DETERMINATION OF *CRYPTOSPORIDIUM* SPP. RISK FACTORS USING MULTILAYER PERCEPTRON NEURAL NETWORK AND RADIAL BASED FUNCTIONAL ARTIFICIAL NEURAL NETWORK METHOD

U. Karaman, and I. Balikci Cicek

Abstract— *Aim:* In the study, it is aimed to compare the estimates of Multilayer artificial neural network (MLPNN) and radial based function artificial neural network (RBFNN) methods, which are among the artificial neural network models in the presence and absence of *Cryptosporidium* spp., and to determine the factors associated with parasite.

Materials and Methods: In the study, "*Cryptosporidium* spp. Dataset," the data set named was obtained from Ordu University. In order to classify the presence and absence of *Cryptosporidium* spp, MLPNN, and RBFNN methods, which are among the artificial neural network models, were used. The classification performance of the models was evaluated with accuracy from the classification performance criteria.

Results: The accuracy, which is the performance criterion obtained with MLPNN, was obtained as 75% of the applied models. The accuracy, which is the performance criterion obtained with the RBFNN model, was achieved as 71.4%. When the effects of variables in the data set in this study on the presence and absence of *Cryptosporidium* spp. are examined, the three most important variables for the MLPNN model were nausea-vomiting, General Puriri, and sex, respectively. For the RBFNN model, age was obtained as cancer and General Puriri.

Conclusion: It was seen that MLPNN and RBFNN models used in this study gave successful predictions in classifying the presence and absence of *Cryptosporidium* spp.

Keywords— Multilayer perceptron neural network, Radial-based function neural network, classification, *Cryptosporidium* spp., risk factors.

1. INTRODUCTION

Intestinal parasites are a significant public health problem worldwide, including in developing countries [1]. Among the intestinal parasites, the coccidian parasites cryptosporidium is one of the obligate intracellular parasites that cause diarrhea in all age groups and individuals with normal immunity [2].

Among the Criptosporidiums (*Cryptosporidium* spp.), C. parvum, the most diseased species in humans, is located in the microvilli of intestinal epithelial cells, causing short-term (about two weeks) spontaneous diarrhea in people with sufficient immunity, and maybe life-threatening in the host whose immune system is suppressed [3]. In immunocompromised individuals, the parasite can spread from the intestinal tract to the bile ducts, pancreas, stomach, respiratory system, and kidneys through the hematogenous pathway. Cryptosporidiums can be transmitted by contaminated water and food, from person to person or from animal to person [4].

Artificial neural networks (ANN) are parallel, distributed information processing models that are developed using the physiology of the human brain and are connected by weighted connections, each consisting of processing elements with their own memory, and are computer programs that imitate biological neural networks [5]. The most essential task of an artificial neural network is to determine an output set that can correspond to an input set shown to it. In order to do this, the network is trained with examples of the relevant event and gained the ability to generalize [6].

Multilayer Artificial Neural Networks (MLPNN) model has been the most used neural network model, especially in medical and engineering applications. This model is widely used because many learning algorithms can be easily used in the training of this network. Multilayer networks consist of an input layer, one or more hidden layers, and an output layer. The information flow is forward, and there is no feedback [7]. The purpose of this method is; It is to make the error between the desired output of the network and the output it produces to a minimum [8].

Feedforward neural networks are widely used in many areas, such as controlling nonlinear systems in addition to modeling. One of the feed-forward neural networks is Radial Based Function Artificial Neural Networks (RBFNN) [9]. RBFNN is a particular case of a multilayer feed-forward artificial neural network and has two distinctive characteristics. The first is that it has only one hidden layer. The second feature is that radial based functions are used as activation functions in the hidden layer. Another essential feature of radial-based artificial neural networks is the transfer of information from input neurons to hidden layer neurons without change [10].

In this study, *Cryptosporidium* spp. by applying MLPNN and RBFNN methods to the data set, it aimed to classify the presence and absence of *Cryptosporidium* spp. and determine the risk factors.

İpek BALIKCI CICEK, Inonu University Department of Biostatistics and Medical Informatics, Faculty of Medicine, Malatya, Turkey, (ipek.balikci@inonu.edu.tr)

Ülkü KARAMAN, Ordu University Department of Medical Parasitology Dep., Faculty of Medicine, Ordu, Turkey, (ukaraman@ordu.edu.tr)

Manuscript received Oct 4, 2020; accepted Oct 29, 2020.
Digital Object Identifier:

2. Material and methods

2.1. Dataset

In this study, *Cryptosporidium* spp. classification process was performed by applying MLPNN and RBFNN methods to the "*Cryptosporidium* spp" data set obtained from Ordu University for the presence-absence situation. There are a total of 497 patients in this dataset. There were 142 (28.6%) people with *Cryptosporidium* spp. and 355 (71.4%) without the parasite. The variables and the descriptive properties of the variables in the relevant data set are given in Table 1.

Table I
Variables in the data set and descriptive properties of variables

Variables	**Variable Explanation**	**Variable type**	**Variable role**
Cryptosporidium spp.	Parasite (0 = absence, 1 = presence)	Qualitative	Dependent/ Target
Age	Age	Quantitative	Independent/ Predictor
Sex	1=male, 2=female	Qualitative	Independent/ Predictor
nausea / vomiting	0= no nausea or vomiting 1= there is nausea and vomiting	Qualitative	Independent/ Predictor
immunosuppressive	0 = absence, 1 = presence	Qualitative	Independent/ Predictor
eosinophilia	0 = absence, 1 = presence	Qualitative	Independent/ Predictor
diabetes	0 = absence, 1 = presence	Qualitative	Independent/ Predictor
cancer	0 = absence, 1 = presence r	Qualitative	Independent/ Predictor
urine syphilis	0 = absence, 1 = presence	Qualitative	Independent/ Predictor
diarrhea	0 = absence, 1 = presence	Qualitative	Independent/ Predictor
neutropenia	0 = absence, 1 = presence	Qualitative	Independent/ Predictor
obesity	0 = absence, 1 = presence	Qualitative	Independent/ Predictor
anemia	0 = absence, 1 = presence	Qualitative	Independent/ Predictor
Aurtiker	0 = absence, 1 = presence	Qualitative	Independent/ Predictor
General puriri	0 = absence, 1 = presence	Qualitative	Independent/ Predictor
Kürtiker	0 = absence, 1 = presence	Qualitative	Independent/ Predictor
urticaria	0 = absence, 1 = presence	Qualitative	Independent/ Predictor
ucolite	0 = absence, 1 = presence	Qualitative	Independent/ Predictor

3. Multilayer Perceptron Neural Network (MLPNN)

This model, which was developed by Rumelhart in 1986, is also called the error propagation model. Multilayer Perceptron neural networks have multiple layers between the input and output layers. It consists of the input layer, output layer, and intermediate layers. The processor elements in the input layer act as a buffer that distributes the input signals to the processor elements in the middle layer. The information flow runs from the input layer to the middle layer and then to the output layer [11].

Training of the multi-layered artificial neural network is carried out according to the "generalized delta rule". In multilayer neural networks, first, an example of the network is introduced. As a result of the example, what kind of result will be obtained is revealed. The examples are applied to the input layer, processed in the intermediate layers, and the error between the desired output is spread back over the weights, changing the weights until the error is minimized. The multi-layer perceptron network is a forward feed network, and the most general result is obtained from the output layer [12].

3.1. Radial Based Function Neural Network (RBFNN)

Radial Based Artificial Neural network first emerged in a multivariate real interpolation solution [13]. RBFNN is a curve-fitting approach in multidimensional space. Training of RBFNN can be called the problem of finding the best surface suitable for data in multidimensional space. RBFNN is feed-forward networks consisting of three layers as input, hidden, and output layer.

The input layer consists of a source of artificial nerve cells. The hidden layer is the hidden layer whose number of artificial nerve cells can be changed and uses a radian-based function as an activation function. The output layer is the part where the output of the network is produced according to the input values. There is a non-linear transformation from the input layer to the hidden layer and a linear transformation from the hidden layer to the output layer [14].

RBFNN working principle; to determine the radial-based function with the appropriate width and center values in the intermediate layer according to the input values in the input layer and to create the linear combinations of the outputs of the radial-based functions with the appropriate weight values and to determine the relationship between the input values and the output values [15].

3.2. Performance Evaluation of Models

In the performance evaluation of the radial-based artificial neural network and multilayer artificial neural network models, which were created to predict the factors that may be associated with the presence or absence of *Cryptosporidium* spp., the performance criteria obtained by using the classification matrix given below were used.

The performance criteria used in the performance evaluation of the models in this study are given below.

Accuracy = (TP+TN)/(TP+TN+FP+FN)

TABLE II
CLASSIFICATION MATRIX FOR CALCULATING PERFORMANCE CRITERIA

		Real		
		Positive	**Negative**	**Total**
Predicted	**Positive**	True positive (TP)	False negative (FN)	TP+FN
	Negative	False positive (FP)	True negative (TN)	FP+TN
	Total	TP+FP	FN+TN	TP+TN+ FP+FN

4. DATA ANALYSIS

Quantitative data are expressed as median (minimum-maximum) and qualitative data as number (percentage). The Kolmogorov-Smirnov test evaluated conformity to normal distribution.

In terms of independent variables, whether there is a statistically significant difference between the "parasite presence " and "parasite absence " groups, which are the categories of the dependent/target variable (*Cryptosporidium* spp.), and whether there is a relationship, Mann-Whitney U test, Pearson chi-square test, Continuity Correction test, and Fisher's Exact test. It was examined using the chi-square test values of $p<0.05$ were considered statistically significant. IBM SPSS Statistics 26.0 package program was used for all analyzes.

For the validity of the model, a 10-fold cross-validation method was used. In the 10-fold cross-validation method, all data is divided into ten equal parts. One part is used as a test set, and the remaining nine parts are used as a training dataset, and this process is repeated ten times.

5. RESULTS

Descriptive statistics for quantitative independent variables examined in this study are given in Table 3, and descriptive statistics for qualitative independent variables are given in Table 4. There is a statistically significant relationship between the dependent/target variable groups ($p<0.05$) in terms of cancer variable.

TABLE III
DESCRIPTIVE STATISTICS FOR QUANTITATIVE INDEPENDENT VARIABLES

Variable	Cryptosporidium		**p-value***
	Parasite absence	Parasite presence	
	Median(min-max)	**Median(min-max)**	
age	22 (1-84)	21 (2-72)	0.619

***: Mann Whitney U test**

TABLE IV
DESCRIPTIVE STATISTICS FOR QUALITATIVE INDEPENDENT VARIABLES

Variables		Cryptosporidium		**p-value**
		Parasite absence	Parasite presence	
		Number (%)	**Number (%)**	
sex	male	184 (51.8)	76 (53.5)	0.733*
	female	171 (48.2)	66 (46.5)	
nausea / vomiting	absence	341 (96.1)	135 (95.1)	0.805**
	presence	14 (3.9)	7 (4.9)	
immunosuppressive	absence	344 (96.9)	136 (95.8)	0.586***
	presence	11 (3.1)	6 (4.2)	
eosinophilia	absence	352 (99.2)	142 (100)	0.561***
	presence	3 (0.8)	0 (0)	
diabetes	absence	353 (99.4)	141 (99.3)	1***
	presence	2 (0.6)	1 (0.7)	
cancer	absence	336 (94.6)	124 (87.3)	**0.009****
	presence	19 (5.4)	18 (12.7)	
urine syphilis	absence	350 (98.6)	142 (100)	0.328***
	presence	5 (1.4)	0 (0)	
diarrhea	absence	280 (78.9)	108 (76.1)	0.493*
	presence	75 (21.1)	34 (23.9)	
neutropenia	absence	350 (98.6)	142 (100)	0.328***
	presence	5 (1.4)	0 (0)	
obesity	absence	353 (99.4)	142 (100)	1***
	presence	2 (0.6)	0 (0)	
anemia	absence	344 (96.9)	134 (94.4)	0.283**
	presence	11 (3.1)	8 (5.6)	
General puriri	absence	336 (94.6)	132 (93.0)	0.607**
	presence	19 (5.4)	10 (7.0)	
Aurtiker	absence	352 (99.2)	141 (99.3)	1***
	presence	3 (0.8)	1 (0.7)	
Kürtiker	absence	349 (98.3)	139 (97.9)	0.719***
	presence	6 (1.7)	3 (2.1)	
urticaria	absence	344 (96.9)	137 (96.5)	0.783***
	presence	11 (3.1)	5 (3.5)	
ucolite	absence	351 (98.9)	138 (97.2)	0.233***
	presence	4 (1.1)	4 (2.8)	

***: Pearson chi-square test, **: Continuity Correction test,**
*****: Fisher's Exact test**

Classification martix of MLPNN and RBFNN models are given in Table 5 and Table 6, respectively.

TABLE V
CLASSIFICATION MATRIX OF MLPNN MODEL

Real / **Predicted**	presence	absence	**Total**
presence	2	38	40
absence	0	112	112
Total	2	150	152

TABLE VI
CLASSIFICATION MATRIX OF RBFNN MODEL

Real Predicted	presence	absence	Total
presence	0	44	44
absence	0	110	110
Total	0	154	154

Table 7, shows the values of the performance criteria calculated from the models created to classify the *Cryptosporidium* spp.

TABLE VII
PERFORMANCE CRITERIA VALUES CALCULATED FROM CREATED MODELS IN THE TESTING PHASE

Model Performance Metric	MLPNN	RBFNN
	Value	Value
Accuracy (%)	75.0	71.4
AUC	0.515	0.547

AUC: Area under the ROC curve; MLPNN: Multilayer Perceptron Neural Network; RBFNN: Radial Based Function Neural Network

In this study, the importance values of the factors associated with the *Cryptosporidium* spp. are given in Table 8, while the values for these importance percentages are shown in Figure 1.

TABLE VIII
IMPORTANCE VALUES OF EXPLANATORY VARIABLES ACCORDING TO MLPNN AND RBFNN MODELS

Explanatory Variables	MLPNN	RBFNN
sex	0.085	0.007
nausea / vomiting	0.185	0.069
immunosuppressive	0.032	0.067
eosinophilia	0.062	0.056
diabetes	0.019	0.055
cancer	0.025	0.076
urine syphilis	0.029	0.053
diarrhea	0.030	0.009
neutropenia	0.031	0.056
obesity	0.051	0.056
anemi	0.072	0.063
General puriri	0.157	0.072
Aurtiker	0.038	0.059
Kürtiker	0.019	0.063
urticaria	0.070	0.065
ucolite	0.075	0.061
age	0.018	0.114
Total	1	1

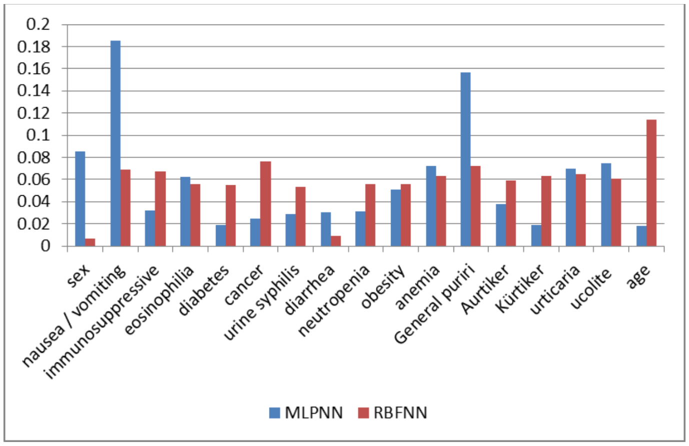

Fig.1. The importance values for possible risk factors

6. DISCUSSION

Cryptosporidium spp. oocysts, which are obligate intracellular parasites, are 4-6 µm in size and spread among living things as a result of consuming water and food contaminated with feces [4, 16, 17].

The parasite has a high prevalence rate in some occupational groups (animal husbandry, veterinarians, laboratory staff, nursery staff), people who travel to endemic areas, those who live in places where hygienic conditions are inadequate, children, the elderly, and those who come into close contact with infected people [18]. *Cryptosporidium* outbreaks have been reported from public swimming pools, common meals, well water, and unhygienic drinking water. It has also been stated that there may be a transition from animals to humans in rural areas [18]. In the spread of the parasite, lack of clean water and sanitation facilities, crowded home environment, and close contact of reservoir animals to individuals are potentially effective [19].Symptoms of cryptosporidiosis differ depending on the type of infected host, the state of the immune system, and age [20].

Artificial neural networks are a method used to estimate the relationship between dependent and independent variables. Artificial Neural Networks in general; It is accepted as a powerful method es such as parameter estimation, classification, and the structure of existing data in many statistical processes parameter estimation, classification, and optimization. Artificial neural networks can reveal complex relationships between predictive variables and make inferences [21].

In this study, the multilayer artificial neural network and radial-based artificial neural network models, which are among the artificial neural network models, were obtained from *Cryptosporidium* spp. It was applied to the data set and aimed to compare the classification estimates of these two models. In this context, *Cryptosporidium* spp. The factors that may be associated with the positivity-negativity (dependent variable) were estimated by multilayer artificial neural network and radial based artificial neural network models. Thus, it has been shown that artificial neural network models can be used in the classification problem.

The imbalanced class problem is one of the important topics in machine learning. In the data set being studied, one of the classes formed by the observations is higher in number than the class or classes formed by other observations, revealing the imbalanced class problem. The problem of bias arises in the classification of sick and non-sick individuals in a two-class data set with an imbalance in the distribution between classes. Because machine learning models used in classification and sensitive to unbalanced class distributions are under the influence of large class (s) and the existence of small classes disappears [22, 23]. There were 142 (28.6%) people with *Cryptosporidium* spp. and 355 (71.4%) people without parasite in this data set. This situation causes the classification algorithms to give biased results and the results to be interpreted incorrectly. Therefore, the accuracy value obtained in the MLPNN and RBFNN models was 75.0 and 71.4, respectively.

In this study, among the performance criteria used to compare classification performances according to the accuracy result, the MLPNN model gave better predictive results than the RBFNN model in the classification of presence-absence of *Cryptosporidium* spp. According to the MLPNN model, the three most important risk factors that may be associated with the presence and absence of *Cryptosporidium* spp. are; nausea-vomiting, general puriri, and sex have been obtained. The RBFNN model estimated as age, cancer, and general puriri.

The presence of the parasite according to the artificial neural network models used according to the findings obtained in the study; They differ in nausea-vomiting, general puriri, age, cancer, and sex variables. Accordingly, it was concluded that *Cryptosporidium* spp. positivity should be investigated in line with the complaints of general puriri, nausea-vomiting, and cancer patients.

REFERENCES

[1] B. Bhattachan, J. B. Sherchand, S. Tandukar, B. G. Dhoubhadel, L. Gauchan, and G. Rai, "Detection of Cryptosporidium parvum and Cyclospora cayetanensis infections among people living in a slum area in Kathmandu valley, Nepal," BMC research notes, vol. 10, pp. 1-5, 2017.

[2] P. Kumar, O. Vats, D. Kumar, and S. Singh, "Coccidian intestinal parasites among immunocompetent children presenting with diarrhea: Are we missing them?," Tropical Parasitology, vol. 7, p. 37, 2017.

[3] D. Dirim Erdoğan, N. Turgay, and M. Z. Alkan, "Bir Cryptosporidiosis olgusunun kinyoun Asit-fast boyası ve polimeraz zincir reaksiyonu (PZR) ile takibi," Türkiye Parazitoloji Dergisi, vol. 27, pp. 237-239, 2003.

[4] D. P. Clark, "New insights into human cryptosporidiosis," Clinical Microbiology Reviews, vol. 12, pp. 554-563, 1999.

[5] [M. Çuhadar, "Türkiye'ye yönelik diş turizm talebinin MLP, RBF ve TDNN yapay sinir aği mimarileri ile modellenmesi ve tahmini: karşilaştirmali bir analiz," Journal of Yasar University, vol. 8, 2013.

[6] S. Tuna, "Şablon eşleme ve çok katmanlı algılayıcı kullanılarak yüz tanıma sisteminin gerçeklenmesi," 2008.

[7] A. Auclair, "Feed-forward neural networks applied to the estimation of magnetic field distributions," 2004.

[8] Ç. Çatal and L. Özyilmaz, "Çok katmanli algilayici ile multiple myeloma hastaliğinin gen ekspresiyon veri çözümlenmesi."

[9] C.-N. Ko, "Identification of non-linear systems using radial basis function neural networks with time-varying learning algorithm," IET signal processing, vol. 6, pp. 91-98, 2012.

[10] O. Akbilgiç, Ğ. Danışman, and H. Bozdoğan, "Hibrit radyal tabanli fonksiyon ağlari ile değişken seçimi ve tahminleme: menkul kiymet yatirim kararlarina ilişkin bir uygulama."

[11] B. K. Wong, T. A. Bodnovich, and Y. Selvi, "Neural network applications in business: A review and analysis of the literature (1988–1995)," Decision Support Systems, vol. 19, pp. 301-320, 1997.

[12] E. Öztemel and Y. S. Ağları, "Papatya yayıncılık," İstanbul, 2003.

[13] M. D. Buhmann, Radial basis functions: theory and implementations vol. 12: Cambridge university press, 2003.

[14] U. Okkan and H. yıldırım Dalkiliç, "Radyal tabanlı yapay sinir ağları ile Kemer Barajı aylık akımlarının modellenmesi," Teknik Dergi, vol. 23, pp. 5957-5966, 2012.

[15] C. Cetinkaya, "Retina Görüntülerinde Radyal Tabanlı Fonksiyon Sinir Ağları İle Damar Tipik Noktalarının Tespit Edilmesi," Ege Üniversitesi Uluslararası Bilgisayar Enstitüsü, YL Tezi, 2011.

[16] S. Özçelik, Ö. Poyraz, K. Kalkan, E. Malatyalı, and S. Değerli, "The investigation of Cryptosporidium spp. prevalence in cattle and farmers by ELISA," Kafkas Üniversitesi Veteriner Fakültesi Dergisi, vol. 18, 2012.

[17] J.K. Griffiths, "Human cryptosporidiosis: epidemiology, transmission, clinical disease, treatment, and diagnosis," in Advances in parasitology. vol. 40, ed: Elsevier, 1998, pp. 37-85.

[18] D. Miron, J. Kenes, and R. Dagan, "Calves as a source of an outbreak of cryptosporidiosis among young children in an agricultural closed community," The Pediatric infectious disease journal, vol. 10, pp. 438-441, 1991.

[19] G. Börekçi, F. Otağ, and G. Emekdaş, "Mersin'de bir gecekondu mahallesinde yaşayan ailelerde Cryptosporidium prevalansı," İnfeksiyon Derg, vol. 19, pp. 39-46, 2005.

[20] Z. Egyed, T. Sreter, Z. Szell, and I. Varga, "Characterization of Cryptosporidium spp.—recent developments and future needs," Veterinary parasitology, vol. 111, pp. 103-114, 2003.

[21] M. Kayri and Ö. Çokluk, "Examining Factors of Academic Procrastination Tendency of University Students by using Artificial Neural Network," International Journal of Computer Trends and Technology, vol. 34, pp. 1-8, 2016.

[22] L. Nanni, C. Fantozzi, and N. Lazzarini, "Coupling different methods for overcoming the class imbalance problem," Neurocomputing, vol. 158, pp. 48-61, 2015.

[23] A. Sarmanova and S. Albayrak, "Alleviating class imbalance problem in data mining," in 2013 21st Signal Processing and Communications Applications Conference (SIU), 2013, pp. 1-4.

.

BIOGRAPHIES

İpek BALIKÇI ÇİÇEK obtained her BSc. degree in mathematics from Çukurova University in 2010. She received MSc. degree in biostatistics and medical informatics from the Inonu University in 2018. She currently continues Ph.D. degrees in biostatistics and medical informatics from the Inonu University. In 2014, she joined the Department of Biostatistics and Medical Informatics at Inonu University as a researcher assistant. Her research interests are cognitive systems, data mining, machine learning, deep learning.

Ülkü KARAMAN she is Associate professor Ordu University Faculty of Medicine, Department of Medical Parasitology, Ordu, Turkey. She has copleted he Ph.D. from Department of Medical Parasitology. Her area of research interest is medical parasitology and microbiolog. He has about 22 years working in health science.