*Research Article*

# Customer Churn Modeling via the Grey Wolf Optimizer and Ensemble Neural Networks

**Maryam Rahmaty** [ID],[1] **Amir Daneshvar,**[2] **Fariba Salahi,**[3] **Maryam Ebrahimi,**[4] **and Adel Pourghader Chobar** [ID][5]

[1]*Department of Management, Chalous Branch, Islamic Azad University, Chalous, Iran*
[2]*Department of Information Technology Management, Electronic Branch, Islamic Azad University, Tehran, Iran*
[3]*Department of Industrial Management, Electronic Branch, Islamic Azad University, Tehran, Iran*
[4]*Department of Information Technology Management, Islamic Azad University-Electronic Branch, Tehran, Iran*
[5]*Department of Industrial Engineering, Faculty of Industrial and Mechanical Engineering, Qazvin Azad University, Qazvin, Iran*

Correspondence should be addressed to Maryam Rahmaty; rahmaty.maryam61@iauc.ac.ir

The customer churn is one of the key challenges for enterprises, and market saturation and increased competition to maintain business position has caused companies to make all attempts to identify customers who are likely to leave and end their relationship with a company in a particular period to become the customer of another company. In recent years, many methods have been developed including data mining for predicting the customer churn and manners that customers are likely to behave in the future and therefore, taking action early to prevent their leaving. This study proposes a hybrid system based on fuzzy entropy criterion selection algorithm with similar classifiers, grey wolf optimization algorithm, and artificial neural network to predict the customer churn of those companies that suffer losses from losing customers over time. The research results are evaluated by other methods in the criteria of accuracy, recall, precision, and F_measure, and it is declared that the proposed method is superior over other methods.

## 1. Introduction

The expansion of business and commerce environment and the advent of various communication platforms such as the Internet has made it an easy experience for customers to be informed of similar services and products offered by different companies in the shortest time. Accordingly, in the event of dissatisfaction, customers can switch to another company, which leads to the rise of customer churn of the prior company [1]. Managers in all types and sizes of organizations in recent years with the help of researchers in diverse fields have used different methods to predict and prevent the loss of their customers. One of the efficient methods that have been considered by researchers beside the expansion of artificial intelligence to predict customer churn is the use of data mining methods including artificial neural

network methods, decision tree, logistic regression, support vector machine, and random forest [2].

Moreover, Amin et al. [3] showed that if companies can decrease the amount of customer losses by almost 5%, they can increase profits by 25 to 95%. Some of the most prominent problems that customer losses cause for companies are as follows [4].

(i) Financial losses due to the decline of profits from lost customers, as well as the high cost of new customers' absorption.

(ii) Negative effects on current customers and these may cause them to leave the company.

(iii) Providing the opportunity for competing companies to expand their business by attracting those customers.

Thus, the main objective of the present study is to provide a hybrid method for predicting customer churn with high accuracy based on fuzzy entropy measures with similarity classifier, grey wolf optimizer (gwo), and ensemble neural network. Thus, the main questions of this research are as follows:

(i) Is it possible to provide a hybrid method based on fuzzy entropy measures with similarity classifier, grey wolf optimizer, and ensemble neural network, which can predict customer churn with high accuracy?

(ii) To what extent can the use of ensemble neural networks instead of conventional neural networks be effective in improving customer predictability?

In the following, in Section 2, the research background is examined. Section 3 describes the research method. In Section 4, the results are numbered and analyzed, and finally in Section 5 conclusions are made.

## 2. Literature Review

In recent years, the issue of predicting and preventing customer losses has become very important for organizations, and researchers have used various methods to develop systems. De Bock and Van den Poel [4] reviewed the application and evaluation of the performance of the hybrid classification method based on the generalized additive model in predicting customer churn. To compare and evaluate the efficiency of the hybrid model based on the generalized additive model, the bagging method, random forest algorithm, random sub-space algorithm, and logistic regression method have been used. Chen et al. [5] proposed a new framework for predicting customer churn using the longitudinal data method. They applied support vector machine methods, boosting, neural networks, decision tree, random forest, logistic regression, and proportional hazard model. The dataset used in their study was from a food and telecommunication company. Sharma et al. [6] proposed a method for predicting customer churn using multilayer perceptron neural networks. The dataset used was from a U.S. mobile telecommunication company. The results indicate that neural network can be used as an efficient method in predicting customer churn, and telecommunication companies can use the proposed method for this aim. Coussement and De Bock [7] presented a comparison between single data mining methods and hybrid methods, which are actually a combination of several data mining methods in predicting customer churn. They compared methods such as generalized additive model, random forest algorithm, decision tree, and the results showed the superiority of hybrid methods. The data used in this study is related to 3729 customers of the online betting company. Tang et al. [8] examined the impact of using derived behavior information on customer attrition in the financial services industry. They used orthogonal polynomial analysis method to obtain the derived information. Proportional hazards model has also been used to predict customer attrition. The dataset used in this article is related to the

customers of a bank. Keramati et al. [9] proposed a hybrid system based on feature selection methods, artificial neural networks, support vector machine, decision tree, and K-nearest neighbor to predict customer churn. The data about 3150 customers of one of the telephone operators in Iran that have been collected during 12 months Moeyersoms and Martens [10] used highly correlated customer information features to predict customer churn. In this study, the decision tree classification methods, support vector machine, and logistic regression methods are used to predict customer losses on customer datasets of different energy industries. Coussement et al. [11] evaluated the efficiency of data preprocessing methods on the efficiency of customer churn prediction methods. In this study, decision tree and logistic regression methods, Bayesian networks, support vector machine, and another method called random boosting gradient and bagging have been used to predict customer churn. The data used is related to 30104 customers of a European telecommunication company. Yu et al. [12] proposed a hybrid method based on error propagation neural networks and a particle swarm optimization algorithm to predict customer churn. The reason for using the particle swarm optimization algorithm is to obtain neural network parameters such as weights and bias values so that the data can be well trained in the neural network. The results of this study indicate that the neural network with particle swarm optimization training algorithm offers better performance in predicting customer churn than the neural network method with error replication algorithm.

Salvi et al. [13]; in a study examined the LSTM neural network and used it to predict the future trend of Brent oil prices based on the previous price of Brent oil. In this study, 4 types of errors have been calculated to check the accuracy of the model and errors. The mean absolute error (MAE) and Root Mean Square Error (RMSE) were 1.1962 and 1.9164, respectively.

Moitra et al. [14] attempted to use short-term memory neural network instead of convolutional neural network to predict crude oil price. Results were promising and showed more accurate forecasts for crude oil prices in the coming days, and a hybrid model was presented for forecasting crude oil price that used sophisticated network analysis and LSTM algorithms. The research results showed that the model is more accurate and has more robustness and reliability. Jafarzadeh Ghoushchi et al. [15]; provided an extended approach to the diagnosis of tumour location in breast cancer using deep learning. This study develops a new machine learning approach based on modified deep learning (DL) to diagnose the tumour location in breast cancer. In this study, the data obtained from the databases (BCDRD01) are developed and resized and divided into datasets. A simple architecture is used for the first group of experiments, one of which utilizes a weighted function to counter the class imbalance. The results indicate that convolutional neural networks (CNNs) are an appropriate option for the separation of breast cancer lesions.

Raju et al. [16]; have developed an approach to forecasting demand in the steel industry using group learning. This study aims to introduce a robust framework for

TABLE 1: Paper classification.

| Author | Year of publication | Goal | Model |
|---|---|---|---|
| Raju et al. | 2022 | Developed an approach to forecasting demand in the steel industry | Regression ensemble framework |
| Jafarzadeh et al. | 2021 | Provided an extended approach to the diagnosis of tumour | New machine learning approach |
| Jnr et al. | 2021 | Hybrid ensemble intelligent model for electricity demand forecasting | Swarm intelligence and artificial neural network |
| Moitra et al. | 2020 | Predict crude oil price | Short-term memory neural network |
| Salvi et al. | 2019 | Predict the future trend of brent oil prices | LSTM neural network |
| Yu et al. | 2018 | Proposed a hybrid method to predict customer churn | Neural networks and a particle swarm optimization algorithm |

forecasting demand, including data preprocessing, data transformation and standardization, feature selection, cross-validation, and regression ensemble framework. In order to maximize the determination coefficient () value and reduce the root-mean-square error (RMSE), hyperparameters are set using the grid search method. Using a steel industry dataset, all tests are carried out under identical experimental conditions. In this context, STACK1 (ELM + GBR + XGBR-SVR) and STACK2 (ELM + GBR + XGBR-LASSO) models provided better performance than other models. As it improves the performance of models and reduces the risk of decision-making, the ensemble method can be used to forecast the demand in a steel industry one month ahead.

Ofori-Ntow et al. [17], have developed an hybrid ensemble intelligent model based on wavelet transform, swarm intelligence, and artificial neural network for electricity demand forecasting. In this study, a three-level hybrid ensemble short-term load forecasting method consisting of discrete wavelet transform (DWT), particle swarm optimization (PSO), and radial basis function neural network (RBFNN) is proposed. The DWT is applied to decompose the data to get a well-behaved requisite series for forecasting since the data becomes stable before using PSO. The statistical analysis revealed that the proposed method performed better based on MAPE, MAD, and RMSE emphasizing its great potential. Table 1 shows the literature review.

## 3. Methodology

In this research, a hybrid system based on fuzzy entropy measures with similarity classifier, grey wolf optimizer and ensemble neural networks to predict customer churn. As can be seen from the diagram of the proposed system in Figure 1, the proposed system consists of five main steps, which we will briefly review in detail below.

### 3.1. Collect Customer Data and Pre-Process the Data.
In the proposed method, the neural network machine learning tool is applied to predict the customer churn in the customer's classification stage. In other words, the neural network must learn the data of previous lost and non-lost customers of the company, which has been collected by the company during a certain period of time. According to patterns and

relationships between customers, the customer churn and their loyalty can be predicted. In other words, it can be predicted which customers would probably leave the company in the future and which customers would remain loyal.

After collecting the customer data, preprocessing is performed on the data to prepare the data. The preprocessing step involves several different operations performed to prepare the data. One of the steps taken is to convert text or character data or string to numeric data. Normalization is another method of data preprocessing which is performed on data before it is used by machine learning algorithms to make predictions. Its purpose is to solve problems caused by large differences in different data properties in some computational processes. That is to say, the aim of normalizing the data is to prevent the dominance of numbers in the larger range to numbers in the smaller range [18]. Another method of preprocessing data is to deal with missing data. Samples of customer data containing missing data are removed from the dataset, by reason of not interfering with the neural network predictor modeling process by this kind of data.

### 3.2. Selection of Optimal Features and Removal of Redundant and Additional Features.
The feature selection process in the system is performed using the fuzzy entropy measures with similarity classifier. This feature selection algorithm was proposed by Luukka [19]; in which the fuzzy entropy criterion is used to select the optimal features and remove the redundant features of the data. At this stage, the customer dataset is given to this algorithm, which specifies some properties that can be removed from the data in its output.

Obtain optimal structures using the grey wolf algorithm: The grey wolf algorithm is used to predict the customer churn to obtain two optimal structures for each of the ensemble neural networks. That is, the grey wolf algorithm starts at this stage and during an optimization process identifies two suitable structures or architectures, namely the number of layers and the number of neurons in each layer. One of the most important issues is how to encode the search agents in the grey wolf optimization algorithm.

The grey wolf algorithm proposed by Mirjalili [20] is a continuous algorithm to solve continuous problems, while the problem of obtaining the appropriate structure for the
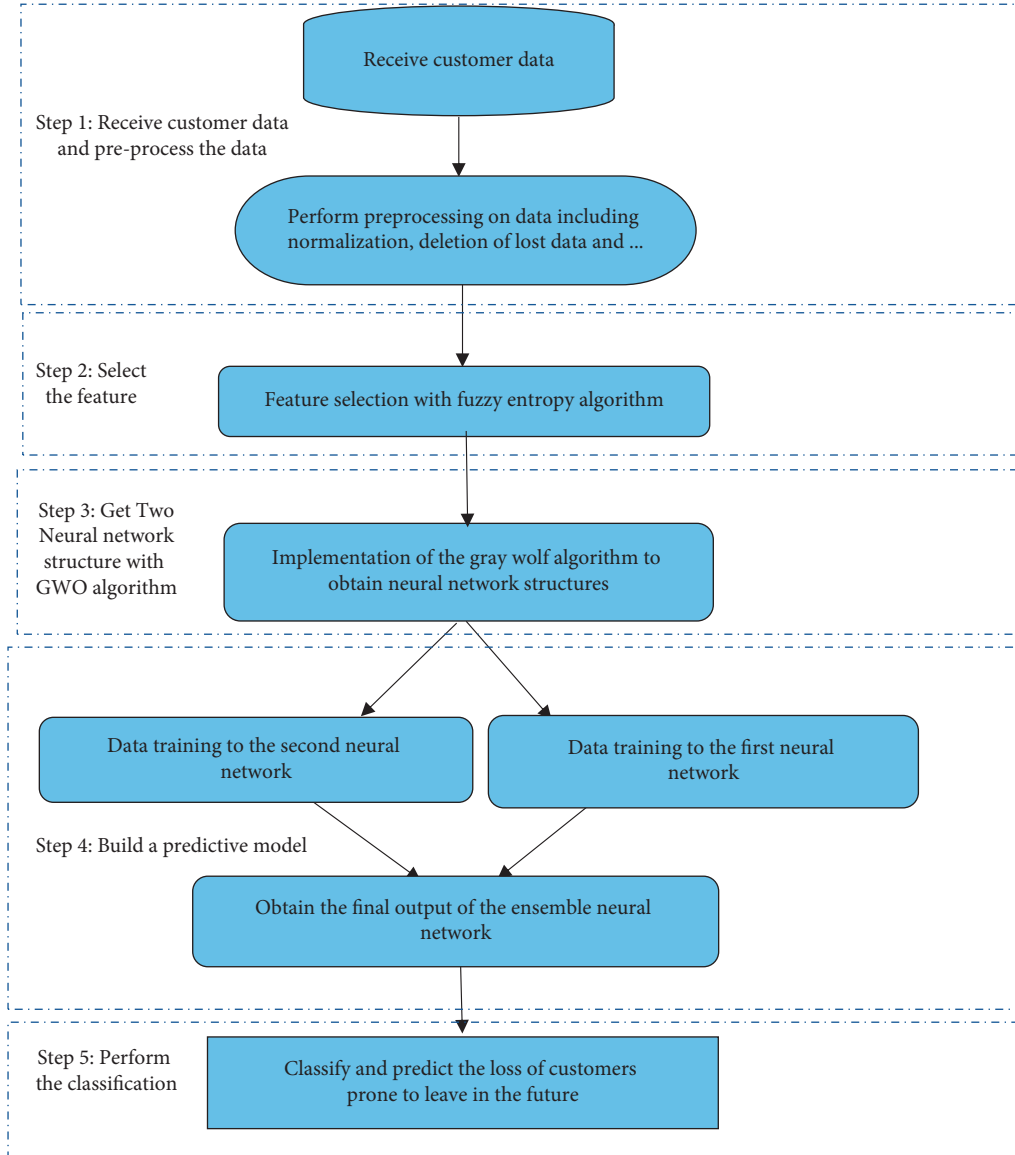
Figure 1: Diagram of the proposed system to predict customer churn.

neural network is a discrete problem. The solution must be a multi-element array with discrete numbers whose number represents the number of layers of the neural network and the value of each cell of the array represents the number of neurons in each layer. Figure 2 presents the solution of the grey wolf algorithm as an array.

On the basis of Figure 2, the array, which consists of four elements, a neural network should be created that consists of four layers in the hidden layer, so that there are 10 neurons in the first layer of this neural network, 20 neurons in the second layer, 8 neurons in the third layer, and 12 neurons in the fourth. Another noteworthy point of this type of encoding is that if the value of one of the cells in the array is 0, that cell should be removed from the array; thus, the number of neural network layers will be reduces.

However, the grey wolf algorithm is a continuous algorithm, and when the search agents are initialized at the



Figure 2: A hypothetical structure for the neural network.

beginning of the optimization process, the values of each search agent will be continuous or decimal numbers. Therefore, because the problem of selecting the optimal structures for the artificial neural network is a discrete problem, so after each iteration of the algorithm and obtaining the best solution by the problem search agents, the resulting solution, which is continuous, must be discretized using the mapping function. For this reason, a mapping function is used as (1)

$$\vec{X}(t)_{i,j} = \left(100 \times \vec{X}(t)_{i,j}\right) \mathrm{mod}\left(h_n\right). \qquad (1)$$

In the (1), $\overrightarrow{X}(t)_{i,j}$ represents the $j^{\text{th}}$ component of the $i^{\text{th}}$ problem search factor in the $t^{\text{th}}$ iteration of the grey wolf algorithm. In other words, $\overrightarrow{X}(t)_{i,j}$ represents the $j^{\text{th}}$ element of the $i^{\text{th}}$ grey wolf. The mod variable is a function for calculating the residual value, and the reason for using this function is to limit the maximum number of neurons in the range 0 to $h_n - 1$. The $h_n$ is the maximum value for the number of hidden layer neurons. Applying this function to any of the problem factors or any of the grey wolves that have continuous solutions will result in discrete solutions. For example, supposing that the value of the first element of one of the solutions to the problem is 0.15 and $h_n$ is equal to 21, the value 15 is obtained.

Another important point about this step of the method is the evaluation function of the search agents in the grey wolf algorithm. That is, in each iteration of the grey wolf algorithm, it is an essential to have a search factor evaluation function to determine which search agent received the best solution. Accordingly, an evaluation function is used to indicate the accuracy of the system for validation data and training data. That is to say, in evaluating each of the population vectors that represent a structure for the neural network, the training data is trained into a neural network with a constructed structure, and after the training process, the accuracy of the neural network prediction is calculated for validation data. Each search engine that maximizes the value of the following relationship is selected as the best member of the population. Equation (2) shows the population members evaluation function.

$$\text{fitness} = 0.7 \times V\_ACC + 0.3 \times T\_ACC. \tag{2}$$

In (2), the variables $V\_ACC$ and $T\_ACC$ indicate the accuracy of forecasting or classifying for the sample in validation and training datasets, respectively. It shows weighted means presenting the accuracy of recognition for validation data and training, so that the importance of accuracy of recognition of validation data is more than the training data.

In this step, two optimal structures must be obtained, so the first solution or the best structure is obtained by alpha wolf, which is the best search agent, and the second solution, or the second best structure is obtained by beta wolf, which represents the second best search engine.

### 3.3. Building Predictive Model with Ensemble Neural Network.
One of the most powerful and famous machine learning tools, namely artificial neural network, has been used to build a model for predicting customer churn. In the proposed system, instead of using a neural network, two neural networks were used to predict the loss of customers with high accuracy by combining the solutions of these two neural networks.

Therefore, at this stage, two multilayer feed neural networks, namely a feed neural network with a solution obtained by alpha wolf and a feed neural network with a solution obtained by beta wolf, were created. Then the dataset was trained to the neural networks with new features obtained from the feature selection stage, and finally the final

solution of the predictive model was calculated using (3), which is a weighted average.

$$\text{Output} = 0.5 \times \text{net } 1 + 0.5 \times \text{net } 2. \tag{3}$$

In (3), net 1 and net 2 are the outputs of the first neural networks, i.e., the neural network made with the alpha wolf solution, and the second neural network, the neural network made with the solution obtained by the beta wolf, respectively. Output also represents the final solution of the neural network, which is obtained by combining the solutions of two neural networks with the same weighted average.

### 3.4. Classify and Predict the Customer Churn.
After building the classification model, the proposed system is ready to predict the customer churn. Therefore, assuming providing a set of data about current customers of a company, the proposed system can classify and predict customers, who are likely to leave the company in the future. Once customers are identified as prone to leave, managers of organizations and companies can prevent the loss of their customers by taking preventive measures in accordance with the policies of their company.

### 3.5. Criteria for Evaluating the Proposed Method.
Predicting customer churn means classifying and categorizing customers into two classes or classes of customers prone to loss as well as customers loyal to the organization. There are various methods for evaluating the efficiency and performance of classification systems, among which criteria such as accuracy, recall, precision, F-measure, and area under curve can be referred. To calculate these classification criteria, a matrix called the configuration matrix should be applied, which represents the performance of classification systems in data classification. Figure 3 shows the configuration matrix for two-class problems.

As shown in Figure 3, the performance of classification systems in data classification using this matrix and with criteria called true positive (TP), true negative (TN), false positive (FP), false negative (FN) is displayed.

(i) True Positive (TP): Equivalent to the percentage of members of a class such as $X$ that the classifier system correctly classifies as a member of Class $X$.

(ii) True Negative (TN): Equivalent to the percentage of members of another class, such as $Y$, whose non-belonging to class $X$ is properly classified.

(iii) False Positive (FP): Equals the percentage of members of another class, such as $Y$, who are incorrectly classified as members of Class $X$.

(iv) False Negative (FN): Equals a percentage of the number of members in class $X$ who are incorrectly classified as members of another class, such as $Y$.

Using the abovementioned criteria, which indicate the performance of classification methods, the performance of classification system can be calculated.

| Ground truth | | | |
|---|---|---|---|
| | + | − | |
| Predicted + | True positive (TP) | False positive (FN) | Precision = TP / (TP + FP) |
| Predicted − | False negative (FN) | True negative (TP) | |
| Recall = TP / (TP + FN) | | | Accuracy = (TP + TN) / (TP + FP + TN + FN) |

FIGURE 3: Confusion matrix and performance criteria of classification methods.

(i) Accuracy: One of the most famous and oldest evaluation criteria of classification systems is the classification accuracy criterion and it is the ratio of all items that are correctly classified. The accuracy of the classification is calculated as

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}}. \quad (4)$$

(ii) Recall: It is the accuracy of the classification system in correctly classifying members of class $X$ who are correctly classified as members of class $X$ and is calculated as equation (5). Another name for the recall evaluation criterion is sensitivity. For example, in the problem of predicting customer churn, if class $X$ is the same as the customer class, this measure indicates the accuracy of the system in correctly classifying the customers.

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}}. \quad (5)$$

(iii) Precision: This criterion indicates what percentage of the members identified as class $X$ members actually belong to class $X$ and is calculated as

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}}. \quad (6)$$

(iv) F-Measure: This criterion is a weighted average between the criteria of accuracy and recall and is used to determine the efficiency of classification systems. This criterion provides more accurate information than the normal average between the accuracy and recall criteria. This criterion is calculated as

$$F - \text{Measure} = 2 * \frac{\text{Precision} * \text{Recall}}{\text{Precision} + \text{Recall}}. \quad (7)$$

(v) Area Under Curve: It indicates the area under the system performance characteristic diagram, and the closer the value of this number to a classification number 1, the more favorable the final performance of the classification method will be evaluated. To calculate the result of the customer churn prediction system in this research, after the system is implemented, it stores the results of its classification in the form of a confusion matrix, so using the results of this matrix, the prior mentioned evaluation criteria are calculated to review and analyze system performance.

## 4. Results

*4.1. Customer Data Collection.* A general dataset has been used to evaluate the efficiency of the proposed method for predicting customer churn. This dataset is for a U.S. telecommunication company and is available at https://www.kaggle.com/blastchar/telco-customer-churn. In order to implement the method of predicting customer churn in this research, MATLAB programming language version 2020a has been used. In this dataset, there are 7043 telecommunication subscribers, each of which has 21 information features. Out of 7043, 11 customer data have lost data in some features, so by deleting the information of this 11 missing data, the remains is a dataset that contains 7032 customer information. Thus, there is a problem with dimensions 21. Out of the total number of customers in this dataset, 1869 people are related to the lost customers, who have left the company after a while and 5174 customers are related to those, who have remained loyal to the company. Table 2 shows some of the features of this dataset.

*4.2. Simulations Results.* After reading the data from the dataset, initial preprocessing was performed on them. This preprocessor involves converting the values of a string variable to numeric variables. Except for properties 3, 6, 19, and 20, which have numeric values, the values of all other properties have string values. So, the values of these variables are converted to numeric values in the preprocessing step. For example, in attribute 4, which has the values yes and no, the value yes is converted to 1 and the value no to 0.

*4.3. Select the Feature Using the Fuzzy Entropy Algorithm.* In this step, before this algorithm is implemented, it is attempted to manually remove the first feature, Customer ID, from the dataset. The reason for this is that logically, an organization's customer identification number is of no importance in terms of classification, so removing this feature from the dataset will increase the number of features to 20. By running this algorithm, two other properties are removed, which are properties 1 and 5.

TABLE 2: Customer characteristics in the U.S. telecommunication company database (https://www.kaggle.com).

| | Attribute name | Description | DataType | Values |
|---|---|---|---|---|
| 1 | customerID | Customer ID | Nominal | 7590-VHVEG, 5575-GNVDE,Etc. |
| 2 | Gender | Customer gender | Nominal | Female, male |
| 3 | SeniorCitizen | Whether the customer is a senior citizen or not | Nominal | 0,1 |
| 4 | Partner | Whether the customer has a partner or not | Nominal | Yes, no |
| 5 | Dependents | Whether the customer has dependents or not | Nominal | Yes, no |
| 6 | Tenure | Number of months the customer has stayed with the company | Numeric | 10,16, etc. |
| 7 | PhoneService | Whether the customer has a phone service or not | Nominal | Yes, no |
| 8 | MultipleLines | Whether the customer has multiple lines or not | Nominal | Yes, no, No phone service |
| 9 | InternetService | Customer's Internet service provider | Nominal | DSL, fiber optic, No |
| 10 | OnlineSecurity | Whether the customer has online security or not | Nominal | Yes, no, internet service |
| 11 | OnlineBackup | Whether the customer has an online backup or not | Nominal | Yes, no, internet service |
| 12 | DeviceProtection | Whether the customer has device protection or not | Nominal | Yes, no, internet service |
| 13 | TechSupport | Whether the customer has tech support or not | Nominal | Yes, no, internet service |
| 14 | StreamingTV | Whether the customer has streaming TV or not | Nominal | Yes, no, internet service |
| 15 | StreamingMovies | Whether the customer has streaming movies or not | Nominal | Yes, no, internet service |
| 16 | Contract | The contract term of the customer | Nominal | Month-to-month, one year, two year |
| 17 | PaperlessBilling | Whether the customer has paperless billing or not | Nominal | Yes, no |
| 18 | PaymentMethod | The customer's payment method electronic check, mailed check, bank transfer | Nominal | Electronic check, mailed heck, bank transfer, credit card |
| 19 | MonthlyCharges | The amount charged to the customer monthly | Numeric | 85.29, 23.16, etc. |
| 20 | TotalCharges | The total amount charged to the customer | Numeric | 85.29, 1899.95, etc. |
| 21 | Churn | Whether the customer churned or not | Nominal | Yes or No |

### 4.4. Obtain Two Optimal Structures Using Grey Wolf.

In order to obtain the appropriate number of layers and the number of neurons in each layer in the proposed method, the grey wolf algorithm is applied. At this stage, the data is randomly divided into three categories: training, validation, and testing. The first category contains 70% of the data used to train the neural network with the structures proposed by the grey wolf algorithm.

First, the grey wolf optimization algorithm starts the optimization process and acquires a structure for the neural network. Then, a neural network is created with this structure and training data is trained to the neural network. After training the neural network with the Levenberg–Marquardt algorithm [21] as one of the best neural network training algorithms, the neural network performance is validated with the dataset.

For example, if the grey wolf algorithm solution is proposed in the first iteration by the best population factor, alpha wolf, it proposes a three-layer neural network with 5, 8, and 10 neurons in the hidden layers one to three, respectively. In this case, a neural network is constructed with such a structure, and the training data is trained to this neural network, and after the completion of the training process, the accuracy of the validation data is calculated based on the evaluation function and the competency of the search agent is determined. This process is performed in different iterations of the algorithm and continues until the number of iterations of the algorithm is exhausted.

After the number of iterations of the algorithm is exhausted, as a result of having used two neural networks in the proposed method, the two best algorithmic solutions are obtained by alpha and beta wolves, as the two structures designated for the neural networks, and two neural networks are created to build a data classifier model.

It is noteworthy that all the parameters of the grey wolf algorithm are similar to the parameters proposed in [20]. Only the number of iterations parameters in this algorithm is considered as 50 iterations and the number of search factors is considered as 30 factors. Additionally, the maximum number of layers considered is a neural network with three layers and also the maximum number of neurons in each layer is a maximum of 20 neurons. The following diagram shows the best structures obtained by the grey wolf algorithm in five different runs of the proposed system.

As can be seen in Figure 4, the grey wolf algorithm has acquired different structures for the neural network in various runs. The reason for this is that each time the algorithm is implemented, the training and validation data are randomly selected, and for this reason, the algorithm obtains different structures appropriate to the training data.

### 4.5. Build a Classifier Model.

After the grey wolf algorithm is implemented in the previous step, the solutions of alpha and beta wolves, which are recognized as the best search agents in the grey wolf algorithm, are as the structures of the feed neural networks. In this step, training of the neural network modules is performed using the Levenberg–Marquardt algorithm.

### 4.6. Predict the Customer Churn.

After building a model for predicting customer churn in the previous step, it is time to evaluate the efficiency and performance of the proposed system. At this stage, the system performance is evaluated using test data. To evaluate the performance of the proposed system at this stage, the confusion matrix and the criteria of accuracy, recall, precision as well as the area under curve are applied. Figures 5–8 show the results of running the
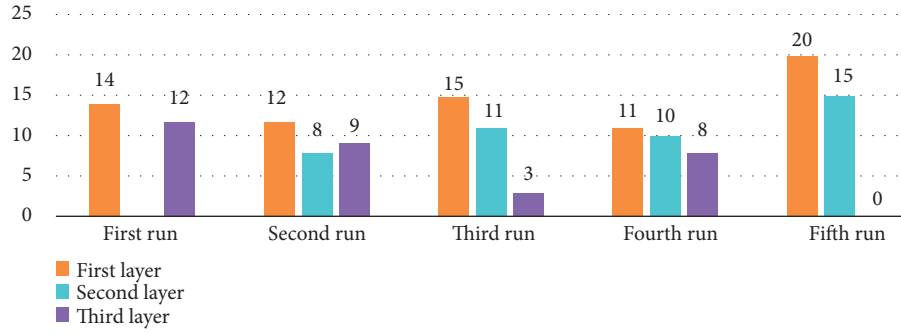
FIGURE 4: The best structures obtained by the grey wolf algorithm in five different implementations.
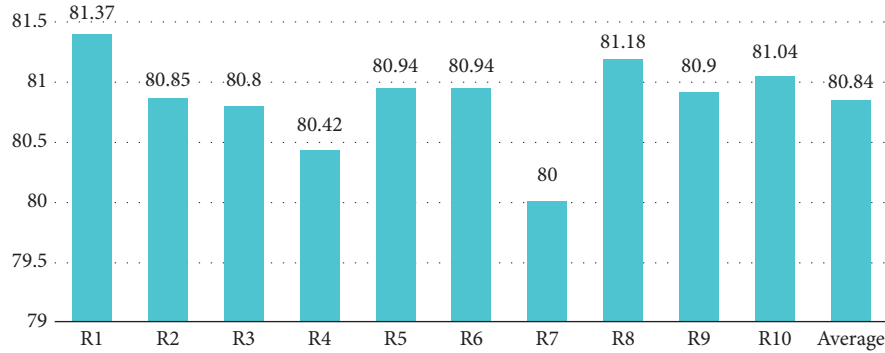


FIGURE 5: Results of 10 runs of the proposed method in the classification accuracy criterion with an average of 10 runs.



FIGURE 6: Results of 10 runs of the proposed method in the classification precision criterion with an average of 10 runs.

proposed method regarding the criteria of accuracy, precision, recall, and F_measure, respectively.

As shown in Figures 5–8, averages of 10 different runs of the proposed system in predicting customer churn in the criteria of accuracy, precision, recall, and F_measure are 80.84, 84.45, 91.08, and 87.64, respectively. Besides, the confusion matrix obtained from the best and worst performance of the proposed system on the test data is presented in Figures 9 and 10.

As can be seen from Figure 9, the accuracy of the proposed system in this performance for the test data is 81.4%. Also, it shows that the performance of the proposed system in the criteria of precision, recall, and F_measure is 84.46%, 91.92%, and 88.03%, respectively.

Figure 10 shows the worst performance of the proposed system, and the accuracy in this run for test data is 0.80%. In

addition, the performance of the proposed system in the criteria of precision, call and F_measure is 83.67%, 90.9%, and 87.14%, respectively.

Another criterion used to evaluate the performance of classifier systems is area under curve. The closer the area under the system performance characteristic curve diagram is 1, the better the system performance is considered. Figure 11 shows the system performance characteristic diagram in the 10 proposed system runs for the test data. Additionally, Figures 12 and 13 present the area under the system performance characteristic curve diagram in the best and worst performance for training and test data, respectively.

Afterwards, the results of the proposed hybrid system with some of the other methods examined on this dataset are compared. Table 3 shows the results of the proposed method and several other methods in the evaluation criteria of
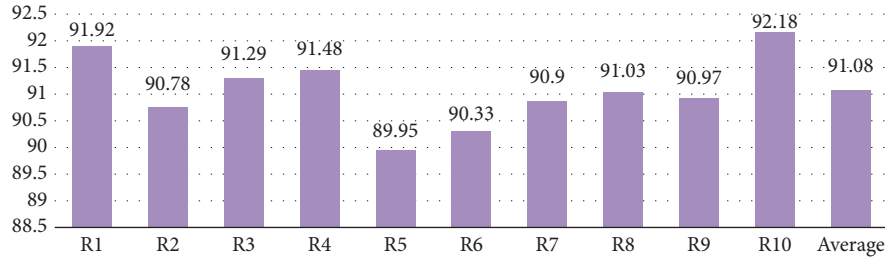
FIGURE 7: Results of 10 runs of the proposed method in the classification recall criterion with an average of 10 runs.
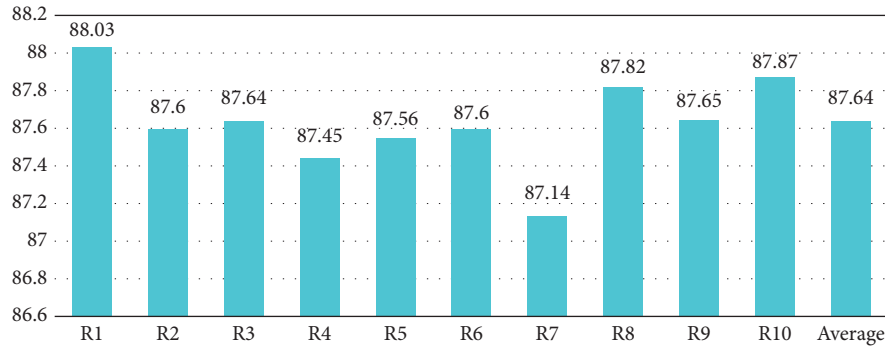


FIGURE 8: Results of 10 runs of the proposed method in the classification recall F_measure with an average of 10 runs.



FIGURE 9: Confusion matrix obtained from the best performance of the proposed system in 10 runs.



FIGURE 10: Confusion matrix resulting from the worst performance of the proposed system in 10 runs.

accuracy, precision, recall, and F_measure. Results of the proposed method are the average of 10 runs of the proposed method. The results of the proposed method are compared with the results of the methods presented in Halibas et al. [22], i.e., Naive Bayes, generalized linear model, logistic regression, deep learning, decision tree, random forest, and Gradient Boosted Tree.

A noteworthy point in these comparisons is that in these methods a sampling method is used and the results of the methods with and without the sampling method have been examined. However, the results in Table 3 for each of the

methods are related to their best results from with and without sampling method.

As can be seen in Table 3, the results of the proposed method are better than other methods in assessing the accuracy, precision, recall, and F_measure criteria. The reasons for the superiority of the proposed method over other methods can be shown as follows

(i) As a result of the use of grey wolf algorithm to select the appropriate structure and architecture for neural networks used in building the predictive model of the proposed system.
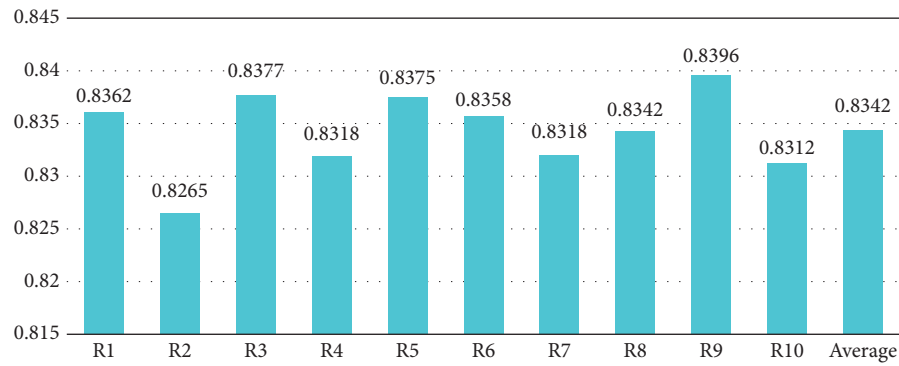
FIGURE 11: Results of 10 runs of the proposed method in the evaluation criterion of area under curve with the average of 10 runs.
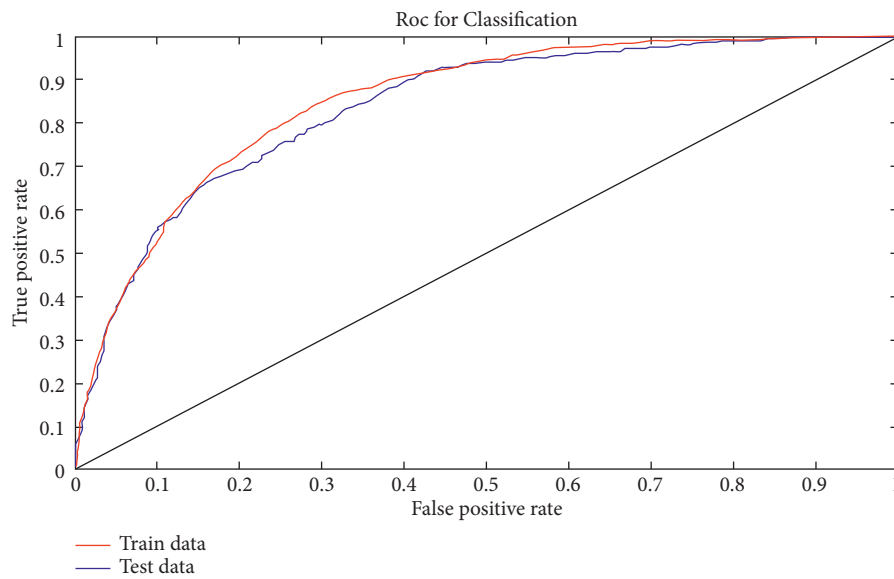


FIGURE 12: The area under the system performance characteristic curve diagram in the best performance.
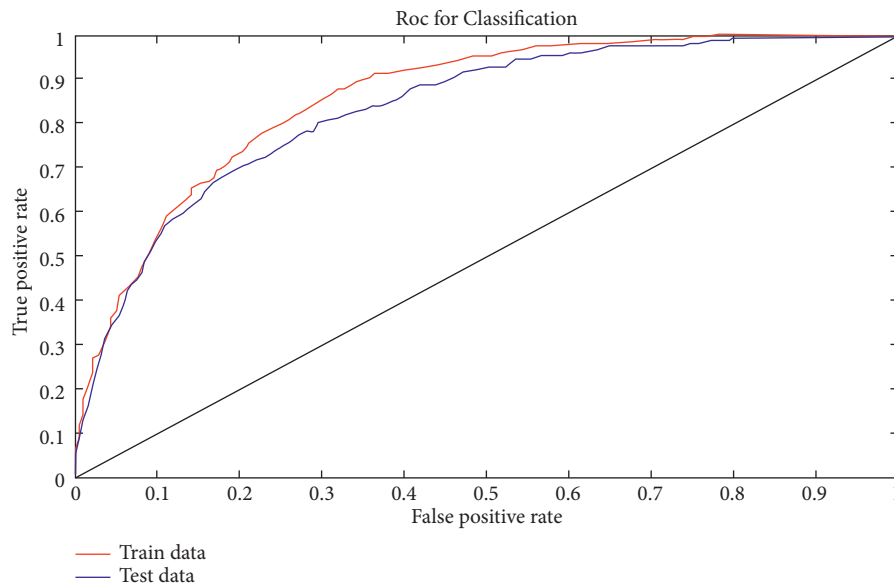


FIGURE 13: The area under the system performance characteristic curve diagram in the worst performance.

TABLE 3: Comparing the performance of the proposed system and other methods in different evaluation criteria.

| No. | Methods | F_measure | Recall | Precision | Accuracy |
|-----|---------|-----------|--------|-----------|----------|
| 1 | Naive Bayes | 72.3 | 83.9 | 63.6 | 73.0 |
| 2 | Generalized linear model | 71.6 | 72.8 | 70.4 | 75.7 |
| 3 | Logistic regression | 71.6 | 72.9 | 70.40 | 75.7 |
| 4 | Deep learning | 72.6 | 80.9 | 66.0 | 74.3 |
| 5 | Decision tree | 70.7 | 84.6 | 61.10 | 70.7 |
| 6 | Random forest | 70.5 | 70.5 | 70.6 | 75.2 |
| 7 | Gradient boosted tree | 76.2 | 79.6 | 73.1 | 79.1 |
| 8 | The proposed method | **87.64** | **91.08** | **84.45** | **80.84** |

(ii) In addition, combining the solutions of the two neural networks and obtaining the final solution makes the proposed method perform better than other methods.

So, it can be used as an efficient method in predicting customer churn of organizations and companies that face the challenge of losing their customers and suffer great financial losses due to the loss of their customers.

## 5. Conclusion

In this research, it was attempted to propose a hybrid system based on fuzzy entropy measures with similarity classifier, grey wolf algorithm, and artificial neural network to predict the customer churn of companies that suffer from the loss of their customers and financial losses. The use of this system can play a decisive role in the survival of companies in the area of market competition. A company can use this system to identify those customers who might leave the company in the future for various reasons, and take preventive measures based on customer interests and policies.

At first, customer's data of the company was collected. Then, various preprocessing methods were performed on the data, such as conversion of string data into numerical data, and data normalization so that the data could be prepared for the selection of features and training of the neural network. So, the feature selection process was performed using the fuzzy entropy measures with similarity classifier on the dataset. In the next step, the grey wolf optimization algorithm was started and during a process of optimization and different iterations, two optimal structures were obtained for the neural network (i.e., determining the number of layers and the number of neurons in each layer). These structures were used in the phase of building the classifier model to predict customer churn. Finally, after building the classifier model, the proposed system was evaluated with test data. The results of the proposed system performance in predicting customer churn were examined on a customer dataset of a telecommunication company. The results showed that the proposed system had a better performance in different evaluation criteria (accuracy, precision, recall, and F_measure) than other data mining methods.

The contribution of this study is the use of the fuzzy entropy measures with similarity classier to select the optimal features, the grey wolf optimization algorithm to obtain optimal structures for neural networks, and ensemble neural networks to build a model for the aim of predicting customer churn.

Therefore, managers of various organizations including banks, insurance companies, telecommunications industry, stores, online games, online betting, various energy industries, food distributors, and dozens of other businesses can use the proposed system to identify those customers who are likely to leave, gain more profit by taking precautionary measures to retain those customers, and avoid other relevant losses such as financial losses.

## Data Availability

The data are contained within the article itself. Upon request, the software file will be provided.

## Conflicts of Interest

The authors declare that they have no conflicts of interest.

## References

[1] A. Amin, F. Al-Obeidat, B. Shah, A. Adnan, J. Loo, and S. Anwar, "Customer churn prediction in telecommunication industry using data certainty," *Journal of Business Research*, vol. 94, pp. 290–301, 2019.

[2] T. Vafeiadis, K. I. Diamantaras, G. Sarigiannidis, and K. C. Chatzisavvas, "A comparison of machine learning techniques for customer churn prediction," *Simulation Modelling Practice and Theory*, vol. 55, pp. 1–9, 2015.

[3] A. Amin, S. Anwar, A. Adnan et al., "Customer churn prediction in the telecommunication sector using a rough set approach," *Neurocomputing*, vol. 237, pp. 242–254, 2017.

[4] K. W. De Bock and D. Van den Poel, "Reconciling performance and interpretability in customer churn prediction using ensemble learning based on generalized additive models," *Expert Systems with Applications*, vol. 39, no. 8, pp. 6816–6826, 2012.

[5] Z.-Y. Chen, Z.-P. Fan, and M. Sun, "A hierarchical multiple kernel support vector machine for customer churn prediction using longitudinal behavioral data," *European Journal of Operational Research*, vol. 223, no. 2, pp. 461–472, 2012.

[6] A. Sharma, D. Panigrahi, and P. Kumar, "A neural network based approach for predicting customer churn in cellular network services," 2013, https://arxiv.org/ftp/arxiv/papers/1309/1309.3945.pdf.

[7] K. Coussement and K. W. De Bock, "Customer churn prediction in the online gambling industry: the beneficial effect of ensemble learning," *Journal of Business Research*, vol. 66, no. 9, pp. 1629–1636, 2013.

[8] L. Tang, L. Thomas, M. Fletcher, J. Pan, and A. Marshall, "Assessing the impact of derived behavior information on customer attrition in the financial service industry," *European Journal of Operational Research*, vol. 236, no. 2, pp. 624–633, 2014.

[9] A. Keramati, R. Jafari-Marandi, M. Aliannejadi, I. Ahmadian, M. Mozaffari, and U. Abbasi, "Improved churn prediction in telecommunication industry using data mining techniques," *Applied Soft Computing*, vol. 24, pp. 994–1012, 2014.

[10] J. Moeyersoms and D. Martens, "Including high-cardinality attributes in predictive models: a case study in churn prediction in the energy sector," *Decision Support Systems*, vol. 72, pp. 72–81, 2015.

[11] K. Coussement, S. Lessmann, and G. Verstraeten, "A comparative analysis of data preparation algorithms for customer churn prediction: a case study in the telecommunication industry," *Decision Support Systems*, vol. 95, pp. 27–36, 2017.

[12] R. Yu, X. An, B. Jin, J. Shi, O. A. Move, and X. Y. Liu, "Particle classification optimization-based BP network for telecommunication customer churn prediction," *Neural Computing & Applications*, vol. 29, no. 3, pp. 707–720, 2018.

[13] H. Salvi, A. Shah, M. Mehta, and S. Correia, "Long short-term model for Brent oil price forecasting," *International Journal for Research in Applied Science and Engineering Technology*, vol. 7, no. 11, pp. 315–319, 2019.

[14] N. Moitra, P. Raj, S. Saxena, and R. Kumar, "Crude oil prediction using LSTM," *International Journal of Innovative Science and Research Technology*, vol. 5, no. 2, pp. 900–905, 2020.

[15] S. Jafarzadeh Ghoushchi, R. Ranjbarzadeh, S. A. Najafabadi, E. Osgooei, and E. B. Tirkolaee, "An extended approach to the diagnosis of tumour location in breast cancer using deep learning," *Journal of Ambient Intelligence and Humanized Computing*, 2021.

[16] S. M. T. U. Raju, A. Sarker, A. Das et al., "An approach for demand forecasting in steel industries using ensemble learning," *Complexity*, vol. 2022, pp. 1–19, 2022.

[17] E. Ofori-Ntow Jnr, Y. Y. Ziggah, and S. Relvas, "Hybrid ensemble intelligent model based on wavelet transform, swarm intelligence and artificial neural network for electricity demand forecasting," *Sustainable Cities and Society*, vol. 66, Article ID 102679, 2021.

[18] J. Jayashree and S. Ananda Kumar, "Hybrid swarm intelligent redundancy relevance (RR) with convolution trained compositional pattern neural network expert system for diagnosis of diabetes," *Health Technology*, vol. 10, no. 1, pp. 197–206, 2019.

[19] P. Luukka, "Feature selection using fuzzy entropy measures with similarity classifier," *Expert Systems with Applications*, vol. 38, no. 4, pp. 4600–4607, 2011.

[20] S. Mirjalili, "How effective is the Grey Wolf optimizer in training multi-layer perceptrons," *Applied Intelligence*, vol. 43, no. 1, pp. 150–161, 2015.

[21] M. T. Hagan and M. B. Menhaj, "Training feedforward networks with the Marquardt algorithm," *IEEE Transactions on Neural Networks*, vol. 5, no. 6, pp. 989–993, 1994.

[22] A. S. Halibas, A. C. Matthew, I. G. Pillai, J. H. Reazol, E. G. Delvo, and L. B. Reazol, "Determining the intervening effects of exploratory data analysis and feature engineering in telecoms customer churn modelling," in *Proceedings of the 2019 4th MEC International Conference on Big Data and Smart City (ICBDSC)*, pp. 1–7, IEEE, Muscat, Oman, 15 January 2019.