomedical and health informatics, 26(6), S 2839–2849.

DOI: 10.34186/klujes.1248062 Geliş Tarihi:05.02.2023 Kabul Tarihi:29.06.2023

[24] Fan, Y., Chen, M., Pan, X., GCRFLDA: scoring lncRNA-disease associations using graph convolution matrix completion with conditional random field. Briefings in bioinformatics, 23(1), 2021.

[25] Zhao, X., Zhao, X., Yin, M., Heterogeneous graph attention network based on metapaths for lncRNA-disease association prediction. Briefings in bioinformatics, 23(1), 2021.

[26] Song, Z., Yang, X., Xu, Z., & King, I., Graph-based semi-supervised learning: A comprehensive review. IEEE Transactions on Neural Networks and Learning Systems, S 21, 2022.

[27] Qu, M., Bengio, Y., Tang, J., Gmnn: Graph markov neural networks, In International conference on machine learning, Long Beach, California, PMLR 97, S 5241-5250, 2019.

[28] Monti, F., Bronstein, M., Bresson, X., Geometric matrix completion with recurrent multi-graph neural networks. Advances in neural information processing systems, 30, 2017.

[29] Wang, Y., Xu, B., Kwak, M., Zeng, X., A simple training strategy for graph autoencoder. In Proceedings of the 2020 12th International Conference on Machine Learning and Computing, S 341-345, 2020.

[30] Nti, I. K., Nyarko-Boateng, O., Aning, J., Performance of Machine Learning Algorithms with Different K Values in K-fold Cross-Validation. Inter. J. Info. Technol. Comp. Sci., 13, S 61-71, 2021.

[31] Chen, G., Wang, Z., Wang, D., Qiu, C., Liu, M., Chen, X., Zhang, Q., Yan, G., Cui, Q., LncRNADisease: a database for long-non-coding RNA-associated diseases. Nucleic acids research, 41(Database issue), S D983–D986, 2013.

[32] Fu, G., Wang, J., Domeniconi, C., Yu, G., Matrix factorization-based data fusion for the prediction of lncRNA–disease associations. Bioinformatics, 34(9), S 1529-1537, 2018.

[33] Huang, Z., Shi, J., Gao, Y., Cui, C., Zhang, S., Li, J., Zhou, Y., Cui, Q., HMDD v3.0: a database for experimentally supported human microRNA-disease associations. Nucleic acids research, 47(D1), S 1013–1017, 2019.

[34] Anguita, D., Ghelardoni, L., Ghio, A., Oneto, L., Ridella, S., The'K'in K-fold Cross Validation. In ESANN, S 441-446, 2012.