



THE JOURNAL OF

COGNITIVE SYSTEMS

an international, peer-reviewed  
indexed, and open-access periodical

VOLUME 05  
NUMBER 01  
YEAR 2020  
ISSN 2548-0650

[www.dergipark.gov.tr/jcs](http://www.dergipark.gov.tr/jcs)





# THE JOURNAL OF COGNITIVE SYSTEMS

An international, peer-reviewed,  
indexed, and open-access periodical

ISSN 2548-0650

## OWNER

ISMAIL KOYUNCU, (RECTOR), Istanbul Technical University, Turkey

## GENERAL PUBLICATION DIRECTOR

S. SERHAT SEKER, Istanbul Technical University, Turkey

## EDITOR-IN-CHIEF

OZER CIFTCIOGLU, Delft University of Technology, Delft, The Netherlands

## EDITORIAL BOARD

T. CETIN AKINCI, Istanbul Technical University, Turkey

## SCIENTIFIC COMMITTEE

ABRAHAM LOMI (Indonesia)  
ALTAN CAKIR (Turkey)  
A.TARIK ZENGİN (Turkey)  
AYKUT HOCANIN (Turkish Republic of Northern Cyprus)  
BELLE R. UPADHYAYA (USA)  
BERK USTUNDAG (Turkey)  
BURAK BARUTCU (Turkey)  
BURCU OZDEMIR (Turkey)  
CEMIL COLAK (Turkey)  
CHANDAN KUMAR CHANDA (India)  
DENİZ TURKPENCE (Turkey)  
ERKAN KAPLANOGLU (USA)  
ESEN YILDIRIM (Turkey)  
GOKHAN ERDEMİR (Turkey)  
HAKAN TEMELTAS (Turkey)  
HASAN DEMIREL (Turkish Republic of Northern Cyprus)  
JELENA DIKUN (Lithuania)  
KUNIHICO NABESHIMA (Japan)  
MURAT OKATAN (Turkey)  
MUSA YILMAZ (Turkey)  
NECDET OSAM (Turkish Republic of Northern Cyprus)  
OKYAY KAYNAK (Turkey)  
OLEKSII TURUTA (Ukraine)  
OSMAN NURI UCAN (Turkey)  
OMER FARUK ERTUGRUL (Turkey)  
RITUPARNA DATTA (South Korea)  
SALIH BARIS OZTURL (Turkey)  
TANJU SURMELI (Turkey)  
UFUK KORKMAZ (Turkey)

## AIM & SCOPE

The Journal publishes original papers in the field of artificial intelligence, biomedical, quantum information, big data analysis and statistical areas, which are related with cognitive science and engineering, as well as the applications of the cognitive studies in social areas. Letter to the editor is also encouraged.

## THE JOURNAL OF COGNITIVE SYSTEMS

THE JOURNAL OF COGNITIVE SYSTEMS (JCS) is published bi-annually. Responsibility for the content rests upon the authors and not upon the JCS or the publisher. **Reuse Rights and Reprint Permissions:** Educational or personal use of this material is permitted without fee, provided such use: i) is not made for profit; and ii) includes this notice and a full citation to the original work on the first page of the copy; and iii) does not imply JCS endorsement of any third-party products or services. Authors and their companies are permitted to post their JCS-copyrighted material on their own web servers without permission, provided that the JCS copyright notice and a full citation to the original work appear on the first screen of the posted copy. Permission to reprint/republish this material for commercial, advertising or promotional purposes or for creating new collective works for resale or redistribution must be obtained from JCS by writing to the JCS General Publication Director Prof. Dr. S. Serhat Seker, Faculty of Electrical and Electronics Engineering, Istanbul Technical University, Istanbul, Turkey; or [sekers@itu.edu.tr](mailto:sekers@itu.edu.tr).  
Copyright © 2016 JCS. All rights reserved.

## CONTACT

Prof. Dr. Özer Ciftcioglu  
Editor-in-Chief of The Journal of Cognitive Systems  
Delft University of Technology, The Netherlands  
Istanbul Technical University, Istanbul, Turkey

## E-MAIL

[cognitive@itu.edu.tr](mailto:cognitive@itu.edu.tr)

## WEB PAGE

[www.cognitive.itu.edu.tr](http://www.cognitive.itu.edu.tr)

## URL

[www.dergipark.gov.tr/jcs](http://www.dergipark.gov.tr/jcs)





# THE JOURNAL OF COGNITIVE SYSTEMS

VOLUME 05, NUMBER 01

J U N E 2 0 2 0

## CONTENTS

<b>A. Acet, and E. Akkaya</b> : A Deep Learning Image Classification Using Tensorflow for Optical Aviation Systems,.....	01-04
<b>E. Guldogan, Z. Tunc, A. Acet, and C. Colak</b> : Performance Evaluation of Different Artificial Neural Network Models in the Classification of Type 2 Diabets Mellitus,.....	05-09
<b>E. Guldogan, Z. Tunc, and C. Colak</b> : Classification of Breast Cancer and Determination of Related Factors with Deep Learning Approach,.....	10-14
<b>M. Kivrak, F. B. Akcesme, and C. Colak</b> : Sample Size Effect on Classification Performance of Machine Learning Models: An Application of Coronary Artery Disease,.....	15-18
<b>M. Kivrak, F. B. Akcesme, and C. Colak</b> : Evaluation of Association Rules Based on Certainty Factor: An Application on Diabetes Data Set,.....	19-22
<b>H. S. Nogay</b> : Prediction of Post-Treatment Survival Expectancy in Head & Neck Cancers by Machine Learning Methods,.....	23-32
<b>A.K. Arslan, Z. Tunc, I. Balikci Cicek, and C. Colak</b> : A Novel Interpretable Web-Based Tool on the Associative Classification Methods: An Application on Breast Cancer Dataset, .....	33-40

# A DEEP LEARNING IMAGE CLASSIFICATION USING TENSORFLOW FOR OPTICAL AVIATION SYSTEMS

A. Acet and A. E. Akkaya

**Abstract**— Deep learning has become very popular in recent years. Great progress has been made in the task of classifying images with the development of deep learning. This research utilized the deep learning methods in TensorFlow to classify the bird and airplane images. In the first step, a general framework for the classification of deep learning images, an image classification network namely airplane images and bird images are built. Then, the images were randomly chosen from the Caltech-UCSD Birds-200-2011 and Caltech 101 datasets. To correctly classify airplane and bird images, a total of 1600 images used. The 1072 images used to train the network and the 528 images used to test built deep learning network. The training phase lasts only 20 epochs to achieve 100% accuracy on the train set. The test data were classified as 99.05%. Overall accuracy is 99.69%. This research has a certain importance to explore the use of cognitive systems approach in aviation safety.

**Keywords**— *Deep learning, TensorFlow, CUDA, Image classification.*

## 1. INTRODUCTION

IN RECENT years deep learning has become a hot topic of research. In the area of artificial intelligence, image recognition, pattern recognition and autonomous driving deep learning have made significant progress. Deep learning has some benefits. These networks are independent of artificial features. Also, deep learning systems can learn adaptively. Thus the algorithm's reliability and versatility are greater than the conventional approach of image processing. Convolutional Neural Networks (CNNs) are a special type of deep learning models widely used in areas such as image classification [1] and natural language processing [2]. Before deep learning architectures, conventional artificial neural networks were used to determine what features an image contains. To achieve this, the image is transformed into a column or row vector and given to the system input. In this type of structure, many features are composed of side-by-side pixel values. However, the human perception system looks at corners, lines and rounded shapes to perceive an image. When an image is flattened and brought into a row or column vector, all the fine details disappear. It is almost impossible to try to solve this structure, which is difficult to perceive even by perfectly functioning human intelligence, using machine learning techniques. CNNs has been developed as a solution to this problem. CNNs use feedforward structure. Unlike conventional artificial neural networks, it has convolution and pooling layers for feature extraction and reducing the size of the input image, respectively. Using both layers, important features in the image can be extracted.

**Ayça ACET.** Inonu University, Malatya, Turkey, (a.aycaacet@gmail.com)

**Abdullah Erhan AKKAYA.** Inonu University, Malatya, Turkey, (abdullahakkaya@gmail.com) 

Manuscript received May 12, 2020; accepted May 29, 2020.  
Digital Object Identifier:

This research utilized the deep learning methods in TensorFlow [3] to classify the bird and airplane images. Firstly, the residual network, a general framework for the classification of deep learning images, an image classification network namely airplane images and bird images are built. Then, the Caltech-UCSD Birds-200-2011 [4] and Caltech 101 [5] datasets are used to train and validate the neural network. To correctly classify airplane and bird images, total of 1600 images consists of birds and airplanes used. The split rate of the network is 0.33. The 67% of images (1072 images) used to train the network. The %33 of the images (528 images) used to validate the network. The training phase lasts only 20 epochs to achieve 100% correctness and the test data were classified as 99.4%.. Although they have similar structures, aircraft and bird images have been successfully distinguished from each other. This research has a certain importance to explore the use of cognitive systems approach in aviation safety.

## 2. DATASET

Caltech-UCSD Birds-200-2011 [4] and Caltech 101 [5] datasets used for train and test steps. Total of 1600 images is chosen from the dataset that has various sizes. These images are labelled as airplane image, and the bird image (Figure 1). All dataset images are divided into two sets. The training set consists of 1072 images and the test set consists of 528 images. Firstly, to train the weights of the built CNN, a training set is used. The validation set is randomly chosen from the set to confirm the model's generalization ability. Therefore, the model weights are selected to be saved on the validation data according to the value of the loss function. Finally, at the test step, the saved weights of the trained model are used to give decisions about the class on the test set.

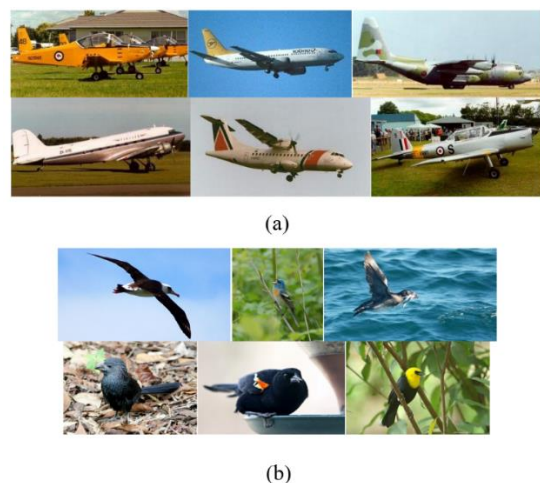


Fig.1. Sample airplane (a) and bird (b) images used for a classification task.

### 3. TOOLS AND NETWORK MODEL

#### 3.1. Development Environment

In this study, CNN using TensorFlow application was performed on GPU using Python language in Anaconda environment. The calculations on the GPU have been performed via the Nvidia CUDA library. CUDA acceleration package and *cudaconv* are installed, and then some related OpenCV-Python, TensorFlow, Scikit-learn and Keras packages are installed. Proposed network has been trained on the Nvidia GeForce GTX 1070 8GB 256-bit graphic card that has 1920 CUDA cores.

Python programming language provides an easy way to solve the problems with coding flexibility [6]. Python is a platform-independent, less time-consuming scripting language [6]. It is a high level interpreted. Many developers use Python for developing easy to code programs. ABC language was popular but shortly after arriving on the market Python took its crown. Python language is a dynamic language and it has garbage collection mechanism.

The Scikit-learn library is widely using for modelling with data mining and data analysis. The Scikit-learn, NumPy, matplotlib, and SciPy libraries contain simple tools for machine learning classification, regression and clustering tasks [7]. Supervised learning algorithms, non-supervised learning algorithms, feature extraction and cross-validation are some features of this library.

Sample Python code for proposed CNN is given below.

```
model.add(Convolution2D(kernel_size=3, strides=1,
filters=32, padding='same', activation='relu', name='conv1',
input_shape=input_shape))

model.add(MaxPooling2D(pool_size=2, strides=2))
...
model.add(Dense(num_classes, activation='softmax'))

model.add(Dropout(0.8))

model.add(Dense(num_classes, activation='softmax'))

optimizer = adam(lr=0.001, beta_1=0.9, beta_2=0.999,
epsilon=1e-08, decay=0.0)

model.compile(optimizer=optimizer,
loss='categorical_crossentropy', metrics=['accuracy'])

hist = model.fit(X_train, y_train, batch_size=64,
epochs=num_epoch, verbose=1, validation_data=(X_test,
y_test), shuffle=True)
```

#### 3.2. Convolutional Neural Networks (CNNs)

CNNs are feedforward neural networks. In feedforward neural networks, the signal flows over a network without loops [8]. A typical CNN model consists of convolutional, activation, pooling, fully connected, and output layers. The convolutional layer has a function composed of multiple convolutional kernels. Each kernel symbolizes a linear function in matching kernel [8]. In the pooling layer, a layer by layer down-sampling non-linear function used for aiming at reducing progressively the size of the feature representation. A fully

connected layer can be considered a type of convolutional layer. The kernel size of fully connected layers is  $1 \times 1$ . To compute the probabilities of input image belonging to which classes, the output or prediction layer is often used at the last fully connected layer.

The proposed CNN consists of eight layers. The first, third and fifth layers are the convolution layers where the basic features of the image detected. The second, fourth and sixth layers are the pooling layers, which reduces the image size by half. In convolution layers, more detailed information about the image was tried to be obtained by using 32,64 and 128 filters of size  $3 \times 3$  in first, third and fifth layers respectively. The seventh layer is a fully connected (dense) layer in which all neurons are connected together. Immediately after this layer, unnecessary nerve cells were deleted using a 0.5 dropout value in order to prevent overfitting problem. The eight and last layer determines which results will be included in the class of airplanes or birds from the previous layers. One of the activation functions the Rectified Linear Unit (ReLU) [9] is used in the first, third, and fifth layers of the convolution layers. The ReLU function used in conjunction with other layers. ReLU activation function has become the most commonly used in deep learning networks and more popular than logistic sigmoid and hyperbolic tangent functions [10]. In this study, ReLU function used as an activation function. The ReLU is, where  $x$  is the input value to the ReLU function, formulated as follows,

$$f(x) = \max(0, x) \quad (1)$$

It truncates all negative values input to zero. Only half of the ReLUs is activated when used in combination with a batch normalization layer. Therefore at a given time, it results in sparse activations. The Softmax activation function used in the eight-layer.

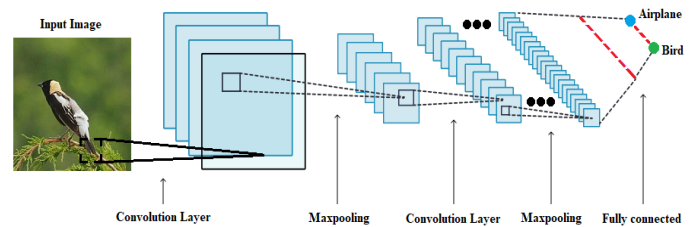


Fig.2. Framework of proposed CNN.

TABLE I  
PARAMETERS OF PROPOSED CNN

Layer name	Layer Parameters
Input Image	128x128 pixels
Conv1	Kernel size=3, strides=1, filters=32x32, ReLU
Maxpooling	Pool size=2, strides=2
Conv2	Kernel size=3, strides=1, filters=64x64, ReLU
Maxpooling	Pool size=2, strides=2
Conv3	Kernel size=3, strides=1, filters=128x128, ReLU
Maxpooling	Pool size=2, strides=2
Dense layer	Neuron size=128, ReLU, Dropout=0.8
Output	Average pooling, fully connected 2 class, Softmax

### 3.3. Categorical Cross-Entropy Loss Function

Categorical Cross-Entropy Loss Function, also called Softmax Loss function, consists of Softmax activation and Cross-Entropy loss functions. When this function used, CNN has trained to output a probability over the classes for each image. Equation (2) and Equation (3) shows Softmax output and Categorical Cross-Entropy loss functions.

$$f(S)_i = \frac{e^{s_i}}{\sum_j^c e^{s_j}} \tag{2}$$

$$CE = - \sum_i^c t_i \log(f(s)_i) \tag{3}$$

## 4. TRAIN AND TEST CNN

### 4.1. Train the CNN Architecture

Adam [11] is selected as an optimization algorithm. The initial learning rate of Adam algorithm is set to 0.001. The batch size is chosen 64 for the Nvidia GTX1070 graphic card. The network model has trained 20 epochs. At the training step, data augmentation is used by the function generator in Keras. The model checkpoint class is used to select optimal model weights with respect to the validation loss value. Figure 3 shows train loss and validation loss values for 20 epochs. Figure 4 show train accuracy and test accuracy curves.

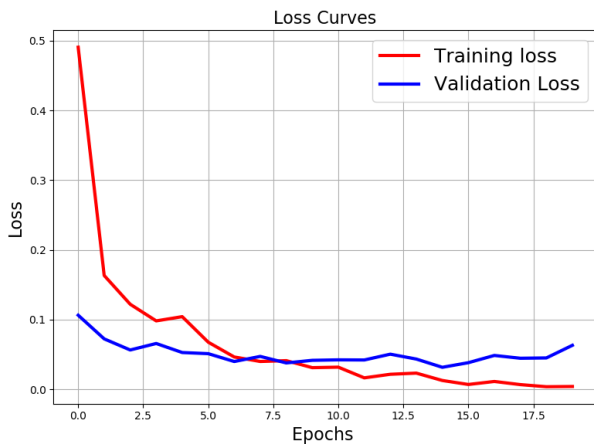


Fig.3. The train and validation loss curves of proposed CNN.

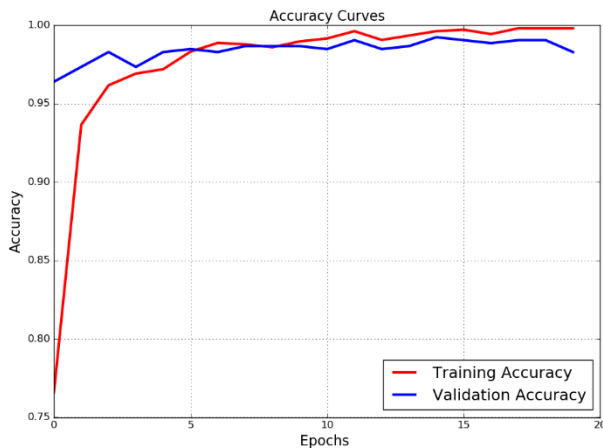


Fig.4. The train and validation accuracy curves of proposed CNN.

### 4.2. Test Result

After the CNN trained, the generalization ability of the trained model should be evaluated. To achieve this the test set which consists of 240 airplane images and 288 bird images is used. For the evaluation of the proposed CNN, we assumed the class containing airplane images as the positive class, and therefore the accuracy, sensitivity, and specificity are computed with reference to this. We report the definitions for these parameters, just for clarity.

- True Positive (TP) is the number of airplane images correctly classified;
- True Negative (TN) is the number of bird images correctly classified;
- False Positive (FP) is the number of bird images incorrectly classified as airplane image;
- False Negative (FN) is the number of airplane images incorrectly classified as bird image.

$$Accuracy = \frac{TP + TN}{(TP + TN) + (FP + FN)} = 1 - error$$

$$Sensitivity = \frac{TP}{TP + FN}$$

$$Specificity = \frac{TN}{TN + FP}$$

$$Precision = \frac{TP}{TP + FP}$$

$$False\ positive\ rate = \frac{FP}{TN + FP} = 1 - specificity$$

Table II shows true and predicted labels after the prediction step on the test set. The sensitivity and specificity are two indexes generally used to evaluate the performance of the classifier. Table II is also called the confusion matrix.

TABLE II  
TEST RESULTS

		Predicted label	
		Airplane (Positive)	Bird (Negative)
True label	Airplane (True)	237 (True Positive)	3 (False Negative)
	Bird (False)	2 (False Positive)	286 (True Negative)

Error rate and accuracy are can be obtained from the confusion matrix. The error rate is calculated by dividing the number of all false estimates by the total number of the data set. The best worst error rate is 1, the best is 0. Accuracy is calculated by dividing the number of all correct estimates by the total number of the data set. The worst accuracy is 0, the best is 1. It can be calculated with (1-error rate). Sensitivity is



calculated by dividing the number of true positive predictions by the total positive number. This is called a recall or true positive rate (TPR). The worst sensitivity is 0, the best is 1. Specificity is calculated by dividing the number of correct estimates by the total number of negatives. This is also called true negative rate (TNR). The worst specificity is 0, the best is 1. Sensitivity is calculated by dividing the number of true positive predictions by the total positive predictions. This is called a positive predictive value (PPV). The worst precision is 0, the best is 1. The false-positive ratio is calculated by dividing the number of false-positive predictions by the total number of negatives. The worst false positive rate is 1, the best is 0. It can also be calculated as  $(1 - \text{specificity})$ .

Table III shows the discrimination ability of the proposed convolutional neural network overtraining and testing steps. Also, whole datasets are shown in the table in terms of accuracy, sensitivity and specificity. The whole dataset split by 0.33 split ratio.

TABLE III  
DISCRIMINATING ABILITY OF THE PROPOSED CNN CLASSIFIER

	Training Set	Testing Set	Whole Data
Accuracy	100%	99.05%	99.69%
Sensitivity	100%	98.75%	99.59%
Specificity	100%	99.30%	99.77%
Precision	100%	99.16%	99.72%

Figure 5 is an example for wrongly classified and labelled airplane image with 97.42% probability. This image is a bird image and for the human optical system, it has different properties. CNN decreases this image size at every max-pooling step. If we zoom out enough we see it looks like a plane because of the wingspan.



Fig.5. Misclassification of bird image at proposed CNN.

## 5. CONCLUSIONS

In this paper, based on binary image classification, airplane and bird classification model is built. Then, equally chosen 1600 airplane and bird images are divided into two sets, train and test. The optimal model weight selected is done by the loss value of the validation set. Finally, the accuracy of the trained model has 99.16% precision and 98.75% sensitivity ratios. This study valuable for exploring the application of deep learning method in optical early bird detection systems for aviation.

## ACKNOWLEDGMENT

The study is selected from Congress Proceedings Abstract Book, International Conference on STEM And Educational Sciences, Muş Alparslan University, May, 2018.

## REFERENCES

- [1] D. Ciresan, U. Meier, and J. Schmidhuber, "Multi-Column Deep Neural Networks for Image Classification," Technical Report, arXiv:1202.2745, 2012.
- [2] R. Collobert and J. Weston, "A Unified Architecture for Natural Language Processing: Deep Neural Networks with Multitask Learning," Proc. Int'l Conf. Machine Learning, 2008.
- [3] Martín Abadi et. al., "TensorFlow: Large-scale machine learning on heterogeneous systems", 2015. Software available from tensorflow.org.
- [4] Wah C., Branson S., Welinder P., Perona P., Belongie S. "The Caltech-UCSD Birds-200-2011 Dataset." Computation & Neural Systems Technical Report, CNS-TR-2011-001.
- [5] L. Fei-Fei, R. Fergus and P. Perona. Learning generative visual models from few training examples: an incremental Bayesian approach tested on 101 object categories. IEEE. CVPR 2004, Workshop on Generative-Model Based Vision. 2004.
- [6] AlizaSarlan, ChayanitNadam, ShuibBasri , "Twitter Sentiment Analysis".
- [7] KUNAL JAIN, 2015, "Scikit-learn in Python – the most important Machine Learning tool I learnt last year!". Online <https://www.analyticsvidhya.com/blog/2015/01/scikit-learn-python-machine-learning-tool/> , Accessed:2020-04-29.
- [8] B. Xu, N. Wang, T. Chen, M. Li, Empirical evaluation of rectified activations in convolutional network, arXiv preprint arXiv:1505.00853 , (2015).
- [9] X. Glorot, A. Bordes, and Y. Bengio, "Deep sparse rectifier neural networks," in Proceedings of the International Conference on Artificial Intelligence and Statistics, pp. 315–323, 2011.
- [10] A. Karpathy, "Commonly used activation functions." <http://cs231n.github.io/linear-classify/#loss>. Accessed:2020-04-26.
- [11] D. Kingma, J. Ba, Adam: a method for stochastic optimization, arXiv preprint arXiv:1412.6980 , (2014).

## BIOGRAPHIES

**Ayca Acet** received the BSc degree in Computer Engineering from TOBB University of Economics and Technology, Ankara, Turkey, in 2017. She currently continues M.S. degree in Computer Engineering, at Inonu University, Malatya, Turkey. Her general research interests include machine learning methods, deep learning and data mining.

**Abdullah Erhan Akkaya** received the BSc, M.S. and Ph.D. degrees, in Computer Engineering from Firat University, Elazig, Turkey, in 2007, Erciyes University, Kayseri, Turkey, in 2012 and Inonu University, Malatya, Turkey in 2017 respectively. He is currently an Assistant Professor of Computer Engineering, at Inonu University, Malatya, Turkey. His general research interests include machine learning methods, image processing, sensor fusion and cognitive systems.

# PERFORMANCE EVALUATION OF DIFFERENT ARTIFICIAL NEURAL NETWORK MODELS IN THE CLASSIFICATION OF TYPE 2 DIABETES MELLITUS

E. Guldogan, Z. Tunc, A. Acet and C. Colak


**Abstract**— Objective: In this study, it is aimed to classify type 2 Diabetes Mellitus (DM), compare the estimates of the Artificial Neural Network models and determine the factors related to the disease by applying Multilayer Perceptron (MLP) and Radial Based Function (RBF) methods on the open-access dataset.


**Material and Methods:** In this study, the data set named “Pima Indians Diabetes Database” was obtained from <https://www.kaggle.com/uciml/pima-indians-diabetes-database>. The dataset contains 768 records with 268 (34.9%) type 2 diabetes patients and 500 (65.1%) people without diabetes, which have 9 variables (8 inputs and 1 outcome). MLP and RBF methods, which are artificial neural network models, were used to classify type 2 DM. Factors associated with type 2 DM were estimated by using artificial neural network models.


**Results:** The performance values obtained with MLP from the applied models were accuracy 78.1%, specificity 81.2%, AUC 0.848, sensitivity 71%, positive predictive value 61.7%, negative predictive value 86.8% and F-score 66%. In relation to RBF model, the performance metrics were accuracy obtained 76.8%, specificity 82.1%, AUC 0.813, sensitivity 66.0%, positive predictive value 64.6%, negative predictive value 83% and F-score 65.3%, respectively. When the effects of the variables in the data set examined in this study on Type 2 DM are analyzed; The three most important variables for the MLP model were obtained as Glucose, BMI, Pregnancies respectively. For RBF, it was obtained as Glucose, Skin Thickness, and Insulin.


**Conclusion:** The findings obtained from this study showed that the models used gave successful predictions for Type 2 DM classification. Besides, unlike similar studies examining the same dataset, the significance values of the factors associated with the models created were estimated.

**Keywords**—Classification, Multilayer perceptron neural network, Radial-based function neural network, Type 2 Diabetes Mellitus.

**Emek GULDOGAN**, Inonu University Department of Biostatistics and Medical Informatics, Faculty of Medicine, Malatya, Turkey, (emek.guldogan@inonu.edu.tr) 

**Zeynep TUNC**, Inonu University Department of Biostatistics and Medical Informatics, Faculty of Medicine, Malatya, Turkey, (zeynep.tunc@inonu.edu.tr) 

**Ayça ACET**, Inonu University, Department of Computer Engineering, Faculty of Engineering, Malatya, Turkey, (a.ayceacet@gmail.com) 

**Cemil COLAK**, Inonu University Department of Biostatistics and Medical Informatics, Faculty of Medicine, Malatya, Turkey, (cemil.colak@inonu.edu.tr) 

Manuscript received May 08, 2020; accepted May 29, 2020.  
Digital Object Identifier:

## 1. INTRODUCTION

**D**IABETES mellitus (DM) is a chronic disease that seriously affects both daily life and quality of life. This disease cannot be cured completely, but when it is managed well and precautions are taken, its negative effects in the short and long term can be minimized [1, 2]. DM is a chronic and metabolic disease characterized by abnormalities in protein, carbohydrate, and fat metabolism caused by absolute or relative insulin deficiency and accompanying clinical, biochemical findings [3]. Type 1 DM is an autoimmune disease and is caused by the destruction of pancreatic beta cells. Type 2 DM is defined as the combination of insulin resistance and impairment of pancreatic beta cells in insulin secretion [4].

Type 2 DM is a heterogeneous disorder caused by a large number of genetic and environmental factors. Although the pathogenesis of type 2 diabetes is quite complex, it is characterized by two main pathophysiological causes. A decrease in insulin sensitivity or insulin resistance.

Dysfunction of pancreatic beta cells in addition to relative insulin deficiency (insulin secretion defect) [5].

Type 2 DM accounts for 80-90% of all diabetes cases. The frequency of the disease is increasing gradually all over the world. The prevalence of type 2 DM increases with age. Factors such as the transition from traditionally accepted lifestyle to western lifestyle, the increase in the number of overweight and obese individuals, decrease in activities such as exercise and sports, and unhealthy diet contributed to the prevalence of the disease [6].

Artificial Neural Networks (ANNs) are computer systems developed to directly realize the features of learning, which is one of the features of the human brain, such as the ability to derive, create and discover new information without any help [7]. ANN can provide nonlinear modeling without needing any prior knowledge between input and output variables, without any assumptions [8]. Artificial neural networks are a successful method in solving many daily life problems such as classification, modeling, and prediction. Multilayer Perceptron (MLP) is a frequently used ANN model for the solution of nonlinear problems. It is a feed-forward, backpropagation network using at least one layer between the input and output layers consisting of at least three layers [9]. In the forward propagation stage, while calculating the output and error value of the network, the link weight values between the layers are updated to minimize the calculated error value in the reverse propagation stage [10].

Radial-based function (RBF) neural networks are feed-forward networks consisting of a 3-layer structure: an input



layer, an output layer, and a single hidden layer. This hidden layer is the layer using radial functions that give the network its name as a transfer function. While the inputs of this network are not linear, the output is linear [11]. The input layer consists of source nodes and provides the connection of the network with the environment. The only hidden layer in the network makes a nonlinear transformation from the input area to the hidden area. The conversion from the input layer to the hidden layer is a nonlinear constant transformation with radial-based transfer functions. The output layer is linear and is the layer that responds to the network, which is the transfer signal applied to the input layer. An adaptive and linear transformation is performed from the hidden layer to the output layer [12].

In this study, it is aimed to compare the classification performance of Type 2 DM and to determine the risk factors that may be associated with Type 2 DM by applying MLP and RBF models on the open-access Type 2 DM data set.

## 2. MATERIAL AND METHODS

### 2.1. Dataset

In this study, the data set named “Pima Indians Diabetes Database” was obtained from <https://www.kaggle.com/uciml/pima-indians-diabetes-database> to examine the working principle of MLP and RBF ANN models and to determine risk factors. In the data set used, there were 768 individuals with 268 (34.9%) type 2 diabetes patients and 500 (65.1%) people without diabetes. DM was described as a concentration of plasma glucose greater than 200 mg/dl two hours after ingestion of a carbohydrate solution with 75 gm. All the subjects were females and  $\geq 21$  years old at the time of the index examination [13, 14]. Explanations about the variables in the data set and their properties are given in Table 1.

TABLE I  
EXPLANATIONS ABOUT THE VARIABLES IN THE DATASET AND THEIR PROPERTIES

Variable	Variable Description	Variable Type	Variable Role
Pregnancies	Number of pregnancies	Quantitative	Input
Glucose	2-hour plasma glucose concentration in the oral glucose tolerance test	Quantitative	Input
Blood Pressure (BP)	Diastolic blood pressure (mmHg)	Quantitative	Input
Skin Thickness (ST)	Triceps skinfold thickness (mm)	Quantitative	Input
Insulin	2-hour serum insulin (mu U / ml)	Quantitative	Input
BMI	Body mass index [weight in kg / (height in m) <sup>2</sup> ]	Quantitative	Input
Diabetes Pedigree Function (DPF)	Diabetes family tree function	Quantitative	Input
Age	Age (years)	Quantitative	Input
Outcome	Class variable (type 2 DM; 0: absent, 1: present)	Qualitative	Output

## 3. ARTIFICIAL NEURAL NETWORK MODELS

In this study, classification performance was compared by applying MLP and RBF methods on artificial neural network models on Type 2 DM data set and risk factors that may be associated with Type 2 DM were determined. Because of its power, flexibility, and ease of use, artificial neural networks are the preferred tool for many predictive data mining applications. Predictive neural networks are particularly useful in applications where the mechanism underlying them is complex. In recent years, interest in the application of neural networks has increased for problems that cannot be solved with classical techniques, and ANN has been used successfully in many medical applications. Unlike traditional spectral analysis methods, artificial neural networks not only model signals but also produce solutions for the classification of signals. Another advantage of artificial neural networks compared to the methods available for the analysis of biomedical signals is that after their training, they are very fast. MLP is a nonparametric artificial neural network technique that performs many detection and prediction operations [15].

Radial Based Function Neural Network (RBFNN) was developed in 1988 inspired by the effect response behaviors seen in biological nerve cells and entered the history of ANN by applying it to the filtering problem. It is possible to see the training of RBFNN models as a curve-fitting approach in multi-dimensional space. For this reason, the training performance of the RBFNN model turns into an interpolation problem, finding the most suitable surface for the data in the output vector space. RBFNN models are defined in three layers as the input layer, hidden layer, and output layer, similar to general ANN architecture. However, unlike conventional ANN structures, RBFNNs use radial-based activation functions and nonlinear cluster analysis in the transition from the input layer to the hidden layer. The structure between the hidden layer and the output layer continues to function as in other ANN types [16].

In the construction of MLP and RBF models, nearly 60% and 40% of the whole dataset were used for training and testing stages, respectively. Rescaling method for the variables was standardized for both models, the number of units in the hidden layer was 6 for MLP and 5 for RBF, hidden layer activation function was hyperbolic tangent for MLP and softmax for RBF, the number of units in output layer was 2 for both models, and output layer activation function was softmax for MLP and identity for RBF. Hyperparameters of the models were optimized by the scaled conjugate gradient method.

### 3.1. Performance Evaluation of the Models

In the performance evaluation of MLP and RBF artificial neural network models for predicting the factors that may be associated with type 2 DM, different metrics that can be calculated from the values in the confusion matrix (Table 2) given below have been obtained.

TABLE II  
CONFUSION MATRIX FOR CALCULATING PERFORMANCE METRICS

		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+F N

The metrics considered in the performance evaluation of the models in this study are given below.

$$\text{Accuracy} = (TP+TN)/(TP+TN+FP+FN)$$

$$\text{Sensitivity} = TP/(TP+FP)$$

$$\text{Specificity} = TN/(TN+FN)$$

$$\text{Positive predictive value} = TP/(TP+FN)$$

$$\text{Negative predictive value} = TN/(TN+FP)$$

$$\text{F-score} = (2*TP)/(2*TP+FP+FN)$$

#### 4. DATA ANALYSIS

Quantitative data are summarized by median (minimum-maximum) and qualitative variables are given by number and percentage. Normal distribution was evaluated with the Kolmogorov-Smirnov test. Whether there is a statistically significant difference between categories of the dependent variable in terms of input variables, the Mann-Whitney U test was used for the analyses.  $P < 0.05$  values were considered statistically significant. In all analyses, IBM SPSS Statistics 26.0 for the Windows package program was used.

#### 5. RESULTS

Descriptive statistics related to the target variable examined in this study are presented in Table 3. There is a statistically significant difference between the dependent variable classes in terms of other variables other than the insulin variable.

In this study, descriptive statistics of the factors examined according to the type 2 DM variable are summarized in Table 3. According to these findings, while there was a difference between the presence and absence of type 2 DM in terms of Pregnancies, Glucose, BP, ST, BMI, DPF and Age ( $p < 0.05$ ), no statistically significant difference was found for the Insulin factor ( $p > 0.05$ ).

TABLE III  
DESCRIPTIVE STATISTICS ABOUT INPUT AND OUTPUT VARIABLES

Variable	Outcome (Diabetes Mellitus)		p value
	Absent (n=500)	Present (n=268)	
Statistics	Median (Min-Max)	Median (Min-Max)	
Pregnancies	2 (0-13)	4 (0-17)	<0.001
Glucose	107 (0-197)	140 (0-199)	<0.001
BP	70 (0-122)	74 (0-114)	<0.001
ST	21 (0-60)	27 (0-99)	0.013
Insulin	39 (0-744)	0 (0-846)	0.066
BMI	30.1 (0-57,3)	34.3 (0-67.1)	<0.001
DPF	0.336 (0.078-2.329)	0.449 (0.088-2.42)	<0.001
Age	27 (21-81)	36 (21-70)	<0.001

BMI: Body mass index; ST: Skin Thickness; DPF: Diabetes Pedigree Function; BP: Blood pressure;

Classification matrices of the testing stages for MLP and RBF models are shown in Tables 4 and 5, respectively.

TABLE IV  
CLASSIFICATION MATRIX OF THE TESTING STAGE FOR THE MLP MODEL

Predicted \ Real	Present	Absent	Total
	Present	66	41
Absent	27	177	204
Total	93	218	311

TABLE V  
CLASSIFICATION MATRIX OF THE TESTING STAGE FOR THE RBF MODEL

Predicted \ Real	Present	Absent	Total
	Present	64	35
Absent	33	161	194
Total	97	196	293

Table 6 presents the performance criteria values calculated from the models to classify type 2 DM in the testing stage.

TABLE VI  
PERFORMANCE METRIC VALUES CALCULATED FROM THE GENERATED MODELS  
IN THE TESTING STAGE

Metric \ Model	ANN type	
	MLP	RBF
Accuracy (%)	78.1	76.8
Specificity (%)	81.2	82.1
AUC	0.848	0.813
Sensitivity (%)	71.0	66.0
Positive predictive value (%)	61.7	64.6
Negative predictive value (%)	86.8	83
F-score (%)	66.0	65.3

AUC: Area under the ROC curve; MLP: Multilayer perceptron neural network; RBF: Radial-based function neural network

The values related to performance criteria obtained from MLP and RBF models are demonstrated in Figure 1.

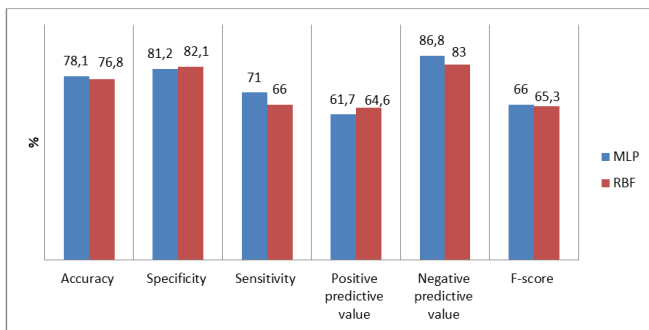


Fig.1. Performance criteria values obtained from MLP and RBF models in the testing stage (MLP: Multilayer perceptron; RBF: Radial-based function)

TABLE VII  
IMPORTANCE VALUES OF EXPLANATORY VARIABLES ACCORDING TO MLP AND RBF MODELS

Explanatory Variables	MLP	RBF
Glucose	0.287	0.175
BMI	0.219	0.144
Pregnancies	0.134	0.074
DPF	0.125	0.135
Age	0.077	0.068
Insulin	0.072	0.159
BP	0.057	0.078
ST	0.03	0.167
Total	100.00	100.00

BMI: Body mass index; ST: Skin Thickness; DPF: Diabetes Pedigree Function; BP: Blood pressure; MLP: Multilayer perceptron; RBF: Radial-based function

In this study, the importance values of the factors related to diabetes mellitus are given in Table 7, while the values for these importance percentages are shown in Figure 2. Table 7: Importance values of explanatory variables according to MLP and RBF models

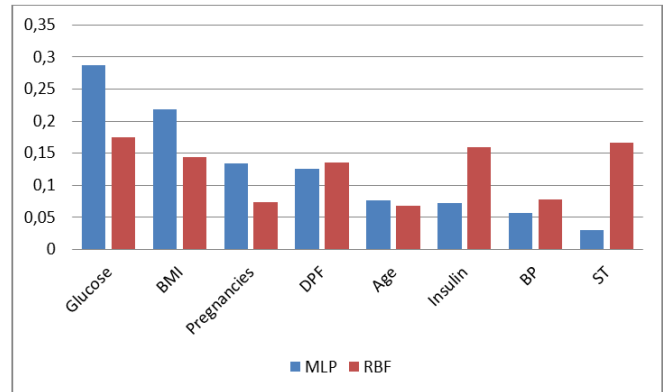


Fig.2. The importance values for possible risk factors (BMI: Body mass index; ST: Skin Thickness; DPF: Diabetes Pedigree Function; BP: Blood pressure; MLP: Multilayer perceptron; RBF: Radial-based function)

## 6. DISCUSSION

An artificial neural network is a very successful technique that solves the classification and prediction problems and is a mathematical model promoted by the regulation and functional feature of artificial neural networks. Neural networks include input and output layers and (in most cases) hidden layers that convert the input to output. When artificial neural network architecture is used to predict any disease, the ANN model can be generally built in two stages: training and testing. First, the ANN model is trained on the specified dataset and the weights of the connections between the neurons are fixed. Second, the model examined is validated to determine the classification of a new data set. The performance of the models constructed is evaluated using different criteria. [17].

In this study, it was aimed to apply multilayer perceptron and radial-based function from artificial neural network models on an open-source type 2 DM dataset and to compare classification estimates. In this framework, various factors (explanatory variables) that may be associated with type 2 DM (dependent variable) are estimated by multilayer perceptron and radial-based function artificial neural network models, and the use of artificial intelligence methods in the classification problem of interest is revealed. Also, the importance levels of factors that may be associated with type 2 DM for use in preventive medicine applications were obtained from these models.

According to the results of the performance criteria (accuracy, AUC, sensitivity, negative predictive value, and F-score) calculated in this study, the MLP model gave better predictive results than the RBF model in the classification of type 2 DM. However, when the criteria of selectivity and positive predictive value are taken into consideration compared to the MLP model, the higher classification rates were obtained. The three most important risk factors that can

be associated with type 2 DM were Glucose, BMI, and Pregnancies according to the MLP model, while the RBF model defined Glucose, Skin Thickness, and Insulin, respectively.

In a study using the same data set, the accuracy, sensitivity, and specificity performance criteria used in the classification made with support vector machines were obtained as 78%, 80%, and 76.5, respectively [18]. In another study using the same data set, classification was made using six different machine learning models. The best classification performances were achieved from the Hoeffding Tree algorithm based on these models, and precision, Recall, F-criterion, and area under the ROC curve values calculated from this algorithm were obtained as 0.757, 0.762, 0.759 and 0.816, respectively. [19]. In this study, the classification of DM was performed with MLP and RBF models on the same data set, and higher classification performances were obtained from the experimental results of the current study.

As a result, the findings obtained from this study showed that the classification of Type 2 DM performed successful predictions. Also, unlike similar studies examining the same dataset, the importance values of the factors associated with the type 2 DM were estimated with the classification models.

#### REFERENCES

- [1] J. A. Cramer, "A systematic review of adherence with medications for diabetes," *Diabetes care*, vol. 27, no. 5, pp. 1218-1224, 2004.
- [2] R. Shobhana, R. Begum, C. Snehalatha, V. Vijay, and A. Ramachandran, "Patients' adherence to diabetes treatment," *The Journal of the Association of Physicians of India*, vol. 47, no. 12, pp. 1173-1175, 1999.
- [3] N. Başkal, "Diabetes Mellitus Tanım, Klasifikasyon, Tanı, Klinik, Laboratuvar ve Patogenez," *Erdoğan G. Klinik Endokrinoloji. Anıttıp AŞ yayınları*, Ankara, pp. 207-233, 2003.
- [4] A. Cameron, "The metabolic syndrome: validity and utility of clinical definitions for cardiovascular disease and diabetes risk prediction," *Maturitas*, vol. 65, no. 2, pp. 117-121, 2010.
- [5] M. Arslan, "Diabetes mellitusta tanı ve sınıflandırma," *İliçin G, Biberoğlu K, Süleymanlar G, Ünal S (editörler). İç Hastalıkları*, vol. 2, pp. 2279-2295, 2003.
- [6] T. Yılmaz, "Diabetes mellitusun tanı kriterleri ve sınıflaması," *Diabetes Mellitus' un Modern Tedavisi*, birinci baskı, İstanbul, Türkiye Diyet Vakfı, 2003.
- [7] E. Öztemel, "Yapay Sinir Ağları, Papatya Yayıncılık, 2," Baskı, İstanbul, pp. 29-57, 2006.
- [8] S. Haykin, "Neural Networks: A comprehensive Foundation. by Prentice-Hall, Inc," Upper Saddle River, New Jersey, vol. 7458, pp. 161-175, 1999.
- [9] H. Batar, "EEG işaretlerinin dalgacık analiz yöntemleri kullanılarak yapay sinir ağları ile sınıflandırılması," *Fen Bilimleri Enstitüsü, Yüksek Lisans Tezi, Kahramanmaraş*, 89s, 2005.
- [10] A. Arı and M. E. Berberler, "Yapay Sinir Ağları ile Tahmin ve Sınıflandırma Problemlerinin Çözümü İçin Arayüz Tasarımı," *Acta Infologica*, vol. 1, no. 2, pp. 55-73, 2017.
- [11] E. Kilic, U. Ozbalci, and H. Ozcalik, "Lineer Olmayan Dinamik Sistemlerin Yapay Sinir Ağları ile Modellenmesinde MLP ve RBF Yapılarının Karşılaştırılması," *ELECO2012 Elektrik-Elektronik ve Bilgisayar Mühendisliği Sempozyumu*, (29.11. 2012-01.12. 2012), 2012.
- [12] S. S. Haykin, "Neural networks and learning machines/Simon Haykin," ed: New York: Prentice Hall, 2009.
- [13] M. Barale and D. J. I. J. o. C. A. Shirke, "Cascaded modeling for PIMA Indian diabetes data," vol. 139, no. 11, pp. 1-4, 2016.
- [14] J. W. Smith, J. Everhart, W. Dickson, W. Knowler, and R. Johannes, "Using the ADAP learning algorithm to forecast the onset of diabetes mellitus," in *Proceedings of the Annual Symposium on Computer Application in Medical Care*, 1988, p. 261: American Medical Informatics Association.
- [15] U. Orhan, M. Hekim, and M. Özer, "Discretization approach to EEG signal classification using Multilayer Perceptron Neural Network model," in *2010 15th National Biomedical Engineering Meeting*, 2010, pp. 1-4: IEEE.
- [16] O. Kaynar, Y. Görmez, Y. E. Işık, and F. Demirköparan, "Değişik Kümeleme Algoritmalarıyla Eğitilmiş Radyal Tabanlı Yapay Sinir Ağlarıyla Saldırı Tespiti," in *International Artificial Intelligence and Data Processing Symposium (IDAP'16)*, 2016.
- [17] I. M. Nasser and S. S. Abu-Naser, "Lung Cancer Detection Using Artificial Neural Network," *International Journal of Engineering and Information Systems (IJEAIS)*, vol. 3, no. 3, pp. 17-23, 2019.
- [18] V. A. Kumari and R. Chitra, "Classification of diabetes disease using support vector machine," *International Journal of Engineering Research and Applications*, vol. 3, no. 2, pp. 1797-1801, 2013.
- [19] F. Mercaldo, V. Nardone, and A. Santone, "Diabetes mellitus affected patients classification and diagnosis through machine learning techniques," *Procedia computer science*, vol. 112, pp. 2519-2528, 2017.

#### BIOGRAPHIES

**Emek GÜLDOĞAN** obtained his BSc. degree in Computer Engineering from Middle East Technical University in 2001. He received MSc. degree in biostatistics and medical informatics from the Inonu University in 2005, and Ph.D. degrees in biostatistics and medical informatics from the Inonu University in 2017. He is currently working as an assistant professor the Department of Biostatistics and Medical Informatics at Inonu University and as the information processing manager at Turgut Özal medical center. His research interests are cognitive systems, data mining, machine learning, deep learning.

**Zeynep Tunç** 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.

**Ayca Acet** received the BSc degree in Computer Engineering from TOBB University of Economics and Technology, Ankara, Turkey, in 2017. She currently continues M.S. degree in Computer Engineering, at Inonu University, Malatya, Turkey. Her general research interests include machine learning methods, deep learning and data mining.

**Cemil Çolak** obtained his BSc. degree in statistics from Ondokuz Mayıs University in 1999. He received MSc. degree in statistics from the Inonu University in 2001, and Ph.D. degrees in the Graduate Department of Biostatistics and Medical Informatics of Ankara University in 2005. He accepted as a Postdoctoral Researcher by the department of biostatistics and medical informatics of Inonu University in 2007. His research interests are cognitive systems, data mining, reliability, and biomedical system, and genetics, and bioengineering. In 2016, he joined the Department of Biostatistics and Medical Informatics at Inonu University as a professor, where he is presently a professor. He is active in teaching and research in the general image processing and data mining, analysis.



# CLASSIFICATION OF BREAST CANCER AND DETERMINATION OF RELATED FACTORS WITH DEEP LEARNING APPROACH

E. Guldogan, Z. Tunc and C. Colak

**Abstract— Aim:** In this study, it is aimed to classify breast cancer and identify related factors by applying deep learning method on open access to breast cancer dataset.

**Materials and Methods:** In this study, 11 variables related to open access to breast cancer dataset of 569 patients shared by the University of Wisconsin were used. The deep learning model for classifying breast cancer was established by a 10-fold cross-validation method. The performance of the model was evaluated with accuracy, sensitivity, specificity, positive/negative predictive values, F-score, and area under the curve (AUC). Factors associated with breast cancer were estimated from the deep learning model.


**Results:** Accuracy, specificity, AUC, sensitivity, positive predictive value, negative predictive value, and F-score values obtained from the model were 94.91%, 91.47%, 0.988, 96.90%, 95.42%, 95.14%, and 96.03%, respectively. In this study, when the effects of the variables in the dataset on breast cancer were evaluated, the three most important variables were obtained as area mean, concave points mean and symmetry mean, respectively.


**Conclusion:** The findings of this study showed that the deep learning model provided successful predictions for the classification of breast cancer. Also, unlike similar studies examining the same dataset, the importance values of cancer-related factors were estimated with the help of the model. In the following studies, breast cancer classification performances can give more successful predictions thanks to different deep learning architectures and ensemble learning approaches.


**Keywords—**Breast cancer, artificial intelligence, deep learning, classification.

## 1. INTRODUCTION

**B**REAST cancer is one of the leading causes of death among women in developed and developing countries. Detection and classification of breast cancer development in the early stages allow patients to receive appropriate treatment. Breast cancer is considered a genetically heterogeneous and biologically diverse disease. Long-known

**Emek GULDOGAN**, Inonu University Department of Biostatistics and Medical Informatics, Faculty of Medicine, Malatya, Turkey, (emek.guldogan@inonu.edu.tr) 

**Zeynep TUNC**, Inonu University Department of Biostatistics and Medical Informatics, Faculty of Medicine, Malatya, Turkey, (zeynep.tunc@inonu.edu.tr) 

**Cemil COLAK**, Inonu University Department of Biostatistics and Medical Informatics, Faculty of Medicine, Malatya, Turkey, (cemil.colak@inonu.edu.tr) 

Manuscript received May 08, 2020; accepted May 29, 2020.

Digital Object Identifier:

clinical and phenotypic differences are associated with differences in gene expression. Previous studies of breast tumors have identified five different types of breast carcinoma subtypes [luminal A (estrogen receptor (ER) +); luminal B (ER +); HER2 overexpression; normal breast-like and basal-like] associated with different clinical outcomes [1, 2].

Artificial intelligence (AI) involves the use of computer systems to achieve set goals by mimicking cognitive abilities. Machine learning (ML) classification is an AI field that allows algorithms or classifiers to learn patterns in large and complex datasets and produce useful predictive outputs. Applying ML algorithms to large datasets can reveal new trends and relationships that may have beneficial effects for clinical practice in medicine. Scientific studies have investigated the application of ML methods in health care and have shown that ML has an important effect on improving health quality and safety [3]. In an actual study, it is reported that artificial intelligence systems that can perform at the level of expert radiologists in digital mammography evaluation increase breast cancer screening accuracy and efficiency [4].

The complex structure of processes such as pretreatment, clustering, feature selection, and extraction, etc. in classical machine learning approaches reduces the performance and accuracy of the system. To solve problems related to traditional machine learning techniques, deep learning strategies are proposed to extract relevant information from raw images and to be used effectively in the classification process. In deep learning, features are determined by the training operations performed from data sets with the help of a general-purpose learning approach. [1].

In this study, it is aimed to classify breast cancer and determine related factors by applying deep learning method on open access to breast cancer data set.

## 2. MATERIAL AND METHODS

### 2.1. Dataset

To analyze the working principle of the deep learning method and to evaluate the model, the open-access dataset called "Breast Cancer Wisconsin (Diagnostic) Data Set" was obtained from UCI Machine Learning Repository [5]. In the data set used, there are 569 people examined for breast cancer. Of the individuals, 357 (62.7%) were diagnosed as benign and 212 (37.3%) were diagnosed as malignant. The explanations about the variables in the data set and their properties are given in Table 1.



TABLE I

EXPLANATIONS ABOUT THE VARIABLES IN THE DATASET AND THEIR PROPERTIES

Variable	Variable Description	Variable Type	Variable Role
diagnosis	The diagnosis of breast tissues (M = malignant, B = benign)	Qualitative	Output
radius_mean	Mean of distances from the center to points on the perimeter	Quantitative	Input
texture_mean	The standard deviation of gray-scale values	Quantitative	Input
perimeter_mean	Mean size of the core tumor	Quantitative	Input
area_mean	-	Quantitative	Input
smoothness_mean	Mean of local variation in radius lengths	Quantitative	Input
compactness_mean	mean of perimeter <sup>2</sup> / area - 1.0	Quantitative	Input
concavity_mean	Mean of the severity of concave portions of the contour	Quantitative	Input
concave points_mean	mean for the number of concave portions of the contour	Quantitative	Input
symmetry_mean	-	Quantitative	Input
fractal_dimension_mean	mean for "coastline approximation" - 1	Quantitative	Input

### 3. DEEP LEARNING MODEL

Deep Learning is based on the multi-layer feed-forward neural network trained with stochastic slope landing using the back-propagation approach. Related network; the hyperbolic tangent (tanh), rectifier, and maxout (a generalization of ReLU and leaky ReLU functions) can contain many hidden layers of neurons with activation functions. Advanced features such as adaptive learning speed, rate annealing, momentum training, dropout, and L1 or L2 regulations can provide high predictive accuracy. L1 is a regularization technique that restrains the absolute valuation of the weights and has the net influence of dropping some weights (setting them to zero) from a model to decrease complexity and refrain overfitting problems. L2 is another regularization technique that restrains the sum of the squared weights. This technique presents bias into the estimates of the parameter; however, it frequently performs considerable gains in modeling as the variance of the estimate is decreased. Each computes node trains a copy of global model parameters on local data in multiple threads (asynchronously) and periodically contributes to the global model through the model average across the network [6].

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 10 equal parts. One part is used as a test set and the remaining 9 parts are used as a training data set and this process is repeated 10 times. Hyperparameters related to the deep learning model were selected as activation function (Maxout linear unit), hidden layer sizes (50), the number of revolutions (10), epsilon ( $1.0 e^{-8}$ ) and rho (0.99). Table 2 shows the hyperparameters used in building the deep learning model [7]. RapidMiner Studio software was used in all modeling and analysis [8].

TABLE II

HYPERPARAMETERS USED TO CONSTRUCT A DEEP LEARNING MODEL

Hyperparameter name	Hyperparameter selection
Activation function	Maxout linear unit
Hidden layer sizes	50
Number of revolutions	10
Epsilon	$1.0 e^{-8}$
Rho	0.99

### 3.1. Performance evaluation criteria

The classification matrix for the calculation of performance metrics is given in Table 3.

TABLE III

CONFUSION MATRIX FOR CALCULATING PERFORMANCE METRICS

		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
		N		

The metrics considered in the performance evaluation of the models in this study are given below.

$$\text{Accuracy} = (TP+TN)/(TP+TN+FP+FN)$$

$$\text{Sensitivity} = TP/(TP+FP)$$

$$\text{Specificity} = TN/(TN+FN)$$

$$\text{Positive predictive value} = TP/(TP+FP)$$

$$\text{Negative predictive value} = TN/(TN+FN)$$

$$\text{F-score} = (2*TP)/(2*TP+FP+FN)$$

### 4. DATA ANALYSIS

Quantitative data are summarized by median (minimum-maximum) and qualitative variables are given by number and percentage. Normal distribution was evaluated with the Kolmogorov-Smirnov test. Whether there is a statistically significant difference between categories of the dependent

variable in terms of input variables, the Mann-Whitney U test was used for the analyses.  $p < 0.05$  values were considered statistically significant. In all analyzes, IBM SPSS Statistics 26.0 for the Windows package program was used.

### 5. RESULTS

Descriptive statistics related to the target variable examined in this study are presented in Table 4. There is a statistically significant difference between the dependent variable classes in terms of other variables other than the fractal\_dimension\_mean variable ( $p < 0.001$ ).

TABLE IV  
DESCRIPTIVE STATISTICS ABOUT INPUT AND OUTPUT VARIABLES

Variables	Diagnosis		p* value
	Benign (n=357)	Malignant (n=212)	
	Median (min-max)	Median (min-max)	
radius_mean	12.2 (6.98-17.85)	17.33 (10.95-28.11)	<0.001
texture_mean	17.39 (9.71-33.81)	21.46 (10.38-39.28)	<0.001
perimeter_mean	78.18 (43.79-114.6)	114.2 (71.9-188.5)	<0.001
area_mean	458.4 (143.5-992.1)	932 (361.6-2501)	<0.001
smoothness_mean	0.09 (0.05-0.16)	0.1 (0.07-0.14)	<0.001
compactness_mean	0.08 (0.02-0.22)	0.13 (0.05-0.35)	<0.001
concavity_mean	0.04 (0-0.41)	0.15 (0.02-0.43)	<0.001
concave points_mean	0.02 (0-0.09)	0.09 (0.02-0.2)	<0.001
symmetry_mean	0.17 (0.11-0.27)	0.19 (0.13-0.3)	<0.001
fractal_dimension_mean	0.06 (0.05-0.1)	0.06 (0.05-0.1)	0.537

\*: Mann Whitney U test

In this study, the classification matrix for the deep learning model used to classify breast cancer is given in Table 5 below.

TABLE V  
CLASSIFICATION MATRIX FOR DEEP LEARNING MODEL

Predicted \ Real	Malignant	Benign	Total
	Present	194	11
Absent	18	346	364
Total	212	357	569

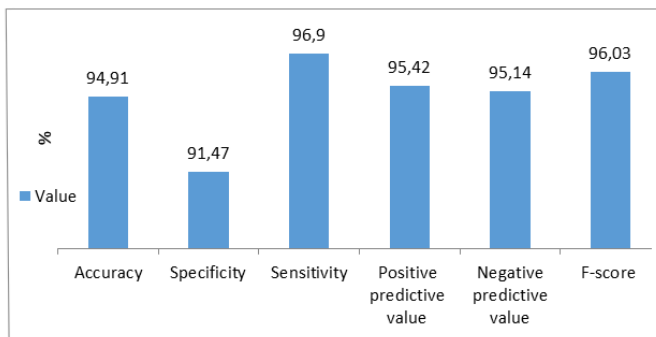


Fig. 1. Values related to performance criteria

The values related to performance criteria are given graphically in Figure 1.

Accuracy, specificity, AUC, sensitivity, positive/negative predictive value, F-score metrics for the deep learning model are summarized in Table 6. Accuracy, specificity, AUC, sensitivity, positive predictive value, negative predictive value, and F-score values obtained from the model were 94.91%, 91.47%, 0.988, 96.90%, 95.42%, 95.14%, and 96.03%, respectively.

TABLE VI  
THE VALUES OF PERFORMANCE METRICS

Performance Criterion	Value
Accuracy (%)	94.91
Specificity (%)	91.47
AUC	0.988
Sensitivity (%)	96.90
Positive predictive value (%)	95.42
Negative predictive value (%)	95.14
F-score (%)	96.03

In this study, the importance values of the factors related to breast cancer are given in Table 7. Among the three most important factors, the most important variable is the area\_mean, followed by the concave points\_mean and symmetry\_mean, respectively.

TABLE VII  
SEQUENCE OF VARIABLES IN ORDER OF IMPORTANCE

Variables	Importance (%)
area_mean	10.99
concave points_mean	10.78
symmetry_mean	10.77
perimeter_mean	10.67
fractal_dimension_mean	10.04
concavity_mean	10.03
radius_mean	9.32
smoothness_mean	9.22
compactness_mean	9.13
texture_mean	9.06

According to the World Health Organization (WHO), early diagnosis of cancer greatly increases the chances of making the right decision on a successful treatment plan [9, 10]. Computer-Aided Diagnosis (CAD) systems are widely applied in the detection and differential diagnosis of many different types of diseases. Therefore, increasing the accuracy of a CAD system has become one of the main research areas. In this study, using the deep learning approach on open access breast cancer data set, computer-aided classification of breast cancer and related factors were determined. Thus, it is possible to prevent the progression of the disease and to implement alternative treatment protocols by diagnosing breast cancer in the early stages [11, 12].

When similar studies were examined, Wisconsin Original Data Set consisting of 569 records and 31 (30 predictors, 1 target) feature/variable was used to increase the accuracy of the diagnosis of breast cancer in different machine learning methods. The accuracy of the proposed support vector machine model was found to be 0.9766 and the study results showed that the proposed model has a high-performance rate and will contribute to improving breast cancer diagnosis accuracy, which is an important problem of today. In this study, the accuracy value was calculated as 0.9491 in the breast cancer classification made using only 11 (10 predictors, 1 dependent) feature/variable on the same data set [13]. In this study, breast cancer classification was made successfully by using fewer variables/features, and similar performance criteria were obtained in the study described. Thus, in this study, the breast cancer classification accuracy rate was obtained very high by using fewer variables, and the importance values related to the investigated features were also revealed with the deep learning technique. Clinicians can evaluate the risk factors that may be effective in the development of breast cancer clinically more effectively with the help of the importance values related to the variables obtained from the deep learning model created. In a similar study, a CAD was developed for the detection of breast cancer

using a back-propagation supervised approach following deep belief networks unsupervised learning. In the model used in this process, weights were obtained from the deep belief network and backpropagation neural network was used with the learning function of Liebenberg Marquardt. The model created was tested on the Wisconsin Breast Cancer Data Set and gave a 99.68% accuracy rate showing promising results compared to previously published studies [14]. In the study summarized, only deep learning algorithms were used to detect breast cancer and no risk factor analysis that could be associated with breast cancer was performed. In this respect, this study shows significant differences from similar studies examining the same data set.

Breast cancer risk prediction provides systematic identification of individuals at the highest and lowest risk. Thus, the detection of high levels of breast cancer risk factors in the general society and women with a family history provides a more accurate decision on disease prevention therapies and screening [15-17]. When the effects of the variables in the data set examined in this study on breast cancer are examined; the three most important variables are as area\_mean (10.99%), concave points\_mean (10.78%), and symmetry\_mean (10.77%) were obtained as a result of calculations.

To sum up, the findings obtained from this study showed that the deep learning model created gave successful predictions in classifying breast cancer. Besides, unlike similar studies examining the same data set, the significance values of cancer-related factors were estimated from the model created. In further studies, the classification performances of different types of deep learning architectures and ensemble learning approaches can provide more successful predictions.

## REFERENCES

- [1] S. Khan, N. Islam, Z. Jan, I. U. Din, and J. J. C. Rodrigues, "A novel deep learning based framework for the detection and classification of breast cancer using transfer learning," *Pattern Recognition Letters*, vol. 125, pp. 1-6, 2019.
- [2] A. C. Peterson and H. Uppal, "Method for predicting response to breast cancer therapeutic agents and method of treatment of breast cancer," ed: Google Patents, 2019.
- [3] Q. D. Buchlak et al., "Machine learning applications to clinical decision support in neurosurgery: an artificial intelligence augmented systematic review," *Neurosurgical review*, pp. 1-19, 2019.
- [4] A. Rodriguez-Ruiz et al., "Stand-alone artificial intelligence for breast cancer detection in mammography: comparison with 101 radiologists," *JNCI: Journal of the National Cancer Institute*, vol. 111, no. 9, pp. 916-922, 2019.
- [5] D. Dua and C. J. U. h. a. i. u. e. m. Graff, "UCI machine learning repository, 2017," vol. 37, 2019.
- [6] M. Hofmann and R. Klinkenberg, *RapidMiner: Data mining use cases and business analytics applications*. CRC Press, 2016.
- [7] G. O. TEMEL, S. ERDOĞAN, and H. ANKARALI, "Sınıflama Modelinin Performansını Değerlendirmede Yeniden Örnekleme Yöntemlerinin Kullanımı," *Bilişim Teknolojileri Dergisi*, vol. 5, no. 3, pp. 1-8, 2012.
- [8] I. Mierswa and R. Klinkenberg, "RapidMiner Studio Version 9.5," ed, 2019.
- [9] F. Bray, J. Ferlay, I. Soerjomataram, R. L. Siegel, L. A. Torre, and A. Jemal, "Global cancer statistics 2018: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries," *CA: a cancer journal for clinicians*, vol. 68, no. 6, pp. 394-424, 2018.

- [10] WHO. (2018). Latest global cancer data: Cancer burden rises to 18.1 million new cases and 9.6 million cancer deaths in 2018. Available: <https://www.who.int/cancer/PRGlobocanFinal.pdf>
- [11] N. ALPASLAN, "MEME KANSERİ TANISI İÇİN DERİN ÖZNİTELİK TABANLI KARAR DESTEK SİSTEMİ," Selçuk Üniversitesi Mühendislik, Bilim Ve Teknoloji Dergisi, vol. 7, no. 1, pp. 213-227, 2019.
- [12] V. Bajaj, M. Pawar, V. K. Meena, M. Kumar, A. Sengur, and Y. Guo, "Computer-aided diagnosis of breast cancer using bi-dimensional empirical mode decomposition," Neural Computing and Applications, vol. 31, no. 8, pp. 3307-3315, 2019.
- [13] H. Kör, "Classification of Breast Cancer by Machine Learning Methods."
- [14] A. M. Abdel-Zaher and A. M. Eldeib, "Breast cancer classification using deep belief networks," Expert Systems with Applications, vol. 46, pp. 139-144, 2016.
- [15] A. Lee et al., "BOADICEA: a comprehensive breast cancer risk prediction model incorporating genetic and nongenetic risk factors," 2019.
- [16] A. Brédart et al., "Clinicians' use of breast cancer risk assessment tools according to their perceived importance of breast cancer risk factors: an international survey," Journal of community genetics, vol. 10, no. 1, pp. 61-71, 2019.
- [17] S. Karadag Arli, A. B. Bakan, and G. Aslan, "Distribution of cervical and breast cancer risk factors in women and their screening behaviours," European journal of cancer care, vol. 28, no. 2, p. e12960, 2019.

## BIOGRAPHIES

**Emek GÜLDOĞAN** obtained his BSc. degree in Computer Engineering from Middle East Technical University in 2001. He received MSc. degree in biostatistics and medical informatics from the Inonu University in 2005, and Ph.D. degrees in biostatistics and medical informatics from the Inonu University in 2017. He is currently working as an assistant professor of the Department of Biostatistics and Medical Informatics at Inonu University and as the information processing manager at Turgut Özal Medical Center. His research interests are cognitive systems, data mining, machine learning, deep learning.

**Zeynep Tunç** 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.

**Cemil Çolak** obtained his BSc. degree in statistics from Ondokuz Mayıs University in 1999. He received MSc. degree in Biostatistics from the Inonu University in 2001, and Ph.D. degree in the Graduate Department of Biostatistics and Medical Informatics of Ankara University in 2005. His research interests are cognitive systems, data mining, reliability, and biomedical system, genetics, and bioengineering. In 2016, he joined the Department of Biostatistics and Medical Informatics at Inonu University as a Professor, where he is presently a professor. He is active in teaching and research in the general image processing, artificial intelligence, data mining, analysis.

# SAMPLE SIZE EFFECT ON CLASSIFICATION PERFORMANCE OF MACHINE LEARNING MODELS: AN APPLICATION OF CORONARY ARTERY DISEASE

M. Kivrak, F.B. Akcesme, and C. Çolak

**Abstract**—Cardiovascular diseases are among the most common causes of death due to their widespread prevalence. Accurate and timely diagnosis of coronary artery disease, one of the fatal cardiovascular diseases, is very important. Angiography, an invasive method, is an expensive and special method used to determine the disease and can cause serious complications. Therefore, cheaper and more efficient data mining methods are used in the diagnosis and treatment of cardiovascular diseases. As an alternative approach, by establishing clinical decision support systems using data modeling and analysis methods such as data mining, errors and costs can be reduced by providing clinicians with computer-aided diagnosis, and patient safety and clinical decision quality can be significantly increased. In this study, the data set on the open-source access website was used to classify cardiovascular disease and consists of patient records of 14 variables created by the Cleveland clinic. Also, machine learning methods (C5.0 Decision Tree, Support Vector Machine, Multilayer Perceptron, and Ensemble Learning) were used to determine the risk of coronary artery disease by deriving 1000 and 10000 data sets from the cardiology data set obtained from original 303 patient records. Performance evaluation of models is compared in terms of accuracy, specificity, and sensitivity. In trying to determine the most successful model in estimating the risk of coronary artery disease, the results are presented comparatively.

**Keywords**—Cardiovascular Diseases, Sample Size, Data Mining, Ensemble Learning.

## 1. INTRODUCTION

THE Cardiovascular diseases (CVD) are caused by pathologies in the heart and blood vessels, and coronary artery disease (CAD), heart failure, cardiac arrest, ventricular arrhythmias, sudden heart death, ischemic stroke, transient ischemic attack, subarachnoid and intracerebral hemorrhage, abdominal aortic aneurysm, can result in diseases and congenital heart diseases [1]. CVD can cause myocardial infarction, heart failure, and

sudden heart death. Nuclear screening, echocardiography, electrocardiogram (ECG), non-invasive (non-invasive) procedures such as exercise stress test, and invasive (interventional) procedures such as angiography are required for the diagnosis of coronary artery disease [2]. For this reason, the angiography diagnostic method, which is an invasive method, is used as a determinant in the definitive diagnosis of coronary artery diseases and in determining the severity of the disease. However, angiography procedure is a diagnostic method that requires a high cost and advanced technical expertise [3]. As an alternative approach, by establishing clinical decision support systems using data modeling and analysis methods such as data mining, errors and costs can be reduced by providing clinicians with computer-aided diagnosis, and patient safety and clinical decision quality can be significantly increased [4].

This study aims to classify cardiovascular disease and consisted of patient records of 14 variables created by the open-source dataset of the Cleveland Clinic. Besides, machine learning methods (C5.0 Decision Tree, Support Vector Machine (SVM), Multilayer Perceptron (MLP), and Ensemble Learning) were used to determine the risk of coronary artery disease by deriving 1000 and 10000 data sets from the cardiology data set obtained from original 303 patient records. Performance evaluation of models is compared in terms of accuracy, specificity, and sensitivity. In order to determine the most successful model in estimating the risk of coronary artery disease, the results are presented comparatively on the open-sourced heart dataset.

## 2. MATERIAL AND METHOD


### 2.1. Data Set


The dataset used for the analysis was obtained from [http://archive.ics.uci.edu/ml/datasets/statlog+\(heart\)](http://archive.ics.uci.edu/ml/datasets/statlog+(heart)) [5]. The data set contains the original 303 heart disease data and 14 variables. In the original 303 heart disease dataset, 1000 and 10000 datasets were derived from the dataset that showed similar distributions from the dataset due to the binomial distribution of the target variable (glass) and the normal, binomial and uniform distribution of the explanatory variables. These variables are class, age, gender, chest pain type, resting blood pressure, serum cholesterol, fasting blood sugar, resting electrocardiographic results, maximum heart rate achieved, painloc, oldpeak, the slope of the peak exercise ST segment, number of major vessels (colored vessels) and thal. The detailed explanations of the variables are given in Table I.

### 2.2. Knowledge Discovery in Databases (KDD)

In the process of KDD; data selection (heart dataset), data preprocessing (extreme and missing value analyses), data

**Mehmet Kivrak**, Dept. of Biostatistics and Medical Informatics Faculty of Medicine, Inonu University, Malatya, Turkey, (e-mail: mehmetkivrak83@gmail.com) 

**F. Berat Akcesme**, Dept. of Biostatistics and Medical Informatics Faculty of Medicine, University of Health Sciences, Istanbul, Turkey, (e-mail: farukberat.akcesme@sbu.edu.tr) 

**Cemil Çolak**, Dept. of Biostatistics and Medical Informatics Faculty of Medicine, Inonu University, Malatya, Turkey, (e-mail: cemilcolak@yahoo.com) 

Manuscript received May 03, 2020; accepted May 30, 2020.  
Digital Object Identifier:



transformation(normalization, etc.), data mining and evaluation, and interpretation of the results were performed.

### 2.3. Classification Method

The most commonly used data mining methods on the analyzed datasets have been applied for the classification of CVD. Performance data obtained by using C5.0 Decision Tree, SVM, MLP, and Ensemble Learning classification methods were comparatively presented to the data sets (303, 1000, and 10000 sample sizes).

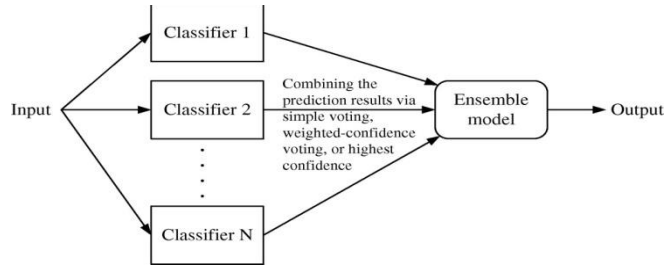


Fig.1. Classification Method and Ensemble Learning Algorithm.

#### 2.3.1. C5.0 Decision Tree

The C5.0 Decision Tree is one of the methods for supervised learning in the form of a tree structure used for classification as well as regression in general. The aim is to build the tree structure that predicts the label of a target variable using the model created.[6]. The C5.0 algorithm uses the concept of knowledge gain and entropy to optimally separate nodes. When there are k probabilities for X variable (attribute)  $P_1, P_2, P_3, \dots, P_k$  respectively, entropy for variable X is given in the equation below [7].

$$Entropy = H(X) = -\sum_{j=1}^k p_j \log_2(p_j) \quad (1)$$

When the target attribute of the sub-clusters  $T_1, T_2, T_3, \dots, T_k$  in the training set is subdivided into sub-compartments, the weighted average of the information required to determine the class of each T is given as the weighted sum of entropies.

$$H_S(T) = \sum_{i=1}^k p_i H_S(T_i) \quad (2)$$

Information gain is calculated to perform the separation process. The C5.0 algorithm realizes the optimal separation process by determining the separation criterion that has the greatest information gain in each decision node. Information gain is given in the equation below[8].

$$IG(S) = H(T) - H_S(T) \quad (3)$$

#### 2.3.2. Support Vector Machine (SVM)

SVM, which is accepted as the latest technology in pattern recognition, aims to increase the predictive performance by finding the Maximal Marginal Hyper Plane (MMH). Sequential Minimum Optimization (SMO) improves the training of the SVM classifier using polynomial nuclei. This generally replaces all missing values and converts the nominal properties to binary values[9,10].To find a decision boundary

between the two classes, SVM tries to maximize the gap between classes, choosing linear separations in a property area. Classification of the k-core function points in space  $x_i$  is  $y_i$ , which varies between -1 and +1. If  $x'$  is a point with an unknown classification, the prediction classification  $y'$  is as in the equation below.

$$y' = Sign(\sum_{i=1}^n \alpha_i y_i K(X_i, X') + d) \quad (4)$$

In the equation, K; core function, n; support vector number,  $\alpha$ ; adjustable weight and d are defined as bias. The classification process is linear in the number of support vectors [11].

#### 2.3.3. Multilayer Perceptron (MLP)

The most widely used artificial neural network model today is the MLP network, which has also been extensively analyzed and many learning algorithms have been developed from it.[12].MLP is a feed-forward, fully artificial neural network model that maps input data sets to an appropriate output set by adjusting the weight between internal data nodes.

$$y = \phi(\sum_{i=1}^n W_i X + b) = \phi(W^T X + b) \quad (5)$$

Equality; W defines the weight vector, X the vector of inputs, b bias (bias), and  $\phi$  activation function [13].

### 2.4. Ensemble Learning

Ensemble learning methods essentially aim to achieve the most accurate result by combining different methods. It can also be applied successfully in various machine learning systems such as feature extraction, error correction, unstable data, learning to deviate in non-stationary distributions, and confidence estimation."Bagging and Boosting" are the most commonly used algorithms for the training of ensemble classifiers. The most common unification rule used to combine individual classifiers is majority voting. The choice of the  $W_c$  class with the majority vote is as inequality [14,15].

$$\sum_{t=1}^T d_{t,c} = \max_c \sum_{t=1}^T d_{t,c} \quad (6)$$

### 2.5. Performance Metrics

Accuracy (AC) is defined as the division of values incompatible eyes by the total number of observations and is indicated by equation 7.

$$AC = \frac{TP+TN}{TP+TN+FN+FP} \quad (7)$$

Sensitivity is the ability of the test to distinguish patients from real patients and is indicated by equation 8.

$$Sensitivity = \frac{TP}{TP+FP} \quad (8)$$

Specificity is the ability of the test to distinguish robots from real robots and is indicated by equation 9 [16].

$$Specificity = \frac{TN}{TN+FN} \tag{9}$$

### 3. RESULTS

#### 3.1. Model Development

In data sets of 303, 1000, and 10000; Due to the low performance of the model, the feature selection model was applied to the data set. Variables 0.8 and above were determined as important contributing variables, while 0.6 and above variables were determined as marginal contributing variables. After the optimization process, data sets were divided into two as 70 % training and 30 % testing. Data analysis was performed by using the IBM SPSS Modeler Version 18.0 package program.

#### 3.2. Evaluation of the Models

After the model development, the evaluation metrics calculated within the scope of the investigation of how the sample size affects the model performance by using different classification methods are shown in Table II. For n = 303, the highest accuracy rate in the train data set was 77.2 %, while the group was ensemble learning, while the lowest classifier was MLP with 60.7 %.

TABLE I

THE DETAIL EXPLANATION OF THE VARIABLES IN THE DATASET

Variables	Explanation
Class	Target(0: healthy,1: disease)
Age	age
Gender	gender(1=male, 0=female)
Chest pain type	chest pain type (1=angina, 2=atypical angina, 3=non-anginal pain, 4=asymptomatic pain)
Resting blood pressure	resting blood pressure
Serum cholesterol	serum cholesterol in mg/dl
Blood sugar	fasting blood sugar, (fasting blood sugar > 120 mg/dl) (1 = true; 0 = false)
Electrocardiographic results	resting electrocardiographic results (0=normal,1= having ST-T wave abnormality, 2= showing probable or definite left ventricular hypertrophy by Estes' criteria )
Max heart rate	maximum heart rate achieved
Pain lock	exercise induced angina (1 = yes; 0 = no)
Oldpeak	Oldpeak= ST depression induced by exercise relative to rest
ST-segment	the slope of the peak exercise ST segment
Vessels	number of major vessels
Thal	Thal(A thalliumstress test; thal: 3 = normal; 6 =

In the test data set after model training, the highest classifier was again ensemble learning with 76.7 %, while the lowest was C5.0 with 63.3 %. For n = 1000, the highest accuracy rate in the train data set was 95.4 %, while the group was ensemble learning, while the lowest classifier was MLP with 66.7 %. In the test data set after model training, the highest classifier was again ensemble learning with 96.8 %, while the MLP was the lowest with 62.4 %. For n = 10000, the highest accuracy rate in the training data set was MLP with

94.2 %, while the lowest classifier was C5.0 with 86.7 %. After model training, the highest classifier was again MLP with 100 % in the test data set, while SVM was the lowest with 96.3 %.

TABLE II  
MODEL PERFORMANCE METRICS

Train (n=303)	Accuracy (%)	Sensitivity (%)	Specificity (%)	Time (Second)
SVM	69.7	63.7	62.6	5
C5.0	75.6	68.8	63.6	4
MLP	60.7	54.9	52.8	6
Ensemble	77.2	68	70.5	7
<b>Test (303)</b>				
SVM	73	70.6	75	2
C5.0	63.3	69.7	100	1
MLP	67.4	72.6	66.7	3
Ensemble	76.7	71.7	81.2	5
<b>Train (n=1000)</b>				
SVM	86	79.5	78.2	15
C5.0	94.1	88.6	83.8	12
MLP	66.7	63.5	58.9	13
Ensemble	95.4	87.8	87.2	11
<b>Test (n=1000)</b>				
SVM	90.9	81	83.3	8
C5.0	95.2	89.9	89.5	7
MLP	62.4	55.7	65	9
Ensemble	96.8	92.6	89.7	6
<b>Train (n=10000)</b>				
SVM	90.4	82.6	82.9	34
C5.0	86.7	84.5	81.7	28
MLP	94.2	88.6	86.2	44
Ensemble	90.5	86.4	82.1	23
<b>Test (n=10000)</b>				
SVM	96.3	90.3	90.2	17
C5.0	96.7	93.3	91	12
MLP	1	1	1	38
Ensemble	99.3	98.5	98.6	11

### 4. CONCLUSION

Diagnosis and treatment of a serious disease, such as cardiovascular diseases, is a very difficult problem and requires many pretreatment experiments and important datasets. The success of the models to be used when applying different classification methods can only be measured by

proving the performance. In this study, increasing the sample size in the data sets positively contributes to the model performance, it was determined that an ensemble learning algorithm is an approach that can be suggested in three data sets in general.

#### ACKNOWLEDGMENT

The study was reported as oral presentation in 1st International Data Science Congress in Health on 05-06 December 2019.

#### REFERENCES

- [1] Wong, N. D. (2014). *Epidemiological Studies of CHD and the Evolution of Preventive Cardiology*. Nature Reviews. Cardiology, 11(5), 276.
- [2] Verma, L., Srivastava, S., Negi, P. C. (2016). A Hybrid Data Mining Model to Predict Coronary Artery Disease Cases Using Non-Invasive Clinical Data. *Journal of Medical Systems*, 40(7), 1-7.
- [3] Alizadehsani, R., Hosseini, M. J., Sani, Z. A., Ghandeharioun, A., & Boghrati, R. (2012). Diagnosis of Coronary Artery Disease Using Cost-Sensitive Algorithms. In *Data Mining Workshops (ICDMW), 2012 IEEE 12th International Conference on* (pp. 9-16). IEEE.
- [4] Srinivas, K., Rani, B. K., & Govrdhan, A. (2010). Applications of Data Mining Techniques in Healthcare and Prediction of Heart Attacks. *International Journal on Computer Science and Engineering (IJCSE)*, 2(02), 250-255.
- [5] E. Smirnov, I. Sprinkhuizen-Kuyper, and G. Nalbantov, "Unanimous voting using support vector machines," in *BNAIC-2004: Proceedings of the Sixteenth Belgium-Netherlands Conference on Artificial Intelligence*, 2004, pp. 43-50.
- [6] Nicholas, E. (2008). *Introduction to Clementine and Data Mining*. Brigham Young University
- [7] Larose, D.T., and Larose, C.D. (2014) *Discovering Knowledge In Data An Introduction To Data Mining*, New Jersey: John Wiley & Sons.
- [8] Quinlan, J.R. (1993) *C4.5: Programs for Machine Learning*, Morgan Kaufmann Publishers Inc. San Francisco, CA, USA.
- [9] Platt, J. C. (1999). Fast training of support vector machines using sequential minimal optimization, *advances in kernel methods*. Support Vector Learning, 185-208.
- [10] Azuaje, F. (2006). Wittenih, frank e: Veri madenciligi: Pratik makine öğrenim araçları ve teknikleri. *Biyomedikal mühendislik çevrimiçi*, 5 (1), 1-2.
- [11] Valdimir, V. N., & Vapnik, N. (1995). The nature of statistical learning theory.
- [12] Rosenblatt, F. (1958). *Two theorems of statistical separability in the perceptron*. United States Department of Commerce
- [13] Miller, D. J., & Pal, S. (2007). Transductive methods for the distributed ensemble classification problem. *Neural computation*, 19(3), 856-884.
- [14] Zhang, C., & Ma, Y. (Eds.). (2012) *Ensemble machine learning: methods and applications*. Springer Science & Business Media.
- [15] Polikar, R. (2012). Ensemble learning. In *ensemble machine learning*, Springer, Boston, MA, 1-34.
- [16] Alpar, R. (2016). Uygulamalı istatistik ve geçerlik-güvenirlilik: spor, sağlık ve eğitim bilimlerinden örneklerle. Detay Yayıncılık.

#### BIOGRAPHIES

**Mehmet Kıvrak** obtained his BSc degree in statistics from Dokuz Eylül University (DEU) in 2001. He received the BSc. and MSc. diploma in Statistics from Dokuz Eylül University in 2001 and 2006 respectively, and Ph.D. degrees in the Graduate Department of Biostatistics and Medical Informatics of Inonu University in 2017. He accepted as an expert statistician Turkish Statistical Institute in 2009. His research interests are data mining, cognitive systems, reliability and genetics and bioengineering, and signal processing. His current research interests are genetics and bioengineering and data mining.

**F. Berat Akçeşme** obtained his BSc degree in biological sciences and bioengineering from the International University of Sarajevo in 2004. He received the BSc. and MSc. diploma in biological sciences and bioengineering from the International University of Sarajevo in 2004 and 2009 respectively, and Ph.D. degrees in Genetics and Bioengineering of the same university in 2012. He accepted as a Postdoctoral Researcher by the department of biostatistics and medical informatics faculty of medicine, university of health sciences in 2012. His research interests are cognitive systems, reliability and biomedical system, and genetics, and bioengineering. In 2017, he joined the Department of Biostatistics and Medical Informatics Faculty of Medicine, University of Health Sciences as an assistant professor, where he is presently an assistant professor. He is active in teaching and research in the general genetics and bioengineering modeling, analysis.

**Cemil Çolak** obtained his BSc. degree in statistics from Ondokuz Mayıs University in 1999. He received MSc. diploma in statistics from the Inonu University in 2001, and Ph.D. degrees in the Graduate Department of Biostatistics and Medical Informatics of Ankara University in 2005. He accepted as a Postdoctoral Researcher by the department of biostatistics and medical informatics of Inonu University in 2007. His research interests are cognitive systems, data mining, reliability, and biomedical system, and genetics, and bioengineering. In 2016, he joined the Department of Biostatistics and Medical Informatics at Inonu University as a professor, where he is presently a professor. He is active in teaching and research in the general image processing and data mining modeling, analysis.

# EVALUATION OF ASSOCIATION RULES BASED ON CERTAINTY FACTOR: AN APPLICATION ON DIABETES DATA SET


M. Kivrak, F.B. Akcesme, and C. Colak


**Abstract**— Data mining is the process of discovering useful information that has not been previously revealed from large amounts of data. Association rules mining is one of the most important techniques used in data mining and artificial intelligence. The first research in the association rules was to find relationships between different products in the customer transaction database and customer purchase models. Based on these relationships, researchers have begun to expand the field of data mining. One of these areas is the application of the rules of association in the field of medicine. Thus, through these applications, the relationship of various features in medical data can be discovered, and the findings obtained can aid medical diagnosis. Support and confidence are the two primary measures employed in the evaluation of association rules. The rules obtained with these two values are often correct; however, they are not strong rules. For this reason, there are many interestingness measures proposed to achieve stronger rules. Most of the rules, especially with a high support value, are misleading. For this reason, there are many interestingness measures proposed to achieve stronger rules. This study aims to establish strong association rules with variables in the open-sourced diabetes data set. In the current study, the Apriori algorithm was used to obtain the rules. As a result of the analysis, only 52 confidence and support criteria were taken into consideration. For more powerful rules, certainty factor was used as one of the interestingness measures proposed in the literature, and it was concluded that only 39 of these rules were strong as a result of the analysis.


**Keywords**— *Machine learning, classification, artificial neural network, support vector machines, decision tree, logistic regression, linear discriminant, nearest neighbor.*

## 1. INTRODUCTION

THE data mining is an important process of extracting previously unknown and useful information from data in databases [1]. Data mining techniques include classification, prediction, associations, and clustering. One of the most important data mining applications is that of mining association rules [2]. Initial research in association rules is about finding relationships between different products in the customer transaction database, as well as customer purchase models.

**Mehmet Kivrak**, Dept. of Biostatistics and Medical Informatics Faculty of Medicine, Inonu University, Malatya, Turkey, (e-mail: mehmetkivrak83@gmail.com) 

**F. Berat Akcesme**, Dept. of Biostatistics and Medical Informatics Faculty of Medicine, University of Health Sciences, Istanbul, Turkey, (e-mail: farukberat.akcesme@sbu.edu.tr) 

**Cemil Colak**, Dept. of Biostatistics and Medical Informatics Faculty of Medicine, Inonu University, Malatya, Turkey, (e-mail: cemil.colak@inonu.edu.tr) 

Based on these relationships, researchers had begun to expand the field of data mining.

One of these areas is the application of association rules in the medical field. Thus, through these applications, the association of various attributes can be discovered in medical data, and this can help medical diagnosis [3]. This study aims to establish strong association rules with variables in the open-sourced diabetes data set.

## 2. MATERIAL AND METHODS

### 2.1. Data Set

The dataset used for the analysis was obtained from the website (<http://datahub.i.o>machine-learning>diabetes>) [4]. The data set contains 768 samples and nine variables. These variables are age, pregnancies, PG concentration, diastolic BP, tri-fold thick, resting electrocardiographic results, serum ins, BMI, DP function, and diabetes. The detailed explanation of the variables are given in Table I. Data analysis was performed by using RStudio Version 1.1.463 programming language.

### 2.2. Knowledge Discovery in Databases (KDD)

In the process of KDD, data selection (diabetes dataset), data preprocessing (extreme and missing value analyses), data transformation (normalization, etc.), data mining and evaluation, and interpretation of the results were performed.

### 2.3. Association Rules Mining

Association rules mining is a very common technique that can define various rules or relationships between variables. [5]. These association rules are composed of two item sets, the antecedent (left-hand side) and consequent (right-hand side), the expression of the form is  $X \Rightarrow Y$ , where X and Y are called antecedent and consequent of the rule respectively [6,7]. There are many algorithms used in association rules such as the Apriori algorithm, Eclat algorithm, and FP-growth algorithm. The most classic and frequently used algorithm is the Apriori algorithm, and it used to find all frequent item sets in a given database [8].

#### 2.3.1. Basic Measures

The support and confidence are basic measures in building strong association rules from the frequent item sets [9].

Support is defined as the probability that transactions in the database contains items both the antecedent and the consequent of the rule, as follows.



TABLE I.  
THE DETAILED EXPLANATION OF THE VARIABLES  
DATA SET

Abbreviation	Explanation
Age	Age
Pregnancies	Number of pregnancy
PG Concentration	Plasma glucose in 2 hours in oral glucose tolerance test
Diastolic BP	Diastolic blood pressure (mm Hg)
Tri-Fold Thick	Three-layer binding layer thickness (mm)
Serum Ins	2 Hour serum insulin (mu U / ml)
BMI	(weight in kg / kg (height in m) ^ 2)
DP Function	Diabetes family tree function
Diabetes	Diabetes (1 = yes; 0 = no)

Support ( $X \Rightarrow Y$ ) = (Transactions containing both X and Y items) / Total number of transactions.

Confidence is a measure that reveals the association between antecedent and consequent of a rule.

Confidence ( $X \Rightarrow Y$ ): Total number of transactions containing items X and Y divided by the total number of transactions containing item X.

Confidence ( $X \Rightarrow Y$ ) = Support (X, Y) / Support(Y) [5].

If a minimum threshold for support is chosen low, large numbers of rules are created, and assessment of such rules is rather complex and time-consuming. Also, selecting the minimum threshold value high causes some rules to be skipped [5]. Therefore, some interestingness measures (IM) are developed to solve this problem [9].

### 2.3.2. Interestingness Measure (IM)

Certainty Factor (CF) is interpreted as a measure of the variation of the probability that Y is in a transaction when we consider only those transactions where X is defined as follows [7].

$$CF(X \Rightarrow Y) = (\text{Conf}(X \Rightarrow Y) - \text{Supp}(Y)) / (1 - \text{Supp}(Y))$$

If  $\text{Conf}(X \Rightarrow Y) > \text{Supp}(Y)$ ,

and,

$$CF(X \Rightarrow Y) = (\text{Conf}(X \Rightarrow Y) - \text{Supp}(Y)) / \text{Supp}(Y)$$

If  $\text{Conf}(X \Rightarrow Y) < \text{Supp}(Y)$ , and 0 otherwise.

## 1. EXPERIMENTAL RESULTS

Since the variables of age, pregnancies, PG concentration, diastolic BP, tri-fold thick, serum ins, BMI, and DP function were continuous variables on the data set, they were transformed into categorical variables by using the program.

For the experimental results, the minimum support value was 1.5 percent (1.5%), and the confidence value was 80 percent (80.0%).

As a result of the analysis, it was found that 52 rules were formed respectively, which consisted of triple, quadruple, and

quintile association rules, which were observed the most as of confidence and support values. It is necessary to take care of the high confidence and support values, which are among the measures used in the interpretation of the obtained rules [10].

Interpretation of the rules obtained only with the measures of confidence and support led to the obtaining of correct but not strong rules. There are many interestingness measures used in the literature to obtain stronger rules. In this context, in addition to the confidence and support measures, the certainty factor was used. Certainty factor approaching 1 indicates that the rules are identified with high accuracy [7]. For this reason, it was concluded that only 39 of the 52 rules that could provide a relationship between the diagnosis and other variables of the patient were strong. The rules obtained are given in Table II. Also, explanations of the variables used in Table II are given in Table I.

The three examples of the interpretation of the rules in Table II are given below:

- Rule 1: Patients with a Pregnancies [10], not having a tri-fold thick, do not have serum insulin with a probability of 100 %.
- Rule 2: Patients with a tri-fold thick [0] and between the age of [27] do not have serum insulin with a probability of 100 %.
- Rule 7: Patients with a Pregnancies[1] and between the age of [27] do not have diabetes with a probability of 0.96%.

Other rules achieved from the association rules mining are similarly interpreted in the related table.

## 2. DISCUSSION

Association rule mining is to discover association rules that satisfy a given database with predefined minimum support and confidence. The problem is usually broken down into two sub problems. One is to find those item sets whose occurrences in the database surpass a predefined threshold; such item sets are called regular or large item sets. The second problem is to create association rules with the minimum confidence constraints from those broad item sets [11].

Interpreting association rules by using only confidence and support measures might create many disadvantages; therefore, it will be more accurate to assess it with certainty factor, proposed by Shortliffe and Buchanan (1984) [12]. Similar to this study, Berzal et al. (2002) found that there were some disadvantages of evaluating association rules just with the measures of confidence and support; and also suggested that especially items with a very high support value might cause misleading rules [13].

There are many alternative interest measures to reach stronger rules in association rules mining. In this study, the Apriori algorithm was used to obtain association rules with open-sourced diabetes data set. In the application, the confidence factor and the rules with high support values are obtained on the right side of the rule. For that reason, in addition to the confidence and support measures, the certainty factor, which is one of the interestingness measures, was used to define the interesting rules in this research.



TABLE II.  
THE GENERATED ASSOCIATION RULE

Row num.	Rule num.	Association Rule (X ⇒Y)	Confidence	Support	Certainty Factor
1	1	{Pregnancies=10,Tri.Fold.Thick=0} => {Serum.Ins=0}	1	0.0156	1
2	2	{Tri.Fold.Thick=0,Age=27} => {Serum.Ins=0}	1	0.0156	1
3	3	{Diastolic.BP=0,Tri.Fold.Thick=0} => {Serum.Ins=0}	0.937	0.0195	0.911
4	4	{Diastolic.BP=0,Diabetes=yes} => {Serum.Ins=0}	1	0.0429	1
5	5	{Diastolic.BP=0,Diabetes=no} => {Serum.Ins=0}	0.942	0.0429	0.918
6	6	{Pregnancies=8,Tri.Fold.Thick=0} => {Serum.Ins=0}	0.947	0.0234	0.925
7	7	{Pregnancies=1,Age=23} => {Diabetes=no}	0.965	0.0208	1
8	8	{Diastolic.BP=76,Tri.Fold.Thick=0} => {Serum.Ins=0}	1	0.0247	1
9	10	{Diastolic.BP=78,Tri.Fold.Thick=0} => {Serum.Ins=0}	1	0.0234	1
10	12	{Pregnancies=6,Tri.Fold.Thick=0} => {Serum.Ins=0}	1	0.0169	1
11	13	{Diastolic.BP=74,Tri.Fold.Thick=0} => {Serum.Ins=0}	1	0.0182	1
12	14	{Diastolic.BP=70,Tri.Fold.Thick=0} => {Serum.Ins=0}	1	0.0195	1
13	15	{Pregnancies=5,Tri.Fold.Thick=0} => {Serum.Ins=0}	1	0.0182	1
14	17	{Pregnancies=4,Tri.Fold.Thick=0} => {Serum.Ins=0}	1	0.0208	1
15	21	{Pregnancies=2,Tri.Fold.Thick=0} => {Serum.Ins=0}	1	0.0273	1
16	22	{Pregnancies=0,Tri.Fold.Thick=0} => {Serum.Ins=0}	1	0.0182	1
17	23	{Pregnancies=1,Tri.Fold.Thick=0} => {Serum.Ins=0}	1	0.0169	1
18	24	{Tri.Fold.Thick=0,Diabetes=yes} => {Serum.Ins=0}	1	0.0325	1
19	28	{Pregnancies=6,Tri.Fold.Thick=0,Diabetes=no} => {Serum.Ins=0}	1	0.0156	1
20	30	{Pregnancies=4,Tri.Fold.Thick=0,Diabetes=yes} => {Serum.Ins=0}	1	0.0351	1
21	31	{Pregnancies=4,Tri.Fold.Thick=0,Diabetes=no} => {Serum.Ins=0}	1	0.0234	1
22	33	{Pregnancies=2,Tri.Fold.Thick=0,Diabetes=no} => {Serum.Ins=0}	1	0.0156	1
23	34	{Pregnancies=0,Tri.Fold.Thick=0,Diabetes=no} => {Serum.Ins=0}	0.967	0.0234	1
24	35	{Pregnancies=1,Age=21} => {Diabetes=no}	1	0.0299	1
25	37	{Diastolic.BP=0,Serum.Ins=0,Diabetes=no} => {Tri.Fold.Thick=0}	1	0.0416	1
26	38	{Diastolic.BP=0,Serum.Ins=0} => {Tri.Fold.Thick=0}	1	0.0208	1
27	39	{Diastolic.BP=0,Diabetes=yes} => {Tri.Fold.Thick=0}	1	0.1145	1
28	40	{Diastolic.BP=0,Serum.Ins=0,Diabetes=yes} => {Tri.Fold.Thick=0}	1	0.1809	1
29	41	{Pregnancies=1,Age=24} => {Diabetes=no}	1	0.0195	1
30	42	{Pregnancies=2,Age=21} => {Diabetes=no}	0.937	0.0195	0.911
31	43	{Pregnancies=1,Age=22} => {Diabetes=no}	1	0.0234	1
32	44	{Serum.Ins=0,Age=21} => {Diabetes=no}	0.947	0.0234	0.925
33	45	{Serum.Ins=0,Age=24} => {Diabetes=no}	1	0.0169	1
34	46	{Diastolic.BP=68,Serum.Ins=0} => {Diabetes=no}	1	0.0208	1
35	47	{Pregnancies=2,Age=25} => {Diabetes=no}	1	0.0182	1
36	49	{Diastolic.BP=60,Serum.Ins=0} => {Diabetes=no}	1	0.0169	1
37	50	{Pregnancies=4,Serum.Ins=0,Diabetes=yes} => {Tri.Fold.Thick=0}	1	0.0169	1
38	51	{Pregnancies=2,Serum.Ins=0} => {Diabetes=no}	1	0.0234	1
39	52	{Pregnancies=8,Diabetes=no} => {Serum.Ins=0}	1	0.0273	1

### 3. CONCLUSION

As a result of this research, in addition to the 52 rules obtained with confidence and support measures, in order to eliminate the misleading rules and to obtain stronger rules, 26 rules were obtained by using certainty factors. In addition to the basic measures, it is recommended that different interestingness measures should also be used to reach more accurate results.

### ACKNOWLEDGMENT

The study was reported as oral presentation in 1st International Data Science Congress in Health on 05-06 December 2019.

### REFERENCES

- [1] M.-S. Chen, J. Han, and P. S. Yu, "Data mining: an overview from a database perspective," *IEEE Transactions on Knowledge and data Engineering*, vol. 8, no. 6, pp. 866-883, 1996.
- [2] M. Ilayaraja and T. Meyyappan, "Mining medical data to identify frequent diseases using Apriori algorithm," in *2013 International Conference on Pattern Recognition, Informatics and Mobile Engineering*, 2013, pp. 194-199: IEEE.
- [3] W.-J. Zhang, D.-L. Ma, and B. Dong, "The automatic diagnosis system of breast cancer based on the improved Apriori algorithm," in *2012 International Conference on Machine Learning and Cybernetics*, 2012, vol. 1, pp. 63-66: IEEE.
- [4] D. Dua and C. J. C. a. C. U. Graff, "UCI machine learning repository [<http://archive.ics.uci.edu/ml>]. <https://archive.ics.uci.edu/ml/datasets>," 2019.
- [5] S. Kumar and N. Joshi, "Rule power factor: a new interest measure in associative classification," *Procedia Computer Science*, vol. 93, pp. 12-18, 2016.
- [6] S. Rao and P. Gupta, "Implementing Improved Algorithm Over APRIORI Data Mining Association Rule Algorithm 1," 2012.
- [7] F. Berzal, I. Blanco, D. Sánchez, and M.-A. Vila, "Measuring the accuracy and interest of association rules: A new framework," *Intelligent Data Analysis*, vol. 6, no. 3, pp. 221-235, 2002.
- [8] D. Jain and S. Gautam, "Implementation of apriori algorithm in health care sector: a survey," *International Journal of Computer Science and Communication Engineering*, vol. 2, no. 4, pp. 22-8, 2013.
- [9] J. Manimaran and T. Velmurugan, "Analysing the quality of association rules by computing an interestingness measures," *Indian Journal of Science and Technology*, vol. 8, no. 15, pp. 1-12, 2015.
- [10] O. Başak, B. Uğur, and M. K. SAMUR, "Kulak Burun Boğaz Epikriz Notlarından Birliktelik Kurallarının Çıkarılması," 2009.
- [11] S. Kotsiantis, D. J. G. I. T. o. C. S. Kanellopoulos, and Engineering, "Association rules mining: A recent overview," vol. 32, no. 1, pp. 71-82, 2006.
- [12] W. van Melle, E. H. Shortliffe, and B. G. J. R.-b. e. s. T. M. e. o. t. S. H. P. P. Buchanan, "EMYCIN: A knowledge engineer's tool for constructing rule-based expert systems," pp. 302-313, 1984.

- [13] F. Berzal, I. Blanco, D. Sánchez, and M.-A. J. I. D. A. Vila, "Measuring the accuracy and interest of association rules: A new framework," vol. 6, no. 3, pp. 221-235, 2002.

### BIOGRAPHIES

**Mehmet Kıvrak** obtained his BSc degree in statistics from Dokuz Eylül University (DEU) in 2001. He received the BSc. and MSc. Diploma in Statistics from Dokuz Eylül University in 2001 and 2006 respectively, and Ph.D. degrees in the Graduate Department of Biostatistics and Medical Informatics of Inonu University in 2017. He accepted as an expert statistician in the Turkish Statistical Institute in 2009. His research interests are data mining, cognitive systems, reliability and genetics and bioengineering, and signal processing. His current research interests are genetics, bio engineering and data mining.

**F. Berat Akçeşme** obtained his BSc degree in Biological Sciences and Bioengineering from the International University of Sarajevo (IUS) in 2009. He pursued his master study at Mediterranean Agronomic Institute of China in the field of Horticultural Genetics and Biotechnology. In 2012, he got accepted to the Ph.D program in Genetics and Bioengineering at IUS where he was working as a research assistant at the same department. He obtained his PhD degree in 2016. He continued to work at IUS as an assistant professor until the end of 2017. His research interests are cognitive systems, bioinformatics, structural bioinformatics. In 2017, he joined the Department of Biostatistics and Medical Informatics at the Faculty of Medicine, University of Health Sciences as an assistant professor. He is active in teaching and research in the genetics and bioengineering. Beside, he is director of Bioinformatics and Biostatistics Application and Research Center at University of Health Sciences.

**Cemil Çolak** obtained his BSc. Degree in statistics from Ondokuz Mayıs University in 1999. He received MSc. Diploma in statistics from the Inonu University in 2001, and Ph.D. degrees in the Graduate Department of Biostatistics and Medical Informatics of Ankara University in 2005. He accepted as a Postdoctoral Researcher by the department of biostatistics and medical informatics of Inonu University in 2007. His research interests are cognitive systems, data mining, reliability, and biomedical system, and genetics, and bioengineering. In 2016, he joined the Department of Biostatistics and Medical Informatics at Inonu University as a professor, where he is presently a professor. He is active in teaching and research in the general image processing and data mining modeling, analysis.

# PREDICTION OF POST-TREATMENT SURVIVAL EXPECTANCY IN HEAD & NECK CANCERS BY MACHINE LEARNING METHODS

H.S. Nogay

**Abstract**— In this study, survival for head and neck cancer disease was estimated using machine learning methods. Starting from the date on which the head and neck cancer disease was diagnosed, without a maximum time limit, at the end of the minimum 8-month period, it is estimated that the patient will be alive or not. Seven classifying machine-learning predictive methods were used in the study. The main goal of this study is to estimate the survivability of head and neck cancer patients and to provide a decision aid for cancer management with applied estimation methods and results. The results obtained by the application of the designed methods are examined and results with extremely high accuracy rates are obtained.

**Keywords**— *Machine learning, classification, artificial neural network, support vector machines, decision tree, logistic regression, linear discriminant, nearest neighbor.*

## 1. INTRODUCTION


HEAD and neck squamous cell carcinoma (HNSCC), including upper aerodigestive tract and anatomic regions, is considered the third leading cause of death worldwide. Progression of the HNSCC is a consequence of both the interaction of environmental factors and the genetic inheritance, and is therefore multi-factorial. Smoking and alcohol dependence are the main risk factors for the development of this disease. Human papillomavirus (HPV) is also thought to be a risk factor for the disease at approximately 25%. The annual incidence of head and neck cancers worldwide, the annual incidence of head and neck cancers worldwide; about 300,000 deaths result in about 550,000 cases per year. The male to female ratio ranges from 2: 1 to 4: 1. HNSCC is the sixth most common cancer worldwide incidence. The overall five-year survival rate of HNSCC patients is approximately 40-50%. Approximately one third of patients are suffering from early stage disease (T1-2, N0). Early HNSCC therapy usually involves single modality therapy with surgery or radiation [1-6]. Treatment for head and neck cancer may include surgery, radiotherapy, chemotherapy, targeted therapy, or a combination of these treatments. The treatment plan depends on various factors such as the precise location of the tumor, cancer stage, age of the patient and general health status [7].

Generally, head and neck cancers at the advanced stage result in the death of the patient despite all kinds of treatment. For this reason, the integration of chemotherapy and radiotherapy is crucial to prolong survival and improve the quality of life of patients. It is necessary to know and follow the general conditions of the patients before starting any treatment [8]. Estimation of survival rate and decision-making of treatment are of great importance both for cancer patients and for

physicians. The World Health Organization has stated that cancer is the second leading cause of death in the world. With early treatment, early detection of cancer will increase prognosis for cancer. At the same time, the prognosis depends on cancer spreading to lymph node drainage sites and metastasizing in different regions. A cancer staging system called TNM (Tumor, Node, Metastasis) is commonly used to determine the cancer status. Spreading to regional nodes or other nodes and distant metastasis reduce survival. Data-driven predictive models for cancer survival can help in prognosis and cancer management [9].

When deciding on the resection surgeon, a number of factors are considered that will affect the quality of life of the patient, including high rates of morbidity and the likelihood that death will occur rapidly. The morbidity rate is reported to be at least 50% and at least 15% mortality during the operation. The effective use of clinical data through the use of machine learning techniques such as artificial neural networks (ANN) can lead to more accurate diagnosis and prediction results, enabling better understanding of complex procedures and improving patient outcomes. Treatment decisions on the oncology not only directly affect survival, but also affect the quality of life of patients [10]. Survival analysis with machine learning methods provides greater convenience in logic implementation than statistical methods [11]. In machine learning, mathematical algorithms used as computer programs are used to recognize patterns in large data sets and to iteratively refine this recognition with additional data. When a specific medical diagnosis is made, predicting survival is crucial in improving patient care and providing information to patients and clinicians. In a data set of specific demographics (eg, age), diagnostic (eg, tumor size), and procedural (eg, radiation and/or surgery) information, it is very important to know that any of this information is sufficient to predict survival for head and neck cancer. Survival analysis is considered clinically important to evaluate the prognosis of the patient. More accurate results can be obtained by applying a correlational approach through machine learning to predict survival [12].

In recent years, significant progress has been made in the development of machine learning. The machine learning method implements a variety of techniques and approaches to analyze and summarize the data obtained from the databases, thus producing relevant information. Artificial neural networks (ANN) have proven to be very effective in disease prediction and survival analysis. Moreover, the unknown relationship between ANN input and output variables can be effectively predicted by repeating the learning and verification process of an ANN in a computer environment until the desired approach is provided. The quality of life of the patient before the treatment and the possible effect of the treatment on the survival of the patient and the associated quality of life affect the perception of treatment value of doctors and patient. It is

**Hidir Selcuk Nogay** is with Electrical and Energy Department, Kayseri University, Kayseri, Turkey, (e-mail: [nogay@kayseri.edu.tr](mailto:nogay@kayseri.edu.tr)). 

Manuscript received May 15, 2020; accepted May 31, 2020.  
Digital Object Identifier:

important that doctors try to protect or improve the quality of life of their patients. Patients generally believe that there is no option other than surgery. Another factor to consider is the potential regret of the doctor or the patient giving a wrong treatment decision [13,14].

To test for survival prediction studies, head and neck cancer disease is the most appropriate disease with some characteristics. Head and neck cancers can be correctly staged using clinical and radiological techniques, distant metastases appear later, and have a short 4 year period of hazard, which facilitates reliable monitoring. In addition, local-regional recurrence in head and neck cancer is easier to detect than other types of cancer and it typically occurs within two years. In addition to predicting survival with artificial neural networks, a hypothesis that ANN produces more successful estimates than other machine learning and classification approaches is tested [15].

In this study, survival for head and neck cancer disease was estimated using machine learning methods. Starting from the date the head and neck cancer disease was diagnosed, the survival of the patient was estimated at the end of a minimum of 8 months of sleep, without a maximum time limit. The data used to perform the study were obtained from the Cancer Imaging Achieve (TCIA) website. The Cancer Imaging Archive (TCIA) is an archive of medical images and clinical data based on the work of Martin Valliere’s at the Department of Medical Physics at McGill University. Machine learning methods used in working; Artificial Neural Network, Decision Tree, Linear Discriminant, Logistic Regression, Nearest Neighbor Classifier, Linear Support Vector Machine, and Quadratic Support Vector Machine. The main goal of this study is to be able to provide a decision aid to understand the survivability of head and neck cancer patients and evidence-based cancer management with the applied prediction methods and results. Evidence-based medicine and evidence-based health care are the focus of modern clinical medicine. This study may also contribute to cancer management. While the scope of this article is limited to cases of head and neck cancers, the machine learning algorithms and methodologies used are also suitable for other cancer management practices [16,17].

2. MATERIALS AND METHODS

2.1 Data set

The data were obtained from the Cancer Imaging Achieve (TCIA) website. The dataset consists of FDG-PET / CT and radiotherapy planning CT imaging and clinical data of 300 head and neck cancer (H & N) patients from four different hospitals in Québec province of Canada. Head and neck cancers of 300 patients in the data set are histologically proven. FDG-PET / BT scans were performed between April 2006 and November 2014 on average 18 days before the start of treatment for all patients. In 93 (31%) of 300 patients, radiotherapy treatment was performed by direct radiation oncologists and FDG-PET / BT imaging was performed. These image data were then used for treatment planning. Radiotherapy (16%) was administered alone to 48 of 300 patients. 252 of 300 patients were treated with chemo + radiation (84%) as part of treatment management for remediation. The median follow-up of all patients is 43 months. During the follow-up period, patients with no local or

recurrent metastases and less than 24 months of follow-up were removed from the study.

TABLE IA  
ENUMERATION OF THE DATASET

	Label	Number
Sex	Male	1
	Female	2
TNM group stage	Stage I	20
	Stage II	21
	Stage IIB	22
	Stage III	23
	Stage IV	24
	Stage IVA	25
	Stage IVB	26
Primary Site	Larynx	3
	Nasopharynx	4
	Oropharynx	5
	Hypopharynx	6
	Unknown	7
HPV Status	-	27
	+	28
	N/A	29
Therapy	chemo + radiation	37
	radiation	38
	TBD	500

TABLE IB  
ENUMERATION OF THE DATASET

	Label	Number
T - Stage	Tx	8
	T1	9
	T2	10
	T3	11
	T4	12
	T4a	30
	T4b	31
N - Stage	N0	13
	N1	14
	N2	33
	N2a	15
	N2b	16
	N2c	17
	N3	32
	N3a	34
	N3b	35
M - Stage	M0	18
	M1	19
	Mx	36
Survival	Dead	1
	Alive	0

TABLE II  
SUMMARY OF DATA SET

	Data Name	Range
Inputs	Sex	1...2
	Age	18...90
	Primary Site	3...7
	Tstage	8...31
	Nstage	13...35
	Mstage	18...36
	TNM group stage	20...26
	HPV status	27...29
	Time diagnosis to PET (days) (TDP)	-203...108
	Time diagnosis to CT sim (days) (TDCT)	-210...500
	Time diagnosis to start treatment (days) (TDS)	-195...128
	Time diagnosis to end treatment (days) (TDE)	-265...458
	Therapy	37...38
	Locoregional	0...1
	Distant	0...1
Output	Death	0...1



During the follow-up period, 45 patients (15%) developed locoregional recurrence. Forty patients developed distant metastases. Fifty-six patients (19%) lost their lives [16,17]. 298 of the dataset were used for training and testing in the estimation models. 15% of the data set was used for testing purposes, 15% for verification purposes and the remaining 70% was used to train the models. To use the data set in the models, numbering is done as in Table 1a and Table 1b. The same data set was used in all models in the study. The summary of the data set used in the study is shown in Table 2 together with the input and output data.

### 2.2 Artificial Neural Networks (ANN)

Artificial neural networks are a mathematical machine learning method that simulates the human brain. ANN is a form of artificial intelligence used for estimation purposes in many application fields. To analyze complex systems, artificial neural networks are generally used. Neural networks are widely used for classification and survival predictions in medical research in the last 20 years. Artificial neural networks are thought to be more influential than statistical methods in that they facilitate not only classification but also decision making [10,15,18]. There are many ANN studies on survival analysis, survival prediction, classification and other diagnostic and therapeutic approaches to diseases. Artificial neural networks provide a more flexible survival time estimate than conventional methods since they can easily account for variable interactions and form a nonlinear prediction model [19-29].

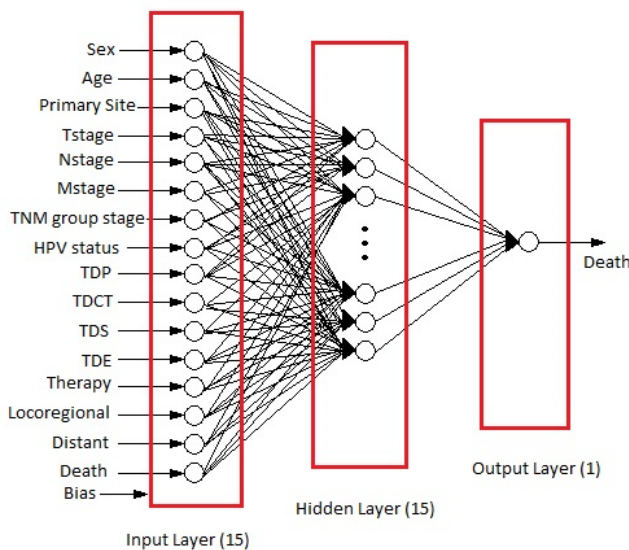


Fig.1. Architecture of the proposed ANN model

The backpropagation learning algorithm is used to train the artificial neural network model used in this study. In the proposed ANN model, the number of neurons in the hidden layer is 15. There are 15 inputs at the input layer and one output at the output layer. The architecture of the proposed ANN model is shown in Figure 1. The transfer function used in the ANN model is the hyperbolic tangent sigmoid transfer function shown in equation (1). 70% of the data set used in the study was used to train the model, 15% of the remaining data was used for

model testing and 15% was used for validation purposes. The test and validation data were randomly selected.

$$\text{tansig}(n) = \frac{2}{1+\exp(-2n)} - 1 \quad (1)$$

### 2.3 Decision Tree (DT)

The decision tree is a commonly used data mining approach to classification and estimation. Although other methods such as neural networks can be used for classification, the decision tree provides an advantage for decision-makers in terms of ease of interpretation and intelligibility. Classification of the data using the DT technique is a two-step process, learning and classification. In the learning step, a previously known training data is analyzed by the classification algorithm to form a model. The learned model is shown as a classification rule or decision tree. In the classification step, the test data is used to determine the correctness of the classification rules or decision tree. If accuracy is acceptable, rules are used to classify new data [30-32].

In this study, CART decision tree algorithm is used in the MATLAB © environment. The CART algorithm can be used as a solution to classification and regression problems since it can accept both numerical and nominal data types as input and estimation variables. The CART decision tree has a structure that is divided into two recursively. The CART tree benefiting from the Gini index as a branching criterion grows continuously by dividing, without any stopping in the establishment phase. In the stage where a new division is not going to take place, pruning starts from the tip to the root. The most successful decision tree possible is tried to be determined by evaluating with a test data independently selected after each pruning operation [30-34]. In the study, 15% of the data set was used for the test.

### 2.4 Linear Discriminant Analysis (LDA)

Discriminant analysis is a classification method. It assumes that different classes produce data based on different Gaussian distributions. To train a classifier, the fit function estimates a Gaussian distribution parameter for each class. To estimate the classes of new data, the trained classifier identifies the class with the lowest false classification cost. Discriminant analysis is a statistical technique that performs the assignment of a unit that is measured over a certain number of known masses. When this assignment is made, an error is made according to the observation value it receives when a unit is assigned to a different mass. In the discriminant analysis, this error is called the error rate or the probability of incorrect classification. The purpose of discriminant analysis is to make the assignment process with a minimum of errors. Linear discriminant analysis is also known as the Fisher separator termed by the inventor Sir R. A. Fisher. Linear Discriminant Analysis (LDA) is a classification method used in statistic, pattern recognition and machine learning to find linear combinations of properties. Although LDA is simple, it is a model that produces good results in complex problems [31,35-39].

LDA is also an important statistical tool for feature extraction and size reduction. The basic tenet of LDA is to reflect the high-dimensional data in a low-dimensional space, to minimize the



distance within the classroom, to maximize the distance between classes, and then to maximize class separation [39, 40]. A number of discrimination vectors are obtained in the LDA method. These discrimination vectors maximize the 'between classes distribution matrix' ( $S_b$ ) while minimizing the in-class distribution matrix ( $S_w$ ) [41].

Suppose that an A data matrix is given as follows;

$$A = [a_1, a_2, a_3, \dots, a_n] \in R^{m \times n}$$

, and;

$a_i \in R^m$  ;  $i = 1, \dots, n$  ;  $a_i$  is the  $i$ th data sample.

Considering the binary classification example;

Let  $n_0$  be the number of samples with zero class, let  $n_1$  be the number of samples in class 1, and we can express the sum of both classes as;  $\sum_{i=0}^1 n_i = n$

A data matrix;  $A = [A_0, A_1]$  and suppose that  $A_i \in R^{m \times n_i}$  covers  $n_i$  data samples of the  $i$ th class. In Linear Discriminant Analysis, two matrices called  $S_b$  and  $S_w$  can be defined as:

$$S_w = \sum_{i=0}^1 \sum_{a_j \in A_i} (a_j - m_i)(a_j - m_i)^T \quad (2)$$

$$S_b = \sum_{i=0}^1 n_i (m_i - m_T)(m_i - m_T)^T \quad (3)$$

Where  $m_i$  is the mean of the  $i$ th class.  $m_T$  is the total average of all data samples. LDA uses these two matrices to find the optimal sequence of discriminant vectors that maximize Fisher's criterion [40-42].

In this study, Fisher's Linear Discriminant Analysis is used in MATLAB Classifier environment. For the LDA prediction model, 15 inputs are assigned as predictors and one data is assigned as outputs or response. 15% of the data set is reserved for testing as it is in other models.

### 2.5 Logistic Regression Classifier (LRC)

Regression analysis in statistics is a method used to determine the causal relationship between a variable and other variables. The variable is divided into X and Y variables. Variable X (x-axis) is named with various terms such as descriptive variables and independent variables. The Y variable is known as the affected variable and the dependent variable. Both of these variables can be random variables, but the affected variables must always be random variables [40-42].

Regression analysis is one of the statistical methods that have proven to be extremely reliable. Logistic regression is a popular, nonlinear, statistical model in which a flexible logistic function is introduced to form the basic mathematical form of the logistic model. Logistic regression analysis is a regression analysis as well as a differential analysis technique. In the logistic regression model, the dependent (binary) variable is a discrete variable such as 0, 1; risk-indicating case 1, other cases 0. In regression problems, the key value is the mean value of the dependent (result) variable, depending on the value of a given independent variable. This value is called the conditional average and is denoted by  $E(Y \setminus x)$ . Here Y is the dependent variable and x is the independent variable. In linear regression analysis, it is assumed that the conditional mean is a linear equation of x. The logistic regression model is very efficient if the outputs are binary. In logistic regression analysis, the

corresponding conditional mean function is as follows when the output Y is binary and the variables X are real numbers [42-44].

$$E(Y \setminus X) = x = \frac{\exp(\alpha^* + \beta x)}{1 + \exp(\alpha^* + \beta x)} \quad (4)$$

Here,  $\alpha^*$  and  $\beta$  are scalar parameters. A multivariate logistic regression model was used in this study. Logistic models with multiple independent variables are called multivariate logistic regressions. Structurally, this model is not different from many other variable regression models, and interpretation of the regression coefficients is different. Interpretation depends on the type of independent variable. Non-continuous variables in a multivariate logistic regression may be nominal (classifiable) and ordinal (sortable) variables. Design variables can be used in order to put intermittent and nominal scale-independent variables into the equation. The model used in the study was obtained in the MATLAB classifier environment and 15 independent variables were used as predictors [43-45].

### 2.6 K-Nearest Neighbor Classification (KNN)

The KNN classification method is one of the classical and popular classification approaches. The KNN classification method is used in different areas due to its simplicity and effectiveness. The K-Nearest Neighbor classification algorithm, which is briefly referred to as KNN, is based on the principle that "objects close to each other probably belong to the same category". An object that is unknown to which class belongs is called a test example. Pre-classified objects are called learning examples. In the KNN algorithm, the distances from the test sample to the learning samples are calculated and if the nearest k samples belong to which class, the test instance is considered to belong to that class. In the KNN algorithm, the samples in the training set are specified by n-dimensional numerical properties. All training samples are held in a n-dimensional sample space so that each sample represents a point in n-dimensional space. When an unknown instance is encountered, the class tag of the new instance is assigned by determining the k instances closest to the relevant instance from the training set, according to the majority vote of the class labels of the nearest neighbour 'k' [46-49]. In this study, KNN classification method was performed in MATLAB environment. In KNN model, k coefficient 1 is taken and Euclidean distance criterion is used for distance. The processing steps of the KNN classification algorithm can be summarized as follows:

*Step 1:* The distances of the test sample to the learning samples are calculated.

*Step 2:* Select the closest k samples.

*Step 3:* If the number of samples belonging to which class is the greatest, the test sample is also assigned to this class.

### 2.7 Support Vector Machines (SVM)

The SVM has the ability to separate with linear separators in two-dimensional space and planar separators in three-dimensional space two or more classes. The working principle of SVM is to estimate the most appropriate decision function that can distinguish between two classes.

In other words, the basic principle is to define the hyperplane, which can distinguish between the two classes in the most appropriate way [50-54]. SVM is used for classification and estimation purposes. Especially in the field of medicine, many articles have been published for purposes such as diagnosis and classification of diseases [58,59].

2.7.1 Linear Support Vector Machines (LSVM)

In SVM classification, it is aimed to separate samples of two classes, which are usually shown as  $\{-1, +1\}$ , with the help of a decision function. By using the decision function, it is necessary to find a hyperplane which can best distinguish the training data. As shown in Fig. 2a, many hyperplanes can be plotted which can distinguish two-class data.

However, the SVM's goal is to find the hyperplane that maximizes the distance between its nearest points. The support vectors and the optimal hyperplane are shown in Figure 2b. The 'optimum hyperplane', which makes the most appropriate difference by raising the limit to maximum, is shown in Figure 2c. In Figure 2c, the points that limit the border width are called support vectors. The support vectors are expressed in the form of  $w \cdot x_i + b = \pm 1$ . The limits of the optimal hyperplane must be maximized. Lagrange equations are used for this. The decision function for LSVM can be written as in equation (5) [58,59].

$$f(x) = \text{sign}(\sum_{i=1}^k \lambda_i y_i(x \cdot x_i) + b) \quad (5)$$

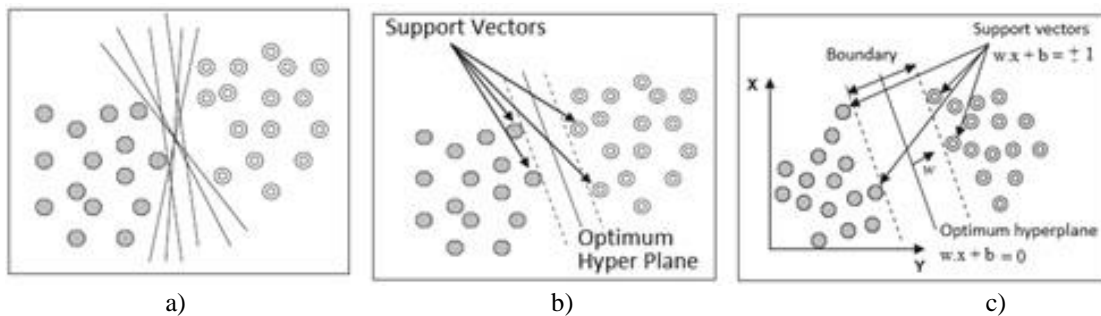


Fig.2. Support vectors and hyperplanes

2.7.2 Non-linear SVM

For many data sets, the data in the two-dimensional state cannot be separated by the help of linear delimiters. This is seen in Figure 3a. In this case, the problem arising from the fact that some of the training data remain on the other side of the optimum hyperplane is solved by defining a positive artificial variable ( $\xi_i$ ) [Fig. 3b]. The balance between maximizing the

boundary and minimizing the classification errors is controlled by an adjustment parameter indicated by C. The C adjustment parameter is always positive ( $0 < C < \infty$ ) [60-65]. As can be seen in Figure 3c, data that cannot be linearly separated in the input space is displayed in a high-dimensional space defined as the property space. Thus, the data can be linearly discriminated and the hyperplane between classes can be determined.

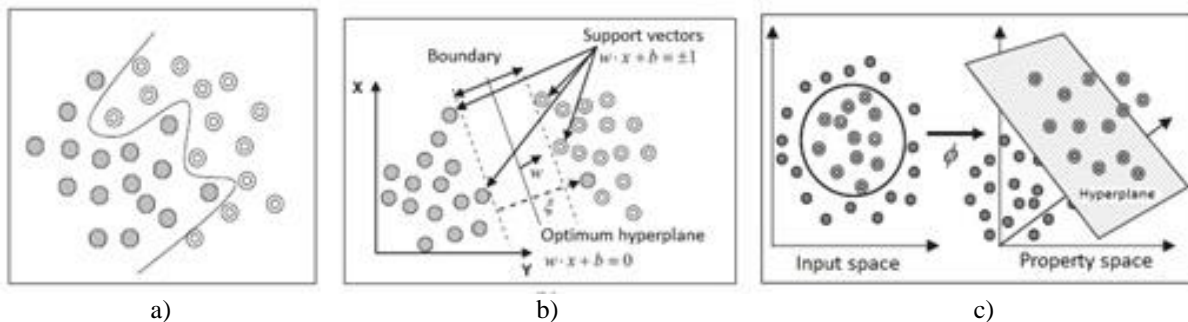


Fig.3. Hyperplanes for non-linear SVM

Equation (6) shows the Kernel function. With kernel function SVM can perform non-linear transformations [58-64].

$$K(x_i, x_j) = \varphi(x) \cdot \varphi(x_j) \quad (6)$$

The decision to solve a two-class problem that cannot be linearly separated using the kernel function can be written as in equation (7) [58-64].

$$f(x) = \text{sign}(\sum_i \alpha_i y_i \varphi(x) \cdot \varphi(x_i) + b) \quad (7)$$

The kernel function to be used for a classification operation with SVM, and the determination of optimum parameters for this function are essential. Besides the parameters specific to the kernel function, the configuration parameter 'C' for all support vector machines must be specified by the user. For this parameter, if too small or too large values are selected, a serious reduction in the classification accuracy is expected, since the optimal hyperplane cannot be determined correctly. On the

other hand, if  $C = \infty$ , the SVM model is only suitable for data sets that can be linearly separated. As can be seen here, the selection of appropriate values for the parameters is a factor that directly affects the performance of the SVM classifier. Despite the use of trial and error strategies, the cross-validation approach allows successful results to be achieved. The purpose of the cross-validation approach is to determine the performance of the generated classification model. For this purpose, the dataset is divided into two parts. The first part is used as training data in the modelling which is the basis of classification, while the second part is processed as test data to determine the performance of the model. As a result of applying the model created by the training set to the test data set, the number of correctly classified samples shows the performance of the classifier. Therefore, by using the cross-validation method, the best classification performance is obtained and the model to be the basis of classification is determined by determining the kernel parameters [58,59].

An important issue to be considered in SVMs is the fact that large data groups have more than one cluster, depending on their particular characteristics. In order to be able to use SVM in multiple class situations, the problem must be transformed into a large number of binary class problems. The most commonly used approaches are the “One vs All” approach and the “One vs. One” approach [60-65]. This study was carried out in MATLAB environment and both "Quadratic SVM" and "Linear SVM" classification were realized. The classification approach used is the "One vs One" approach.

### 3. RESULTS

Seven machine learning models were designed to predict the survival time for head and neck cancer in the study. The ROC (Receiver Operating Characteristic) curves obtained for these seven models are shown in Figures 4a, b, c, d, e, f, and g.

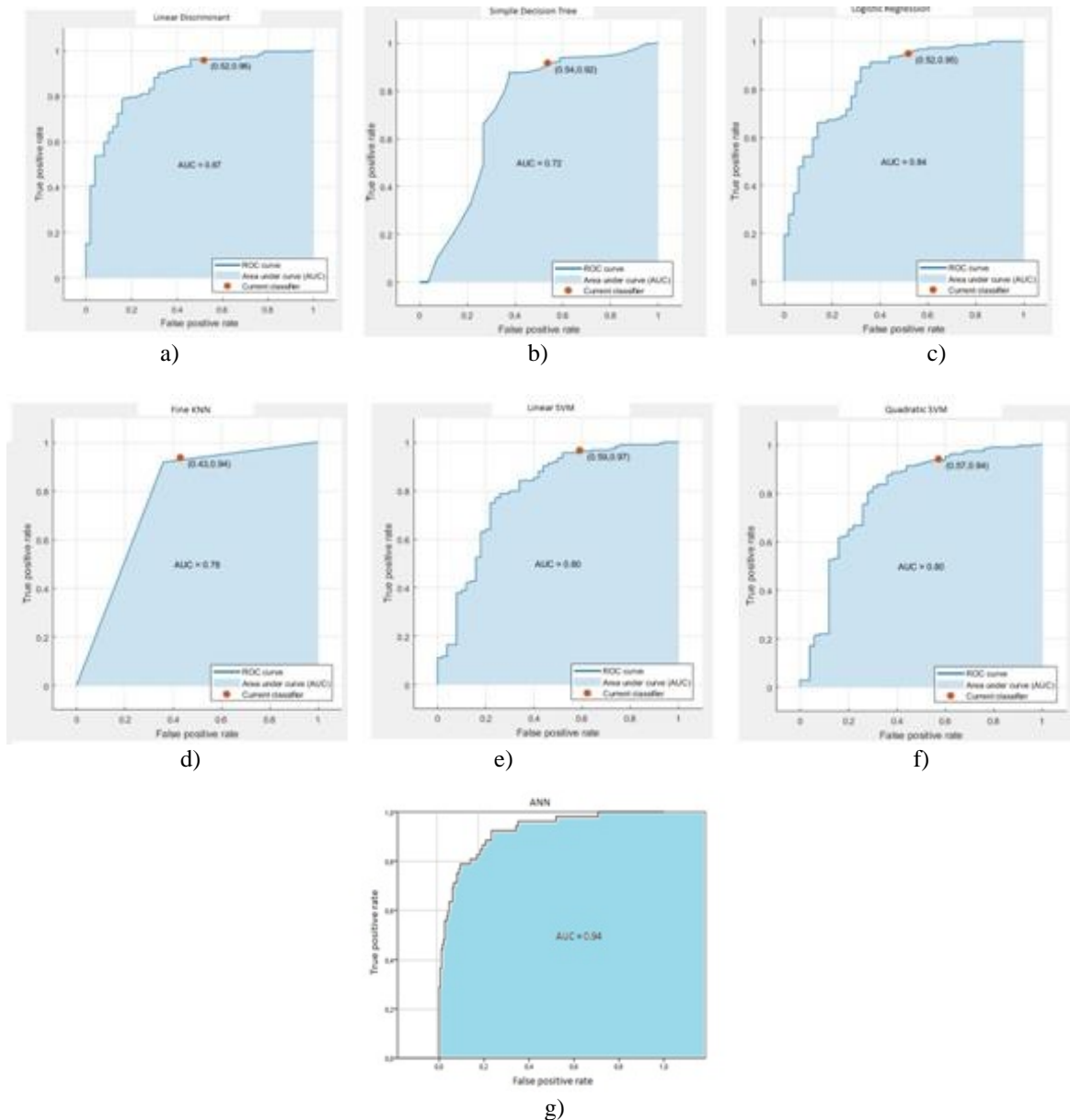


Fig.4. The ROC curves; a) LDA, b) DT, c) LRC, d)KNN e) LSVM, f) QSVM, g) ANN.

In Figure 5, ROC curves of all models are shown on the same axis. In addition, the confusion matrices obtained from the classification models designed in Figure 6 are shown. The classification results obtained are shown in Table 3.

TABLE III  
CLASSIFICATION RESULTS OF THE MODELS

Model Parameters	Machine learning methods						
	LDA	DT	LR	KNN	LSVM	QSVM	ANN
AUC	0.87	0.72	0.84	0.78	0.80	0.80	0.94
Accuracy (%)	86.9	83.2	86.2	86.9	86.2	84.6	90
Training Time (s)	3.1346	0.8772	11.667	1.9018	4.6418	1.3179	0.4

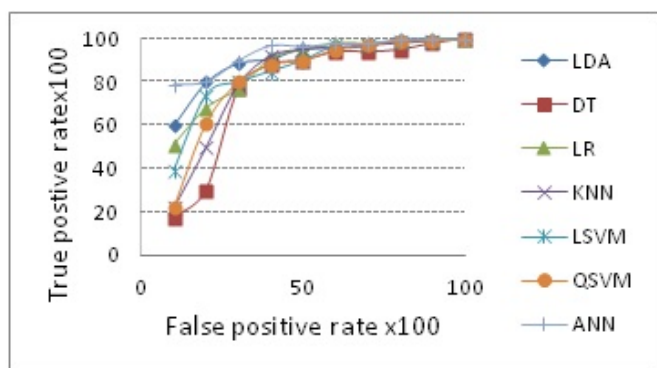


Fig.5. ROC curves of all models

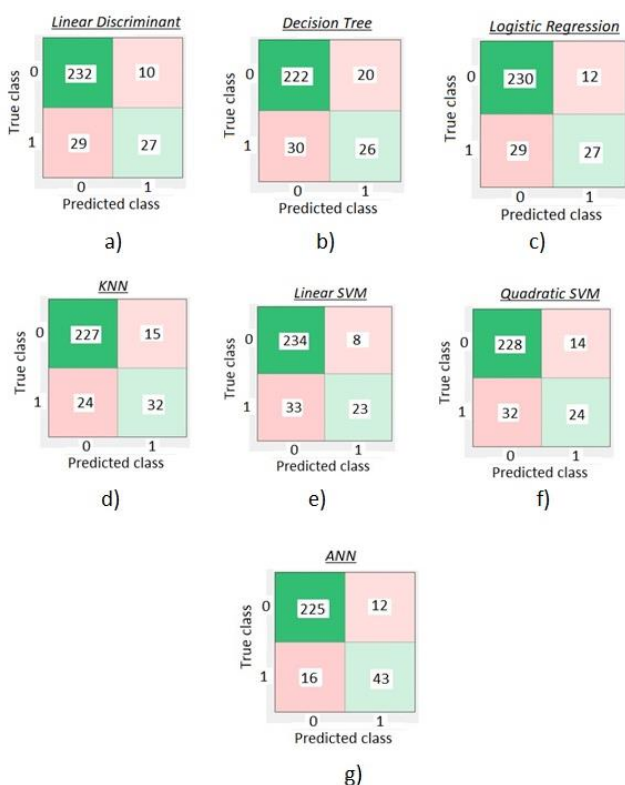


Fig.6. Confusion matrices for all classification models

A performance curve was also drawn to evaluate the success of the designed artificial neural network model. The performance

curve drawn is shown in Figure 7b. From this curve, the mean square error can also be observed.

TABLE IV  
RESULTS OF ANN MODEL

	Samples	MSE	R
Training	208	7.67845e-2	7.10192e-1
Validation	45	1.58386e-1	6.20647e-1
Testing	45	9.50916e-2	7.04439e-1

In addition, mean square error and predicted values are shown in Table 4. The regression curves of the test and validation data obtained from the artificial neural network model are given in figure 7a. From Figure 8, the success status of the ANN model in the testing process can be interpreted.

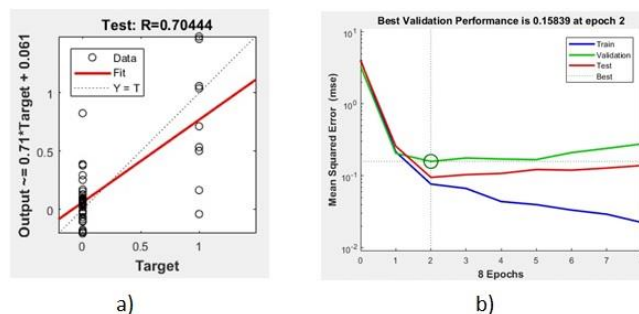


Fig.7. a) Regression and b) performance, curves of proposed ANN model

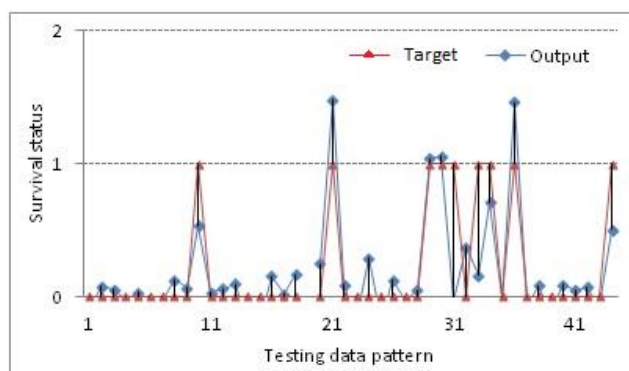


Fig.8. Comparison of the target and output of ANN for testing

#### 4. CONCLUSIONS

When the obtained graphs and confusion matrices were examined, survival estimates were made for 298 cases in total. Separate results can be obtained from the application of each method.

First, from the linear discriminant analysis, it is clear that the prediction that 232 people will survive is correct. It was also correctly predicted that 27 people would die. Despite these correct estimates, the claim that it will survive for 29 people has also come out as a false prediction. In the same way, the prediction that 10 people will die is not true. In this case, for the head and neck cancer patients, survival was estimated with an accuracy of 86.9% by the "linear discriminant classification" method. This rate is extremely satisfactory.

When estimating and classifying by decision tree method was examined, it was estimated that 222 people would survive correctly. It was also estimated correctly that 26 people will die.



Though 20 people were estimated to die, they lived and were estimated incorrectly. Though 30 people were thought to live, they died. In this case, the ratio of correct estimates is 83.2% as can be understood from Table 3. Accuracy and other results may be considered to be generally successful, albeit worse than LDA. It is also very important how many seconds these results are reached and how long these estimates are valid for. As Table 3 reveals, these results were achieved at an extremely short time of 0.8 seconds. These estimates were made for the time until the first control of all patients, i.e. for a minimum of 8 months. In this case, it is easy to say that the decision tree method is also successful.

For another classification method, logistic regression, there are 230 cases that are thought to be alive and predicted correctly. The number of patients correctly estimated to die is 27. In this method, it can be seen that there are mispredictions, as shown in figure 6c. In this case, accuracy is close to LDA, can be said to be extremely successful with 86.2%

The number of surviving patients correctly estimated by the KNN method is 227, and the number of correctly estimated and lost patience is 32. The survival of 39 patients was estimated incorrectly. In this case, it can be seen in Table 3 that the accuracy rate is the same as the linear discriminant with 86.9%. However, when the ROC curve is examined, it can be observed that AUC is 87% in linear discriminant and 78% in KNN. In this case, it is possible to say from the ROC analysis that LDA is more preferable than KNN if it is considered that the KNN method is successful.

Considering the linear SVM, 234 patients in the confusion matrix are correctly predicted to survive. This is the largest number in all other methods. The number of patients who are correctly predicted to lose their lives is 23. This is the smallest number among other methods. The number of patients who are living despite being estimated incorrectly, that is, estimated to die, is only 8. This figure is the lowest and successful figure among other methods. Despite this successful outcome, however, one negative outcome is 33 patients who died despite the fact that they were expected to survive. In this case, the rate of accurate estimates is very successful with 86.2%.

For the Quadratic SVM model, the number of surviving patients correctly estimated is 228. Figure 6f shows that other estimation results are similar to those of other methods. It can be said that the accuracy rate is very good with 84.6%.

When the results obtained from the ANN approach are considered; the number of patients correctly estimated to live is 225, the number of patients who are correctly predicted to lose their lives is 43. Despite being predicted to live, the number of patients who actually lost their lives is 16, which is the most successful value. Despite being estimated to die, the number of patients living is 12. A correctly estimated survival rate of 90% can be said to be the most successful. However, even though the estimates appear to be close to real values, there can be observed differences in the ANN's figure 9 comparative curve. In order to perform ROC analysis, it is necessary to round the estimated values according to the output. The Confusion matrices are the result of these rounds. The accuracy rate is only 70% for the test data, as can be seen from the regression curve in figure 7a and the performance curve in figure 7b without rounding. However, when the outputs are considered to be 1

and 0, the accuracy rate is 90% since values close to these output values are rounded to these values.

These seven classification and prediction models used in the study can be compared with each other. In this case, it can be concluded that the most successful model is ANN and the other methods can be considered successful with at least 80% accuracy. In conclusion, using clinical data of head and neck cancers, survival estimates with machine learning approaches were obtained with at least 83% accuracy and at most 90% accuracy.

#### ACKNOWLEDGMENT

This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors. The author would like to thank Dr. Martin Vallières from the Medical Physics Unit of McGill University, and TCIA

#### REFERENCES

- [1] Allison, D. B., Maleki, Z. (2016). HPV-related head and neck squamous cell carcinoma: Anupdate and review, *Journal of the American Society of Cytopathology*, 5, pp.203-215.
- [2] Maund, I., Jefferies, S. (2015). Squamous cell carcinoma of the oral cavity, oropharynx and upper oesophagus, *Medicine*, 43, 197-201.
- [3] McGurk, M., Goodger, N. M. (2000). Head and neck cancer and its treatment: historical review, *British Journal of Oral and Maxillofacial Surgery*, 38, pp.209-220.
- [4] Galbiatti, A. L. S., Junior, J. A. P., Maniglia, J. V., Rodrigues, C. D. S., Pavarino, É. C., Bertollo, E. M. G. (2013). Head and neck cancer: causes, prevention and treatment, *Braz J Otorhinolaryngol*, 79, pp.239-47.
- [5] Young, D., Xiao, C. C., Murphy, B., Moore, M., Fakhry, C., Day, T. A. (2015). Increase in head and neck cancer in younger patients due to human papillomavirus (HPV), *Oral Oncology*, 51, pp.727-730.
- [6] Jemal, A., Bray, F., Center, M. M., Ferlay, J., Ward, E., Forman, D. (2011). Global cancer statistics, *CA Cancer J Clin*, 61, 69-90.
- [7] Zini, E. M., Lanzola G., and Quaglini, S. (2017). Detection and Management of Side Effects in Patients with Head and Neck Cancer, *IEEE 3rd International Forum on Research and Technologies for Society and Industry (RTSI)*, pp.1-6.
- [8] Drago, G. P., Setti, E., Licitra, L., and Liberati, D. (2002). Forecasting the Performance Status of Head and Neck Cancer Patient Treatment by an Interval Arithmetic Pruned -Perceptron, *IEEE Transactions on Biomedical Engineering*, 49, 782-787.
- [9] Shukla, N., Hagenbuchner, M., Wi, K. T., Yang, J. (2018). Breast cancer data analysis for survivability studies and prediction, *Computer Methods and Programs in Biomedicine*, 155, 199-208.
- [10] Walczak, S., Velanovich, V. (2017). An Evaluation of Artificial Neural Networks in Predicting Pancreatic Cancer Survival, *J Gastrointest Surg.*, 21, pp.1606-1612.
- [11] Wróbel, L., Gudy's A., and Sikora, M. (2017). Learning rule sets from survival data, *BMC, Bioinformatics* 285 DOI 10.1186/s12859-017-1693-.
- [12] Lynch, C. M., Abdollahib, B., Fuquac, J. D., de Carloc, A. R., Bartholomaic, J. A., Balgemann, R. N., van Berkeld, V. H., Frieboesc, H. B. (2017). Prediction of lung cancer patient survival via supervised machine learning classification techniques, *International Journal of Medical Informatics*, 108, pp.1-8.
- [13] Wu, C., Wu, Y., Liang, P., Wu, C., Peng, S. F., Chiu, H. W. (2017). Disease-free survival assessment by artificial neural networks for hepatocellular carcinoma patients after radiofrequency ablation, *Journal of the Formosan Medical Association* 116, pp.765-773.
- [14] Walczak, S., Velanovich, V. (2017). Improving prognosis and reducing decision regret for pancreatic cancer treatment using artificial neural networks, *Decision Support Systems*, doi: 10.1016/j.dss.2017.12.007.
- [15] Walczak, A. S., Taktak, A. G. F., Helliwell, T. R., Fenton, J. E., Birchall, M. A., Husband, D. J., Fisher A. C. (2006). An artificial neural network improves prediction of observed survival in patients with laryngeal squamous carcinoma, *Eur Arch Otorhinolaryngology*, 263, pp.541-547.



- [16] Vallières, M. et al. (2017). Radiomics strategies for risk assessment of tumour failure in head-and-neck cancer, *Sci Rep* 10117 doi: 10.1038/s41598-017-10371-5.
- [17] Cancer Imaging Archive (2018). <http://www.cancerimagingarchive.net/>, last accessed date: February 10, 2018.
- [18] Ravdin, P. M., and Clark, G. M. (1992). A practical application of neural network analysis for predicting outcome of individual breast cancer patients, *Breast Cancer Research and Treatment*, 22, 285-293.
- [19] Chi, C. L., Street, W. N., Wolberg, W. H. (2007). Application of Artificial Neural Network-Based Survival Analysis on Two Breast Cancer Datasets, *AMIA Annu Symp Proc.*, pp.130-134.
- [20] Dolgobrodov, S. G., Moore, P., Marshall, R., Bittern, R., Steele, R. J. C., Cuschieri, A. (2007). Artificial Neural Network: Predicted vs. Observed Survival in Patients with Colonic Cancer, *Diseases of the Colon & Rectum*, 50, pp.184-191.
- [21] Ahmed, F. E. (2005). Artificial neural networks for diagnosis and survival prediction in colon cancer, *Molecular Cancer*, 29, doi:10.1186/1476-4598-4-29.
- [22] Devi, M. A., Ravi, S., Vaishnavi, J., and Punitha, S. (2016). Classification of Cervical Cancer using Artificial Neural Networks, *Procedia Computer Science*, 89, 465-472.
- [23] Ripley, R. M., Harris A. L., and Tarassenko, L. (1998). Neural Network Models for Breast Cancer Prognosis, *Neural Comput & Applic*, 7, pp.367-375.
- [24] Shukla, R. S., Aggarwal, Y. (2017). Nonlinear Heart Rate Variability based artificial intelligence in lung cancer prediction, *Journal of Applied Biomedicine*, Vol.16, No.2, pp.145-155. doi:10.1016/j.jab.2017.
- [25] De Laurentiis, M., and Ravdin, P. M. (1994). Survival analysis of censored data: Neural network analysis, detection of complex interactions between variables, *Breast Cancer Research and Treatment*, 32, pp.113-118.
- [26] Ochi, T., Murase, K., Fujii, T., Kawamura, M., Ikezoe, J. (2002). Survival prediction using artificial neural networks in patients with uterine, cervical cancer treated by radiation therapy alone, *Int J Clin Oncol* 7, pp.294-300.
- [27] Asria, H., Mousannif, H., Al Moatassime, H., Noel, T. (2016). Using Machine Learning Algorithms for Breast Cancer Risk, Prediction and Diagnosis, *Procedia Computer Science*, 83, pp.1064 - 1069.
- [28] Francis, N. K., Luther, A., Salib, E., Allanby, L., Messenger, D., Allison, A. S., Smart, N. J., Ockrim, J. B. (2015). The use of artificial neural networks to predict delayed discharge and readmission in enhanced recovery following laparoscopic, colorectal cancer surgery, *Tech Coloproctol*, 19, pp.419 - 428.
- [29] Iraj, M. S. (2017). Prediction of post-operative survival expectancy in thoracic lung cancer surgery with soft computing, *Journal of Applied Biomedicine*, 15, pp.151-159.
- [30] Chien, C. F., Chen, L. F. (2008). Data Mining to Improve Personnel Selection and Enhance Human Capital: A Case Study in High-Technology Industry, *Expert Systems with Applications*, 34, pp.280-290.
- [31] Discriminant analysis (2018). <https://www.mathworks.com/help/stats/discriminantanalysis.html>, last accessed date: February, 14, 2018.
- [32] Zheng, H., Chen, L., Han, X., Zhao, X., Ma, Y. (2009). Classification and regression tree (CART) for analysis of soybean yield variability among fields in Northeast China: The importance of phosphorus application rates under drought conditions, *Agriculture, Ecosystems & Environment*, 132, pp.98-105.
- [33] Breiman, L., Friedman, J. H., Olshen R. A., and Stone, C. J. (1984). *Classification and Regression Trees*, Chapman and Hall, New York, USA.
- [34] Stephen, E. F., Hsieh, Y., Rivadineria, A., Beer, T. M., Mori, M., Garzotto, M. (2006). Classification and Regression Tree Analysis for the Prediction of Aggressive Prostate Cancer on Biopsy, *The Journal of Urology*, 175, pp.918-922.
- [35] Fisher, R. A. (1936). The Use of Multiple Measurements in Taxonomic Problems, *Annals of Eugenics*, 7, pp.179-188.
- [36] Lachenbruch, P. A. (1975). *Discriminant analysis*, Hafner Press, New York, USA.
- [37] Lachenbruch, P. A., and Mickey, M. R. (1968). Estimation of error rates in discriminant analysis, *Technometrics* 10, pp.1-11.
- [38] Elkhali, K., Kammoun, A., Couillet, R., Al-Naffouri, T. Y., and Alouini, M. S. (2017). Asyptotic Performance of Regularized Quadratic Discriminant Analysis Based Classifiers, 2017 IEEE International Workshop on Machine Learning for Signal Processing Tokyo, pp.25-28.
- [39] Cai, J., and Huang, X. (2018). Modified Sparse Linear-Discriminant Analysis via, Nonconvex Penalties, *IEEE Transactions on Neural Networks and Learning Systems*, Early Acces, pp.1-10.
- [40] Lee, Y., Madayambath, S. C., Liu, Y., Da-Ting, L., Chen, R. and Bhattacharyya, S., S. (2017). Online Learning in Neural Decoding Using Incremental Linear Discriminant Analysis, *IEEE2017 IEEE International Conference on Cyborg and Bionic Systems Beijing China*, pp.173-177.
- [41] Lawi, A., La Wungo, S., Manjang, S. (2017). Identifying Irregularity Electricity Usage of customer Behaviors using Logistic Regression and Linear Discriminant Analysis, *IEEE 3rd International Conference on Science in Information Technology (ICSITech)*, pp.552-557.
- [42] Tsangaratos, P., Iliia, I. (2016). Comparison of a logistic regression and Naïve Bayes classifier in landslide susceptibility assessments: The influence of models complexity and training dataset size, *Catena*, 145, pp.164-179.
- [43] Geng, P., Sakhnenko, L. (2016). Parameter estimation for the logistic regression model under case-control study, *Statistics and Probability Letters*, 109, 168-177.
- [44] Razanamahandry, L. C., Andrianisa, H. A., Karoui, H., Podgorski, J., Yacouba, H. (2018). Prediction model for cyanide soil pollution in artisanal gold mining area by using logistic regression, *Catena*, 162, pp.40-50.
- [45] Zhou, C., Wang, L., Zhang, Q., Wei, X. (2014). Face recognition based on PCA and logistic regression analysis, *Optik*, 125, pp.5916-5919.
- [46] Duca, A., Bacciu, C., Marchetti, A. (2017). A K-Nearest Neighbor Classifier for Ship Route Prediction, *IEEE OCEANS - Aberdeen*, pp.1 - 6.
- [47] Yu, Z., Chen, H., Liu, J., You, J., Leung, H., and Han, G. (2016). Hybrid, k-Nearest Neighbor Classifier, *IEEE Transactions Cybernetics*, 46, pp.1263-1275.
- [48] Li, W., Du, Q., Zhang, F., Hu, W. (2014). Collaborative Representation Based K-Nearest Neighbor Classifier for Hyperspectral Imagery , *Hyperspectral Image and Signal Processing: Evolution in Remote Sensing* , WHISPERS DOI: 10.1109/WHISPERS.2014.8077601.
- [49] Cover, T. M., and Hart, P. E. (1967). Nearest neighbor pattern classification, *IEEE Transactions on Information Theory*, 13, pp.21-27.
- [50] Support Vector Machines (2018). <https://www.mathworks.com/help/stats/support-vector-machines-for-binary-classification.html>, last accessed date February 14, 2018.
- [51] Cortes, C., Vapnik, V. (1995). Support-Vector Network, *Machine Learning*, 20, pp.273-297.
- [52] Vapnik, V. N. (2000) *The Nature of Statistical Learning Theory*, 2nd Edition, Springer-Verlag, New York.
- [53] Kavzaoglu, T., Colkesen, I. (2010). Investigation of the Effects of Kernel Functions in Satellite Image Classification Using Support Vector Machines, *Map Journal* July, 144, 73-82.
- [54] Ilias, S., Tahir, N. M., Jailani, R. (2016). Feature extraction of autism gait data using principal component analysis and linear discriminant analysis, *IEEE Industrial Electronics and Applications Conference IEACon.*, pp.275 - 279.
- [55] Gao, L., Ye, M., Lu, X., Huang, D. (2017). Hybrid Method Based on Information Gain and Support Vector Machine for Gene Selection in Cancer Classification, *Genomics, Proteomics & Bioinformatics*, 15, pp.389-395.
- [56] Wang, H., Zheng, B., Yoon, W., Ko, H. S. (2018). A support vector machine-based ensemble algorithm for breast cancer diagnosis, *European Journal of Operational Research*, 267, pp.687-699.
- [57] Ghaddar, B., Sawaya, J. N. (2018). High dimensional data classification and feature selection using support vector machines, *European Journal of Operational Research*, 265, pp.993-1004.
- [58] Madadum, H., Becerikli, Y. (2017). The implementation of Support Vector Machine (SVM) using FPGA for human detection, 10th International Conference on Electrical and Electronics Engineering ELECO, pp.1286 - 1290.
- [59] Nefedow, A., Ye, J. Ye, Kulikowski, C., Muchnik, I., Morgan, K. (2009). Comparative Analysis of Support Vector Machines Based on Linear and Quadratic Optimization Criteria, *IEEE International Conference on Machine Learning and Applications*, pp.288 - 293.
- [60] Vapnik, V. N. (1995). *The Nature of Statistical Learning Theory*, 1st Edition, Springer-Verlag, New York USA.

- [61] Machine learning methods, (2018). <https://www.mathworks.com> , last accessed date: February 12, 2018.
- [62] Osuna, E. E., Freund, R., Girosi, F. (1997). Support Vector Machines: Training and Applications, Massachusetts Institute of Technology and Artificial Intelligence Laboratory 144, Massachusetts.
- [63] Whsu, C., Lin, C. J. (2002). A Comparison of Methods for Multiclass Support Vector Machines, IEEE Transactions On Neural Networks, 13, pp.415-425.
- [64] Nogay, H.S. (2018). Classification of Different Cancer Types by Deep Convolutional Neural Networks, Balkan Journal of Electrical and Computer Engineering, Vol.6, pp.56-59.
- [65] Zini, E. M., Lanzola G., and Quaglini, S. (2017). Detection and Management of Side Effects in Patients with Head and Neck Cancer, IEEE 3rd International Forum on Research and Technologies for Society and Industry (RTSI), pp.1-6.

## B IOGRAPHIES

**H. Selçuk Noğay** received B.S degrees in Electrical Education from Kocaeli University, and M.S. and Ph.D degrees in Electrical Education from Marmara University respectively 2002, 2003 and 2008. His research interests include Artificial Neural Network, Deep Learning and signal processing technique He has been working as a Professor in Vocational Scholl of Kayseri University in Kayseri, Turkey

# A NOVEL INTERPRETABLE WEB-BASED TOOL ON THE ASSOCIATIVE CLASSIFICATION METHODS: AN APPLICATION ON BREAST CANCER DATASET

A. K. Arslan, Z. Tunc, I. Balıkcı Cicek and C. Colak


**Abstract— Aim:** The second-largest cause of cancer mortality for women is breast cancer. The main techniques for diagnosing breast cancer are mammography and tumor biopsy accompanied by histopathological studies. The mammograms are not detective of all subtypes of breast tumors, particularly those which arise and are more aggressive in young women or women with dense breast tissue. Circulating prognostic molecules and liquid biopsy approaches to detect breast cancer and the death risk are desperately essential. The purpose of this study is to develop a web-based tool for the use of the associative classification method that can classify breast cancer using the association rules method.

**Materials and Methods:** In this study, an open-access dataset named “Breast Cancer Wisconsin (Diagnostic) Data Set” was used for the classification. To create this web-based application, the Shiny library is used, which allows the design of interactive web-based applications based on the R programming language. Classification based on association rules (CBAR) and regularized class association rules (RCAR) are utilized to classify breast cancer (malignant/benign) based on the generated rules.


**Results:** Based on the classification results of breast cancer, accuracy, balanced accuracy, sensitivity, specificity, positive predictive value, negative predictive value, and F1-score values obtained from the CBAR model are 0.954, 0.951, 0.939, 0.964, 0.939, 0.964, and 0.939 respectively.


**Conclusion:** In the analysis of the open-access dataset, the proposed model has a distinctive feature in classifying breast cancer based on the performance metrics. The associative classification software developed based on CBAR produces successful predictions in the classification of breast cancer. The hypothesis established within the scope of the purpose of this study has been confirmed as the similar estimates are achieved with the results of other papers in the classification of breast cancer.

**Keywords—** Artificial intelligence, association rules, associative classification, web-based software, breast cancer.

**Ahmet Kadir ARSLAN**, Inonu University Department of Biostatistics and Medical Informatics, Faculty of Medicine, Malatya, Turkey, (arslan.ahmet@inonu.edu.tr) 

**Zeynep TUNC**, Inonu University Department of Biostatistics and Medical Informatics, Faculty of Medicine, Malatya, Turkey, (zeynep.tunc@inonu.edu.tr) 

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

**Cemil COLAK**, Inonu University Department of Biostatistics and Medical Informatics, Faculty of Medicine, Malatya, Turkey, (cemil.colak@inonu.edu.tr) 

Manuscript received May 14, 2020; accepted June, 11, 2020.  
Digital Object Identifier:

## 1. INTRODUCTION

The second most common reason for death in adult women is breast cancer, which is diagnosed annually in more than 2 million new cases. While in particular, in developing countries where 5-year survival rates have been 90 percent or higher for invasive breast cancer, breast cancer survival has increased dramatically over the previous three decades as a result of improved early detection and improved treatment. There is a sharp increase in the global incidence of breast cancer, according to the World Health Organization as a result of improvements in lifestyles, reproductive factors, and life expectancy. Among middle and low-income countries, 58 percent of all breast cancer deaths occur. While breast cancer survival rates in developed countries are approximately 80%, the rate drops to 60% in the middle- and 40% in low-income nations due to lack of early-screening programs which lead to incurable diagnoses in late-stage 80% of these tumors. Mammography and other expensive and technically complex methods cannot be done in middle and low-income countries due to high costs and shortages of trained staff. Furthermore, ER-positive breast cancer is more likely to be identified by mammograms and not indicated for younger people. Therefore, earlier diagnosis using traditional methods for all race classes are not predicted; for example, a small-sized African-American woman with metastases is more probable than a Caucasian woman. Hence, there is a significant ethnic difference in the survival of breast cancer with higher breast cancer mortality rates [1, 2].

Today, the development of computer technologies thanks to technological possibilities has led to the collection of large amounts of data in databases and to access these data more easily. As the amount of data collected grows and the complexity in the data structure collected increases, the need for much better analysis techniques also increases simultaneously. At this point, the concept of Knowledge Discovery in Databases has emerged in the past decades. Knowledge discovery is the process of finding new, previously unknown, and useful information in databases. The knowledge discovery process includes data selection, data preprocessing, data conversion, data mining, and evaluation stages. One of the important stages of the information discovery process is called data mining. Data mining is an interdisciplinary field defined as revealing the relationships and patterns hidden in the data. Data mining is the search for the relations and rules that will allow us to make predictions of a large amount of data using computer programs. According to the definition of data mining, the main purpose is to keep a large amount of data in the data warehouse and obtain meaningful information from these data [3, 4].

There are many models used in data mining. These models are examined under four main categories: association rules (ARs) analysis, classification, clustering, and predictive models [5]. ARs, which are one of the data mining models, are under the name of "association rules analysis", which has a wide usage area in many fields such as economy, education, telecommunication, and medicine. ARs are widely utilized in data mining due to their usefulness and easy understanding and are the process of finding associations, relationships, and patterns among the data as rules. ARs express the occurrence of events together with certain possibilities [6, 7].

Associative classification is a branch of scientific work, known as data mining in artificial intelligence. Associative classification combines the association rules and classification, two known methods of data mining, to create a model for predictive purposes. Specifically, associative classification is a type of classification approach that is built with a set of rules obtained by the association rule mining to form classification models. While the classification and association rules are the prediction of the class labels of the main purpose of the classification, they have similar tasks in data mining, except that the association rule is a method used to find common patterns, correlations, associations, or causal structures from datasets. In the past few years, association rules methods have been successfully used to create correct classifiers in associative classification. [8].

One of the main advantages of using a classification based on association rules according to classical classification approaches is that the output of an associative classification algorithm is represented by simple if-then rules, making it easier for the user to understand and interpret it. [8]. For this reason, the current study aims to develop a new user-friendly web-based software to realize the use of the associative classification method that can classify breast cancer using the association rules method. For this purpose, the main hypotheses of this study are to determine whether classification-based association rules models are successful in predicting breast cancer on the open-access dataset and evaluate the classification performance.

## 2. MATERIAL AND METHODS

### 2.1. Dataset

The open-access dataset named "Breast Cancer Wisconsin (Diagnostic) Data Set" was obtained from the UCI machine learning repository. The dataset consists of 569 samples examined for breast cancer with the ten predictors/inputs and one response/output variables. Of the individuals, 357 (62.7%) were diagnosed as benign, and 212 (37.3%) were diagnosed as malignant. A digitized image of a fine needle aspirate (FNA) of a breast mass is used to measure the characteristics. The characteristics of the presented cell nuclei are described in the image [9]. The explanations about the variables in the data set and their properties are given in Table 1.

TABLE I  
EXPLANATIONS ABOUT THE VARIABLES IN THE DATASET AND THEIR PROPERTIES

Variable	Variable Explanation	Variable type	Variable role
Diagnosis	The diagnosis of breast tissues (M = malignant, B = benign)	Qualitative	Output
Radius	Mean distances from the center to perimeter points	Quantitative	Predictor
Texture	The standard deviation of gray-scale values	Quantitative	Predictor
Perimeter	Mean size of the core tumor	Quantitative	Predictor
Area	-	Quantitative	Predictor
Smoothness	Mean of local variation in radius lengths	Quantitative	Predictor
Compactness	$(\text{mean of perimeter})^2 / (\text{area} - 1)$	Quantitative	Predictor
Concavity	Mean of the severity of concave portions of the contour	Quantitative	Predictor
Concave points	mean for the number of concave portions of the contour	Quantitative	Predictor
Symmetry	-	Quantitative	Predictor
Fractal dimension	mean for "coastline approximation" - 1	Quantitative	Predictor

## 3. ASSOCIATION RULES AND ASSOCIATIVE CLASSIFICATION

Data mining is often the analysis of large datasets to find unexpected relationships with the principle of being both understandable and useful for the owner of the data and summarizing the data with new methods. The relationships and summaries obtained from a data mining application are often called models or patterns. Linear equations, rules, sets, graphs, tree structures, and repetitive patterns in the time series are some of these Association rules that are among the most popular representatives of regional patterns in data mining [10].

Association rules mining is one of the unsupervised data mining tasks that look for the relationship between records in a data set. Association rules are often expressed as if it happens, then this happens. Mostly used for finding interpretable trends and relationships among variables [11]. Association rules are rules with support and confidence measurements in the form of "IF- precursor expression-, IF-successor expression" [12]. The value of support and confidence can be evaluated as units of measure showing the strength of an association rule. Confidence and support values, which are the measures of interestingness for association rules, are shown as follows [13].



D: Data,

$t_i$ : Records in data,  $D = \{t_1, t_2, \dots, t_i, \dots, t_n\}$

X, Y: Items in rules (precursor and successor)

$X \rightarrow Y$ , where X is the precursor and Y is the successor ( $X \cap Y = 0$ );

Support value (SV):

$$SV(X) = \frac{|t \in D, X \subset t|}{|D|}$$

Confidence value (CV):

$$CV(X \rightarrow Y) = \frac{SV(X \rightarrow Y)}{SV(X)}$$

As can be understood from the formulas, the support value is the ratio of repeated records in the data to the whole data. The confidence value is known as the ratio of the support value of a rule to the support value of the predecessors. The fact that the established rules are strong requires high trust and support values. At the beginning of the procedures, the rules that will remain above the minimum support and minimum confidence values to be determined by the researcher should be taken into consideration, and other information should be eliminated.

Association rules can be considered as a two-step process:

1. Find all common product sets. The predetermined minimum support value defines the frequency of these determined product clusters.
2. Establish strong association rules from common product sets. These rules are defined as the rules that provide minimum support and trust value. [14].

Some algorithms used and developed for association rules are; AIS [13], SETM [15], Apriori [16], Partition [17], RARM (Rapid Association Rule Mining) [18], and CHARM [19]. Among these algorithms, the first one is AIS, and the best known is the Apriori algorithm.

### 3.1. Apriori Algorithm

The name of the Apriori algorithm is Apriori, meaning "prior" since it gets the information from the previous step [16]. This algorithm is essentially iterative and is used to discover sets of passing items. It is necessary to browse the database many times to find frequently passing sets of items. In the first scan, there are frequently passed item sets that provide the minimum support criteria with one element, and in the following scans, the frequent element sets found in the previous scan are used to produce new potential favorite item sets called candidate sets. The support values of the candidate

sets are calculated during scanning, and the sets that provide the minimum support criteria from the candidate sets are the frequent sets of items produced in that transition. Frequent sets of items become candidate sets for the next pass. This process continues until there is no new set of frequent [14]. According to the essence of the Apriori Algorithm, if the k-item set (item set with k elements) provides the minimum support criterion, the subset of this set also provides the minimum support criterion. That is, the support value of a set of items is not greater than the support value of the subset [14].

Association rules share many common features with classification. Both use rules to characterize regularities in a dataset. However, these two methods differ greatly in their goals. While classification focuses on prediction, association rules focus on providing information to the user. In particular, it focuses on detecting and characterizing unexpected relationships between data items [20].

### 3.2. Associative Classification

Associative classification is a data mining method that combines classification and association rules methods to make predictions. Particularly, an associative classification is an approach that uses rules obtained with association rules to create classification models. Associative classification is a special association rule mining with the target/response/dependent/class variable to the right of the rule obtained. In a rule such as  $X \rightarrow Y$ , Y must be the target / response / dependent / class variable. One of the primary advantages of employing a classification based on association rules according to classical classification approaches is that simple if-then rules represent the output of an associative classification algorithm. This rule makes it easier for the user to understand and interpret the result [8].

Associative classification and association rules are different methods. Relational classification takes into account only the class attribute in the relevant rules. However, association rules allow multiple attribute values in related rules. In other words, there is no class feature in association rules, an example of unsupervised learning, and a class is given in associative classification, an example of supervised learning. The purpose of the association rules is to discover the relationship between the items in the transaction database, while in the associative classification; the aim is to create a classifier that can predict the classes of test data objects. While association rules can have more than one attribute as a result of a rule, in associative classification, there is an only class attribute as a result of a rule. In association rules, over-fitting is not a problem, but in associative classification, over-fitting is a problem. Overfitting occurs when it performs well in the training data set and poorly in the test data set. Overfitting may be due to a variety of reasons, such as a small amount of training data object or noise [8]. The relational classification consists of three steps [21].

- 1) Determine the smallest support and confidence values,
- 2) Create rules and pruning,
- 3) Classification is made in the light of the meta-rules.

There are many algorithms for associative classification. Some of the methods are; CBAR (Classification based on



association rules), wCBAR (Weighted classification based on association rules), CARGBA (Classification based on association rule generated in a bidirectional approach), HMAc (Hierarchical Multi-label Associative Classification), GARC (Gain based association rule classification), and RCAR (Regularized class association rules).

### 3.3. Developed web-based software

To create this web-based application, the Shiny library was used to allow the design of interactive web-based applications based on the R programming language [22]. Also, in the development of the interface, shinythemes [23], shinyBS [24], shinyLP [25], shinyalert [26], shinyjs [27] were used. Boruta [28], arules [29], arulesCBAR [30], caret [31], visNetwork [32] packages were used to make the analysis. The main and submenus of the software are described below. The developed software includes three main sections: "Introduction", "Data" and "Analysis".

#### 3.3.1. Introduction

This section includes an information section with general information about the software and information about the packages used during the software development phase. With the "Start" tab on the page, the "Data Transactions" menu is passed.

#### 3.3.2. Data

There are three submenus under the "Data Transactions" main menu: "File upload", "Data viewing", "Variable Types". In the "File upload" menu, the file containing the data set is loaded. This developed software supports data files with ".xls", ".xlsx", ".sav" and ".arff" extensions. After uploading the file, the "Data viewing" sub-tab becomes active, and we have the opportunity to see the data set. With the "Variable Types" tab, we can determine the type and role of the variables in the data set. If the response/output variable is not determined while determining the variable roles, the error screen "Missing variable definitions" appears in the developed software.

#### 3.3.3. Analysis

The analysis will be made with the Response / Output variable and the predictive variable (s) with the "Analysis" tab. To carry out the analysis, it should be decided whether to select the variable with the section "Apply variable selection". Then, "Support" and "Confidence" values should be determined. If no selection is made, the analysis is made by accepting the "Support" value as 0.2 and the "Confidence" value as 0.8. If there are numerical variables in the loaded data set, the "Discretization method" tab will be displayed in this tab. The conversion of numerical variables into categorical variables is performed by selecting one of the discretization methods included in this tab. If no selection is made, numerical variables are converted into categorical variables by using the "Ameva" method. Finally, with the "Classification algorithm", one of the CBAR (Classification based on association rules) and RCAR (Regularized class association rules) methods included in the software is selected, and analyzes are done with the "Analysis" button. The results can be printed with the "Print page".

### 3.3.4. Developed web-based software accessibility

The developed interactive web-based software can be accessed free of charge at <http://biostatapps.inonu.edu.tr/ACS/>.

## 4. RESULTS

Open access dataset named "Breast Cancer Wisconsin (Diagnostic) Data Set" was used to analyze how the web-based software developed in this study works and evaluates its outputs. First, the data set was loaded into the software. After this process, the "Analysis" step is started with the "Next" button. The classification performance metrics of the model are given in Table 2.

TABLE II  
THE METRICS OF THE MODEL'S CLASSIFICATION PERFORMANCE

METRIC	MODELS	
	CBAR	RCAR
ACCURACY	0.954	0.951
BALANCED ACCURACY	0.951	0.951
SENSITIVITY	0.939	0.953
SPECIFICITY	0.964	0.95
POSITIVE PREDICTIVE VALUE	0.939	0.918
NEGATIVE PREDICTIVE VALUE	0.964	0.971
F1-SCORE	0.939	0.935

CBAR: Classification based on association rules; RCAR: Regularized class association rules

According to the findings of performance metrics, accuracy, balanced accuracy, sensitivity, specificity, positive predictive value, negative predictive value, and F1-score values obtained from the CBAR model are 0.954, 0.951, 0.939, 0.964, 0.939, 0.964, and 0.939, respectively.

Association rules using the classification algorithm are given in Table 3. When radius=[15,28.1) and texture=[19.5,39.3) are considered, the probability of a woman getting breast cancer is about 100%. Similarly, as texture=[19.5,39.3) and area=[696,2.5e+03) are taken into account, the probability of a female having breast cancer is nearly 100%, and when texture=[19.5,39.3) and perimeter=[98.8,188) are regarded, the probability of a woman with breast cancer is almost 100%. In contrast to the above rules, as texture=[9.71,19.5), area=[144,696) and compactness=[0.0194,0.102) are reckoned, the probability of a female not having breast cancer is 99.5%. The other rules generated from the CBAR model can be interpreted as the rules described earlier (Table 3).

**TABLE III**  
ASSOCIATION RULES USED TO CONSTRUCT THE BEST PERFORMING MODEL (CBAR)

Left-hand side rules	Right-hand side rules	Support	Conf.	Freq.
{radius=[15,28.1), texture=[19.5,39.3)}	{diagnosis=Malignant}	0.206	1	117
{texture=[19.5,39.3), area=[696,2.5e+03)}	{diagnosis=Malignant}	0.206	1	117
{texture=[19.5,39.3), perimeter=[98.8,188)}	{diagnosis=Malignant}	0.204	1	116
{texture=[9.71,19.5), area=[144,696),compactness=[0.0194,0.102)}	{diagnosis=Benign}	0.348	0.995	198
{area=[144,696), smoothness=[0.0526,0.0895),concavepoints=[0,0.0514)}	{diagnosis=Benign}	0.279	0.994	159
{area=[696,2.5e+03), concavity=[0.0933,0.427)}	{diagnosis=Malignant}	0.243	0.993	138
{perimeter=[43.8,98.8), smoothness=[0.0526,0.0895), fractal_dimension=[0.0553,0.0665)}	{diagnosis=Benign}	0.232	0.992	132
{texture=[9.71,19.5), area=[144,696), concavepoints=[0,0.0514)}	{diagnosis=Benign}	0.406	0.991	231
{texture=[9.71,19.5), area=[144,696),c oncavity=[0,0.0933)}	{diagnosis=Benign}	0.404	0.991	230
{texture=[9.71,19.5), perimeter=[43.8,98.8),compactness=[0.0194,0.102)}	{diagnosis=Benign}	0.351	0.99	200
{area=[144,696), smoothness=[0.0526,0.0895)}	{diagnosis=Benign}	0.285	0.988	162
{area=[144,696), concavity=[0,0.0933),concavepoints=[0,0.0514),symmetry=[0.172,0.304)}	{diagnosis=Benign}	0.246	0.986	140
{perimeter=[98.8,188), concavity=[0.0933,0.427)}	{diagnosis=Malignant}	0.244	0.986	139
{area=[696,2.5e+03), compactness=[0.102,0.345)}	{diagnosis=Malignant}	0.232	0.985	132
{texture=[19.5,39.3), smoothness=[0.0895,0.163),concavepoints=[0.0514,0.201)}	{diagnosis=Malignant}	0.223	0.984	127
{texture=[9.71,19.5), compactness=[0.0194,0.102),concavepoints=[0,0.0514),fractal_dimension=[0.0553,0.0665)}	{diagnosis=Benign}	0.281	0.982	160
{texture=[9.71,19.5), compactness=[0.0194,0.102),concavepoints=[0,0.0514)}	{diagnosis=Benign}	0.353	0.98	201
{area=[696,2.5e+03),concavepoints=[0.0514,0.201)}	{diagnosis=Malignant}	0.262	0.98	149
{texture=[9.71,19.5), area=[144,696),fractal_dimension=[0.0553,0.0665)}	{diagnosis=Benign}	0.327	0.979	186
{area=[144,696), compactness=[0.0194,0.102),concavity=[0,0.0933),concavepoints=[0,0.0514)}	{diagnosis=Benign}	0.466	0.978	265
{area=[144,696), concavity=[0,0.0933), concavepoints=[0,0.0514)}	{diagnosis=Benign}	0.534	0.977	304
{texture=[19.5,39.3), smoothness=[0.0895,0.163),compactness=[0.102,0.345),concavity=[0.0933,0.427)}	{diagnosis=Malignant}	0.209	0.975	119
{area=[144,696), compactness=[0.0194,0.102),concavepoints=[0,0.0514),symmetry=[0.106,0.172)}	{diagnosis=Benign}	0.278	0.975	158

## 5. DISCUSSION

The most common type of cancer among women is breast cancer (BC). Each year around the world, over half a million people die because of BC. BC is more prevalent in women and middle-aged people. If early diagnosis is not made promptly, cancer cells start spreading across the body. Operational intervention and intensive chemotherapy processes can become important for the treatment of patients with BC in the next step. Early diagnosis is so critical for those reasons. Advances in artificial intelligence technology predict that the efficiency of automated systems will be more dominant than the human factor in this field. In other words, the experts' decision-making processes should be converted through technical means. Nowadays, automated systems based on artificial intelligence models are commonly used to diagnose various diseases [33].

Studies of developing interpretable/explainable machine learning models and making black-box models interpretable/explainable have gained importance recently. In particular, the classification of traditional medical datasets with satisfactory accuracy and interpretation of model outputs may be the reason for these models to be preferred over classical statistical hypothesis tests that require many assumptions. CBAR and RCAR rule-based interpretable models used in this study are determined to create rules with a high ability to interpret with negligible classification performance losses compared to models (such as support vector machine, random forest, neural network-based model, etc.) employed in classifying breast cancer data in other studies.

In the current study, to predict breast cancer early, it is intended to develop a new user-friendly web-based software to realize the use of the associative classification method, which uses the association rules method. Association rules, one of the descriptive models of data mining, are methods that analyze the coexistence of events. Association rules use combinations such as statistical analysis, data mining, and database management to reveal existing hidden relationships. These relationships are based on the coexistence of data elements and express the co-occurrence of events together with certain possibilities [34].

Classification analysis is one of the basic methods of machine learning and is used by a large scientific community. Classification is an estimation process that assigns each observation in the dataset to the predetermined classes under certain rules [35]. Associative classification makes classification by combining two common data mining methods, association rules, and classification methods. In recent years, association rules methods have been successfully used to create correct classifiers in associative classification. The important advantage of using this method is that simple if-then rules represent its output by using a classification based on association rules according to classical classification approaches, and it facilitates to understand and interpret the rules [8].

In the current study, the proposed software based on the studied models generated promising predictions in classifying breast cancer for malignant and benign according to the metric values of the classification performance on the open-access

“Breast Cancer Wisconsin (Diagnostic) Data Set”. For this purpose, the main hypotheses of this study are to determine whether classification-based association rules models are successful in predicting breast cancer on the open-access dataset and evaluate the classification performance. According to the experimental results, the calculated accuracy metric was quite high (0.954), and the other metrics of balanced accuracy, sensitivity, specificity positive predictive value, negative predictive value, and F1-score were similarly so large (>0.930) from the proposed model and web-based software. As of 2020, several studies have been conducted to investigate the classification of breast cancer using machine learning and data mining techniques. A novel paper offers a comparative analysis by applying various machine learning algorithms such as Support Vector Machine, Naïve Bayes, Decision Tree, K-Nearest Neighbor, k-means clustering, and Artificial Neural Networks on Wisconsin Diagnostic Data Set to predict early breast cancer. The authors conclude, after analyzing all the implemented algorithms, that artificial neural network provides better prediction as 97.85% compared to all the other methods [36]. Another newly published work performed the experiments on the dataset Wisconsin Diagnostic Breast Cancer (WDBC), and the technique of k-fold cross-validation is used for model assessment. The proposed two-layer nested ensemble classifiers were compared with single classifiers (i.e., BayesNet and Naïve Bayes) in terms of classification precision, accuracy, recall, F1 score, the area under the ROC curve, computational durations of single and nested ensemble classifiers. The results show that the accuracy of SV-BayesNet-3-MetaClassifier and SV-Naïve Bayes-3-MetaClassifier was 98.07 percent, and the proposed two-layer nested ensemble models outperform the single classifiers and much of the preceding research [37]. Another up-to-date paper introduces a genetic algorithm for mutual information (MIGA), where MIGA is a combination of two algorithms: mutual information (MI) and genetic algorithm (GA) for detecting breast cancer using the Breast Cancer Wisconsin Diagnostic dataset. The results of the MIGA algorithm show that the highest accuracy (99 percent) with GA-based MI features was achieved [38]. The novel paper offers six classification algorithms of the medical diagnostic methods used in machine learning on the UCI three medical data sets, including the “Diagnostic Breast Cancer dataset for Wisconsin”. Overall, three medical datasets, the experimental results indicate that the SVM classification algorithm has achieved the most promising prediction [39]. In another recent study, the dataset for Wisconsin Diagnostic Breast Cancer (WDBC) is analyzed with Support Vector Machine (SVM), k-Nearest Neighbors (k-NN), Naïve Bayes (NB), Decision-Tree (DT) and Logistic Regression (LR) using 5-fold cross-validation method. Classification efficiency is calculated through the use of the confusion matrix through performance assessment parameters, i.e., precision, sensitivity, and specificity. The best result in that study by SVM is a 99.12 percent accuracy after the phase of normalization [40]. When the classification performances of the previous studies are outlined, the performance metrics values of the current study are so high (>0.930 for all the metrics evaluated) and similar to the reported other papers on the classification of breast

cancer. Besides, the present study develops free web-based software to classify the breast cancer, and defines the associated rules of any data sets (e.g., Breast Cancer Wisconsin (Diagnostic) Data Set) achieved from the associative classification methods. This research has important features compared to other studies in that it includes open access web-based software and association rules based on the classification of diseases (e.g., breast cancer).

As a result, in the analysis of the open-access dataset, the proposed model (CBAR) has a distinctive feature in classifying breast cancer based on the performance metrics. The associative classification software developed based on CBAR produces successful predictions in the classification of breast cancer. The hypothesis established within the scope of the purpose of this study has been confirmed as the similar estimates are achieved with the results of other papers in the classification of breast cancer.

#### REFERENCES

- [1] K. Oktay et al., "A Computational Statistics Approach to Evaluate Blood Biomarkers for Breast Cancer Risk Stratification," *Hormones and Cancer*, vol. 11, no. 1, pp. 17-33, 2020.
- [2] J. Ping et al., "Differences in gene-expression profiles in breast cancer between African and European-ancestry women," *Carcinogenesis*, 2020.
- [3] H. Akpınar, "Veri tabanlarında bilgi keşfi ve veri madenciliği," *Ü İşletme Fakültesi Dergisi*, vol. 29, no. 1, pp. 1-22, 2000.
- [4] A. Koyuncugil and N. Özgülbaş, "Veri madenciliği: Tıp ve sağlık hizmetlerinde kullanımı ve uygulamaları," *Bilişim Teknolojileri Dergisi*, vol. 2, no. 2, 2009.
- [5] L. T. Moss and S. Atre, *Business intelligence roadmap: the complete project lifecycle for decision-support applications*. Addison-Wesley Professional, 2003.
- [6] Y.-L. Chen, J.-M. Chen, and C.-W. Tung, "A data mining approach for retail knowledge discovery with consideration of the effect of shelf-space adjacency on sales," *Decision support systems*, vol. 42, no. 3, pp. 1503-1520, 2006.
- [7] A. K. Pujari, *Data mining techniques*. Universities press, 2001.
- [8] F. Thabtah, "A review of associative classification mining," *The Knowledge Engineering Review*, vol. 22, no. 1, pp. 37-65, 2007.
- [9] D. Dua and C. Graff, "UCI machine learning repository. School of Information and Computer Science, University of California, Irvine, CA," ed, 2019.
- [10] A. S. Kumar and R. Wahidabanu, "Data Mining Association Rules for Making Knowledgeable Decisions," in *Data Mining Applications for Empowering Knowledge Societies*: IGI Global, 2009, pp. 43-53.
- [11] U. M. Fayyad, G. Piatetsky-Shapiro, P. Smyth, and R. Uthurusamy, "Advances in knowledge discovery and data mining," 1996: American Association for Artificial Intelligence.
- [12] D. T. Larose and C. D. Larose, *Discovering knowledge in data: an introduction to data mining*. John Wiley & Sons, 2014.
- [13] R. Agrawal, T. Imieliński, and A. Swami, "Mining association rules between sets of items in large databases," in *Proceedings of the 1993 ACM SIGMOD international conference on Management of data*, 1993, pp. 207-216.
- [14] J. Han and M. Kamber, "Data Mining: Concepts and Techniques. ISBN 13: 978-1-55860-901-3," ed: Elsevier, USA, 2008.
- [15] M. Houtsma and A. Swami, "Set-oriented mining for association rules in relational databases," in *Proceedings of the eleventh international conference on data engineering*, 1995, pp. 25-33: IEEE.
- [16] R. Agrawal and R. Srikant, "Fast algorithms for mining association rules," in *Proc. 20th int. conf. very large data bases, VLDB*, 1994, vol. 1215, pp. 487-499.
- [17] A. Savasere, E. R. Omiecinski, and S. B. Navathe, "An efficient algorithm for mining association rules in large databases," *Georgia Institute of Technology* 1995.
- [18] A. Das, W.-K. Ng, and Y.-K. Woon, "Rapid association rule mining," in *Proceedings of the tenth international conference on Information and knowledge management*, 2001, pp. 474-481.
- [19] M. J. Zaki and C.-J. Hsiao, "CHARM: An efficient algorithm for closed itemset mining," in *Proceedings of the 2002 SIAM international conference on data mining*, 2002, pp. 457-473: SIAM.
- [20] N. Ye, *The handbook of data mining*. CRC Press, 2003.
- [21] M. Azmi, G. C. Runger, and A. Berrado, "Interpretable regularized class association rules algorithm for classification in a categorical data space," *Information Sciences*, vol. 483, pp. 313-331, 2019.
- [22] W. Chang, J. Cheng, J. Allaire, Y. Xie, and J. McPherson, "Shiny: web application framework for R," *R package version*, vol. 1, no. 5, 2017.
- [23] W. Chang, T. Park, L. Dzedzic, N. Willis, and M. McInerney, "shinythemes: Themes for Shiny," *R package version*, vol. 1, no. 1, p. 144, 2015.
- [24] E. Bailey, "shinyBS: Twitter bootstrap components for shiny," *R package version* 0.61, 2015.
- [25] J. Dumas, "shinyLP: Bootstrap Landing Home Pages for Shiny Applications," *R package version*, vol. 1, p. 2, 2019.
- [26] D. Attali and T. Edwards, "shinyalert: Easily Create Pretty Popup Messages (Modals) in Shiny," *R package version* 1.0, 2018.
- [27] D. Attali, "Shinyjs: Easily improve the user experience of your shiny apps in seconds," *R package version* 0.9, 2016.
- [28] M. B. Kursa and W. R. Rudnicki, "Feature selection with the Boruta package," *J Stat Softw*, vol. 36, no. 11, pp. 1-13, 2010.
- [29] M. Hahsler et al., "Package 'arules'," ed, 2019.
- [30] I. Johnson, "arulesCBA: Classification for Factor and Transactional Data Sets Using Association Rules."
- [31] M. Kuhn, "The caret package," *R Foundation for Statistical Computing, Vienna, Austria*. URL <https://cran.r-project.org/package=caret>, 2012.
- [32] B. Almende, B. Thieurmel, and T. Robert, "visNetwork: Network Visualization using 'vis.js' Library," ed: CRAN, 2016.
- [33] M. Toğaçar, B. Ergen, and Z. Cömert, "Application of breast cancer diagnosis based on a combination of convolutional neural networks, ridge regression and linear discriminant analysis using invasive breast cancer images processed with autoencoders," *Medical Hypotheses*, vol. 135, p. 109503, 2020/02/01/ 2020.
- [34] İ. Perçin, F. H. Yağın, E. Güldoğan, and S. Yoloğlu, "ARM: An Interactive Web Software for Association Rules Mining and an Application in Medicine," in *2019 International Artificial Intelligence and Data Processing Symposium (IDAP)*, 2019, pp. 1-5: IEEE.
- [35] İ. PERÇİN, F. H. YAĞIN, A. K. ARSLAN, and C. ÇOLAK, "An Interactive Web Tool for Classification Problems Based on Machine Learning Algorithms Using Java Programming Language: Data Classification Software," in *2019 3rd International Symposium on Multidisciplinary Studies and Innovative Technologies (ISMSIT)*, 2019, pp. 1-7: IEEE.
- [36] G. Rawal, R. Rawal, H. Shah, and K. Patel, "A Comparative Study Between Artificial Neural Networks and Conventional Classifiers for Predicting Diagnosis of Breast Cancer," in *ICDSMLA 2019*: Springer, 2020, pp. 261-271.
- [37] M. Abdar et al., "A new nested ensemble technique for automated diagnosis of breast cancer," vol. 132, pp. 123-131, 2020.
- [38] N. Vutakuri and A. U. J. I. J. o. A. I. P. Maheswari, "Breast cancer diagnosis using a Minkowski distance method based on mutual information and genetic algorithm," vol. 16, no. 3-4, pp. 414-433, 2020.
- [39] P. S. Nishant, S. Mehrotra, B. G. K. Mohan, and G. Devaraju, "Identifying Classification Technique for Medical Diagnosis," in *ICT Analysis and Applications*: Springer, 2020, pp. 95-104.
- [40] N. Panwar, D. Sharma, and N. J. A. a. S. Narang, "Breast Cancer Classification with Machine Learning Classifier Techniques," 2020.

#### BIOGRAPHIES

**Ahmet Kadir ARSLAN** received his BSc degree in Maths from Afyon Kocatepe University and MSc degree in Biostatistics and Medical Informatics from Inonu University. He is currently in his second year of his PhD in Biostatistics and Medical Informatics at Inonu University. His research interest are interpretable machine learning, decision support systems, neural networks, data preprocessing, image classification and dimension reduction.

**Zeynep TUNÇ** 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.

**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.

**Cemil ÇOLAK** obtained his BSc. degree in statistics from Ondokuz Mayıs University in 1999. He received MSc. degree in Biostatistics from the Inonu University in 2001, and Ph.D. degree in the Graduate Department of Biostatistics and Medical Informatics of Ankara University in 2005. His research interests are cognitive systems, data mining, reliability, and biomedical system, genetics, and bioengineering. In 2016, he joined the Department of Biostatistics and Medical Informatics at Inonu University as a Professor, where he is presently a professor. He is active in teaching and research in the general image processing, artificial intelligence, data mining, analysis.





## PUBLICATION ETHICS

All who participate in producing The Journal of Cognitive Systems conduct themselves as authors, reviewers, editors, and publishers in accord with the highest level of professional ethics and standards. Plagiarism or self-plagiarism constitutes unethical scientific behaviour, and is never acceptable. By submitting a manuscript to this journal, each author explicitly confirms that the manuscript meets the highest ethical standards for authors and co-authors. **The undersigned hereby assign(s) to The Journal of Cognitive Systems (JCS) copyright ownership in the above paper, effective if and when the paper is accepted for publication by JCS, and to the extent transferable under applicable national law. This assignment gives JCS the right to register copyright to the paper in its name as claimant, and to publish the paper via any print or electronic medium.**

Authors, or their employers, in the case of works made for hire, retain the following rights.

- + all proprietary rights other than copyright, including patent rights
- + the right to make and distribute copies of the Paper for internal purposes
- + the right to use the material for lecture or classroom purposes
- + the right to prepare derivative publications based on the Paper, including books or book chapters, journal papers, and magazine articles, provided that publication of a derivative work occurs subsequent to the official date of publication by JCS.
- + the right to post an author-prepared version or an official version (preferred version) of the published paper on an internal or external server controlled exclusively by the author/employer, provided that (a) such posting is non-commercial in nature, and the paper is made available to users without charge; (b) a copyright notice and full citation appear with the paper, and (c) a link to JCS's official online version of the abstract is provided using the Document Object Identifier (DOI) link



# THE JOURNAL OF COGNITIVE SYSTEMS

an international, peer-reviewed, indexed, and  
open-access periodical

VOLUME 05, NUMBER 01

J U N E 2 0 2 0

## CONTENTS

<b>A. Acet, and E. Akkaya</b> : A Deep Learning Image Classification Using Tensorflow for Optical Aviation Systems,.....	01-04
<b>E. Guldogan, Z. Tunc, A. Acet, and C. Colak</b> : Performance Evaluation of Different Artificial Neural Network Models in the Classification of Type 2 Diabets Mellitus,.....	05-09
<b>E. Guldogan, Z. Tunc, and C. Colak</b> : Classification of Breast Cancer and Determination of Related Factors with Deep Learning Approach,.....	10-14
<b>M. Kivrak, F. B. Akcesme, and C. Colak</b> : Sample Size Effect on Classification Performance of Machine Learning Models: An Application of Coronary Artery Disease,.....	15-18
<b>M. Kivrak, F. B. Akcesme, and C. Colak</b> : Evaluation of Association Rules Based on Certainty Factor: An Application on Diabetes Data Set,.....	19-22
<b>H. S. Nogay</b> : Prediction of Post-Treatment Survival Expectancy in Head & Neck Cancers by Machine Learning Methods,.....	23-32
<b>A.K. Arslan, Z. Tunc, I. Balikci Cicek, and C. Colak</b> : A Novel Interpretable Web-Based Tool on the Associative Classification Methods: An Application on Breast Cancer Dataset, .....	33-40

[www.dergipark.gov.tr/jcs](http://www.dergipark.gov.tr/jcs)



ISSN 2548-0650