Classification of Down Syndrome of Mice Protein Dataset on MongoDB Database

F. G. Furat and T. İbrikçi

Abstract—**There are samples both with Down Syndrome and without in mice protein expression data set. It is important to define the reason of Down Syndrome treatment by means of mice protein for the same treatment seem human being. In the present study, mice protein expression data set from UCI repository are classified using Bayesian Network algorithm, K- Nearest Neighbor, Decision Table, Random Forest and Support Vector Machine which are some of classification methods. The classification algorithms with 10-fold cross validation and by splitting equally as test and train data are tested to classify on the mice protein data set. The classification of the data set was succeeded with 94.3519% accuracy in 0.06 seconds using Bayesian Network, with 99.2593% accuracy in 0.01 seconds using KNN, with 95.4630 % accuracy in 1.2 seconds using Decision Table, with 100% accuracy in 0.58 seconds using Random Forest and with 100% accuracy in 1.17 seconds using SVM, with 10-fold cross validation. On the other hand, the classification of the data set was succeeded with 95.3704% accuracy in 0.22 seconds using Bayesian Network, with 98.3333% accuracy in 0 seconds using KNN, with 98.3333% accuracy in 0.72 seconds using Decision Table, with 100% accuracy in 0.77 seconds using Random Forest and with 100% accuracy in 1.48 seconds using SVM, by equally train-test data partition.**

Index *Terms*—**Bayesian Network, KNN, Decision Table, Random Forest, SVM, Classification, MongoDB, NoSQL.**

I. INTRODUCTION

IN RECENT YEARS, as data collections expand, the need to find meaningful data increases. Hence, as interest on information technology increases, the popularity of data processing fields such as data mining, big data, machine learning and artificial intelligence increases.

Classification that is one of popular information technology methods is a machine learning technique used to predict class labels. The classification consists of two steps, model construction and model usage. Model construction is that relationships are discovered with a training set. Model usage is that test set are used to evaluate success of model. Classification has many application areas such as medical diagnosis, credit approval, target marketing and fraud detection, etc. [1].

NoSQL databases have been used to analyze big data when relational databases were not sufficient to be stored and analyzed such amount of large data. NoSQL which is abbreviation of "Not Only SQL" overcomes the data without structured in contrast to conventional relational databases [2].

Although the common property of NoSQL databases is non-relational based structure, there are a number of different technologies such as MongoDB, Cassandra and Neo4j etc [3].

MongoDB database which is a document based NoSQL databases is used to store in order to store mice protein expression data in this study. The database in this study is preferred due to update of stored data being easy.

Bayesian Network that is one of them classification methods has been remarkably successful in many studies for classification such as [4-7] on WEKA. In [4], breast cancer data set is classified using Bayesian Network with 89.71% accuracy. Furthermore, when Bayesian Network classifier is compared to other methods such as Radial Basis Function, Single Conj. Rule Learner, Decision Tree and Pruning and Nearest Neighbors in terms of correct classification, Bayesian Network has been classified with less error [4]. Basic classification such as Bayesian Networks, decision tree and k-nearest neighbor and clustering algorithms such as k-means, partional clustering, hierarchical clustering and fuzzy clustering are compared using Iris data set on WEKA tool [5]. Decision tree, Bayesian Network, Random Forest, k-nearest neighbor and Bagging algorithms are compared using email header fields for test spam classification. Emails are correctly classified with 97.87% accuracy using Bayesian Network [6]. Various classification algorithms are compared for intrusion detection on WEKA tool. KDDCUP99 data set is classified with 90.62% accuracy

F.G. FURAT, is Ph.D. Student at Department of Electrical and Electronics Engineering, Cukurova University, Adana, Turkey,

T. İBRİKÇİ, is with Department of Electrical and Electronics Engineering Cukurova University, Adana, Turkey, (e-mail: ibrikci@cu.edu.tr)

Manuscript received August 10, 2017; accepted Nov 16, 2017.
DOI: 10.17694/bajece.419553

using Bayesian Network [7].

In the present study, mice protein expression data set from UCI repository are used to classify on WEKA tool. In [8], mice protein expression data set together with 7 datasets –totally 8 data sets- from UCI are used to cluster using three different clustering algorithms as Harm-ELM, US-ELM and K-Harmonic Mean. Mice protein expression data set are clustered with 82.97% accuracy, 77.51% accuracy and 77.51% accuracy using Harm-ELM, US-ELM and K-Harmonic Mean, respectively. Four data sets including mice protein expression data set from UCI repository are used to analyze elephant search algorithm (ESA) that is a new improved algorithm in [9] and compare performance of the ESA with k-means, GMM-EM and DBSCAN algorithms [9].

In [10], 1000 samples of medical data set have classified to guess future disease of patients using SVM with 82.542% accuracy in 0.0642 seconds and KNN with 79.225 accuracy in 0. 261 seconds. SVM and KNN are commonly used in areas such as education, industry and medicine where information extraction is necessary [10, 11]. When classifications with data at different size are performed, it appears that KNN algorithm is more successful for data of small size [10]. KNN and genetic algorithm are used together to handle complexity of large data for heart disease diagnosis in [11]. The KNN with genetic algorithm increases 5% accuracy of diagnosis.

Many classification methods such as SVM and Random Forest are applied to discover future health disease risks. Healthcare Cost and Utilization Project (HCUP) dataset is trained using Random Forest, SVM, bagging and boosting to predict disease risk. Random Forest algorithm yields better results than other algorithms depending on the ROC curve [12].

In section 2, mice protein data set used to classify, Bayesian Network, K- Nearest Neighbor, Decision Table, Random Forest and Support Vector Machine which are some of classification methods, WEKA and MongoDB are introduced. The experimental results of the study such as classification accuracy, time taken and confusion matrix are given in section 3. Information results are concluded and future work is provided in section 4.

II. METHOD

A. Mice Protein Expression Data Set

Mice Protein Expression data set is obtained from UCI Repository [13]. The data set is a collection of 1080 protein measurements where type of 570 measurements of them are control mice (without Down Syndrome), and type of the rest 510 measurements are trisomic mice (down syndrome). Both control mice measurements and Down Syndrome measurements are divided into 4 classes. Hence, eight classes are obtained from measurements of protein as shown on the Table 1. The data set contains 82 features of each sample. The combination of these features is used to find the type that each sample belongs to [13, 14].

TABLE I.
CLASSES OF MICE PROTEIN EXPRESSION DATA SET

Classes	Features	Samples per class
c-CS-s	control mice, motivated to learn, infused with saline	150
c-CS-m	control mice, motivated to learn, infused with memantine	150
c-SC-s	control mice, not motivated to learn, infused with saline	135
c-SC-m	control mice, not motivated to learn, infused with memantine	135
t-CS-s	trisomy mice, motivated to learn, infused with saline	135
t-CS-m	trisomy mice, motivated to learn, infused with memantine	135
t-SC-s	trisomy mice, not motivated to learn, infused with saline	105
t-SC-m	trisomy mice, not motivated to learn, infused with memantine	135

B. Bayesian Network

Bayesian Network is also named as Bayesian network and belief network which is a probabilistic graphical model. Bayes Network comprises of a directed acyclic graphs (DAG) in which nodes represent random variables and conditional probability tables (CPT) in which distribution for each node given its parents: P(xi|parents(xi)) based on DAG [15-17].

The probability of class given the particular sample of x1…..xn features is computed to use Bayes rule. The class with the highest posterior probability based on Bayes rule is assigned as class of the sample. Bayesian Network aims to predict correct class of given sample. There is a sample a Bayesian Network in Figure 1 [16].

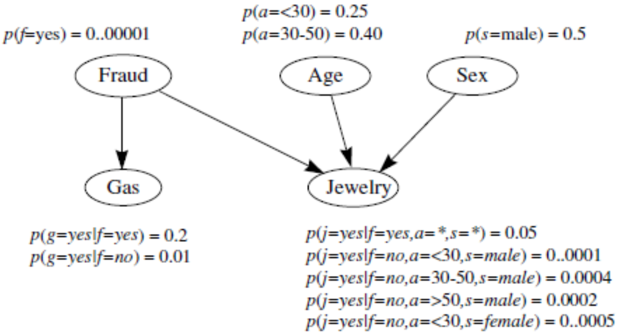
p(a=<30) = 0.25
p(a=30-50) = 0.40
p(f=yes) = 0..00001
p(s=male) = 0.5
Fraud
Age
Sex
Gas
Jewelry
p(g=yes|f=yes) = 0.2
p(g=yes|f=no) = 0.01
p(j=yes|f=yes,a=*,s=*) = 0.05
p(j=yes|f=no,a=<30,s=male) = 0..0001
p(j=yes|f=no,a=30-50,s=male) = 0.0004
p(j=yes|f=no,a=>50,s=male) = 0.0002
p(j=yes|f=no,a=<30,s=female) = 0..0005

Figure 1. Motion Scenario

C. K- Nearest Neighbor a (KNN)

KNN algorithm is a simple supervised learning classification algorithm which used in many areas such as medical data analysis, statistical estimation and pattern recognition [10,11]. KNN algorithm is called as different tags like lazy learning and instance based learning etc [11].

KNN algorithm is roughly classified into two types based on Nearest Neighbor (NN) techniques. One of them is structure-based NN and the other is structure-less NN. Structure-based NN handles memory limitation and on the other hand structure-less NN decreases the computational complexity [18].

The KNN algorithm is based on the assumption that the new sample will include the class that has the closest properties to it. The KNN algorithm proceeds with the following steps [19]:

a. The distance between the new sample and all the samples in the training set is calculated using distance functions such as Euclidean and Manhattan.
b. The closest k samples to the new sample are selected from the training set.
c. The new sample is assigned the highest class among the nearest k neighbors.

D. Decision Table

Rule based classification algorithm is an iterative process that is known as separate-and-conquer method. The Rule based classification algorithm creates a rule which covers a training examples' subset, firstly. After that, all samples covered by the rule are moved out of training set. This procedure is repeated until there is no sample moved out of the training set [20].

The rule based algorithms are OneR, Decision Table, DTNB and Ridor algorithm. Decision Table is of them that builds simple decision table that includes the same number of features as the real dataset. After that, a new data sample is assigned a class by discovering the line in the decision table that matches the out of class values of the data sample. [20]

E. Random Forest (RF)

Random Forest is an ensemble method that builds many decision trees. Each tree in RF will cast a vote for some input. After that, the decision trees are used to classify a new sample with majority vote [12].

RF use many trees to overcome high dimensionality of data. Some notable features of RF algorithm are following:

a. There is an effective way to guess missing data in RF.
b. There is a method for balancing faults in unbalanced data is the weighted random forest (WRF) in RF.
c. The significance of the variables processed in classification is predicted in RF.

F. Support Vector Machine (SVM)

Firstly, Vapnik designed Support Vector Machine (SVM) as an efficient statistical learning algorithm to be used in classification in 1998 [21].

SVM is a supervised learning algorithm to use for classification and regression [22,23].

SVM has two types named as Linear SVM and Non-Linear SVM classification using to classify binary and multiclass problems respectively.

SVM represents that samples as points in space. SVM uses decision planes to classify the points. A decision plane is a plane to separate points having different class.

SVM finds an optimal hyperplane to classify new samples.

G. Waikato Environment for Knowledge Analysis (WEKA) Tool

The University of Waikato in New Zealand develops WEKA which is a data mining tool written in java language. The tool performs data preprocessing, classification, regression, clustering and association rules, also visualization [24-26].

H. MongoDB

It is an efficient non-relational database with high performance. It is under development with the following features [3].

- MongoDB is a document based database which is independent schema.
- MongoDB is easy scalable with rich queries and fast in-place updates. Hence data insertion, deletion and update processes can be performed effortless.
- Documents are stored in BSON format that is binary-encoded format of JSON documents on MongoDB.
- MongoDB makes features such as auto shading, consistency fault tolerance, persistence, aggregation, indexing, replication and high availability.

III. EXPERIMENTAL RESULTS

In the present study, mice protein expression data set are preprocessed from UCI data repository. Then, the data set are stored with 82 features and 1080 samples in MongoDB database. Due to the easy update feature of MongoDB, it is preferred to store data in this study.

Five classification algorithms as Bayesian Network, KNN, Decision Table, Random Forest and SVM are chosen to classify into 8 classes. Two different processes for preferring test and train data is applied. One of them is 10-fold cross validation and the other is that data set is split. in half as the test and the remaining half as train data. The five classification algorithms are used to classify the data set.

Classification performance results of the data set using five different algorithms with 10-fold cross validation and 50–50% train-test data partition is given in the Table II and Table III.

When these algorithms are evaluated according to classification accuracy, Random Forest and SVM have left the other three algorithms for this data set with 100% accuracy while Bayesian Network shows the lowest accuracy among the other chosen four algorithms. If algorithms are compared according to the time of building the classification, KNN is the algorithm that performs the operation in the shortest time compared to the other selected algorithms.

Since the Kappa values are between 0.9 and 1 for all algorithms, it is seen that the operations performed are very reliable results.

When examining the effect on the results of selecting 50-50% train-test data partition and 10 fold cross validation, it is unacceptable that one of them is more successful than the other. Because, selecting 50-50% train-test data partition gives more successful for Bayesian Network, while hand selecting 10 fold cross validation gives more successful for KNN.

TABLE II.
CLASSIFICATION RESULTS OF MICE PROTEIN EXPRESSION DATA SET WITH 10 FOLD CROSS VALIDATION

Evaluation Methods	Bayesian Network	KNN	Decision Table	Random Forest	SVM
The classification accuracy (%)	94.3519	99.2593	95.463	100	100
Time Taken to build (seconds)	0.06	0.01	1.2	0.58	1.17
Kappa Value	0.9354	0.9915	0.948	1	1
Mean Absolute Error	0.0149	0.0036	0.0792	0.0909	0.1875
Root Mean Squared Error	0.1093	0.0429	0.1455	0.1458	0.2912
Relative Absolute Error (%)	6.8214	1.6581	36.2553	41.6138	85.8255
Root Relative Squared (%)	33.0619	12.9928	44.0169	44.1280	88.1166

TABLE III.
CLASSIFICATION RESULTS OF MICE PROTEIN EXPRESSION DATA SET WITH 50–50% TRAIN-TEST DATA PARTITION

Evaluation Methods	Bayesian Network	KNN	Decision Table	Random Forest	SVM
The classification accuracy (%)	95.3704	98.3333	98.3333	100	100
Time Taken to build (seconds)	0.22	0	0.72	0.77	1.48
Kappa Value	0.947	0.9809	0.9809	1	1
Mean Absolute Error	0.0134	0.0073	0.0391	0.1053	0.1875
Root Mean Squared Error	0.0994	0.0643	0.0792	0.1676	0.2912
Relative Absolute Error (%)	6.1103	3.3393	17.867	48.1819	85.7782
Root Relative Squared (%)	30.0569	19.4266	23.9572	50.6836	88.0446

The following confusion matrixes to detect Down Syndrome treatment are produced using the classification algorithms by 10 fold cross validation and 50–50% train-test data partition. TP and FP rates are given in Table IV that have been obtained from the following confusion matrixes.

1) Confusion Matrix for classification using Bayesian Network with 10 fold cross validation

```
  a    b    c    d    e    f    g    h   <-- classified as
131    0   13    0    6    0    0    0 |  a = c-CS-m
  0  139    0    0    0   11    0    0 |  b = c-SC-m
  6    0  129    0    0    0    0    0 |  c = c-CS-s
  0    1    0  132    0    2    0    0 |  d = c-SC-s
  8    0    1    0  125    0    1    0 |  e = t-CS-m
  0    6    0    1    0  128    0    0 |  f = t-SC-m
  0    0    2    0    1    0  100    2 |  g = t-CS-s
  0    0    0    0    0    0    0  135 |  h = t-SC-s
```

2) Confusion Matrix for classification using Bayesian Network with 50–50% train-test data partition

```
 a  b  c  d  e  f  g  h   <-- classified as
67  0  3  0  0  0  0  0 | a = c-CS-m
 0 74  0  0  0  3  0  0 | b = c-SC-m
 6  0 66  0  0  0  0  0 | c = c-CS-s
 0  0  0 67  0  1  0  0 | d = c-SC-s
 7  0  0  0 55  0  0  0 | e = t-CS-m
 0  1  0  1  0 63  0  0 | f = t-SC-m
 1  0  0  0  1  0 48  1 | g = t-CS-s
 0  0  0  0  0  0  0 75 | h = t-SC-s
```

3) Confusion Matrix for classification using Random Forest with 10 fold cross validation

```
  a    b    c    d    e    f    g    h   <-- classified as
 150   0    0    0    0    0    0    0 |  a = c-CS-m
  0  150    0    0    0    0    0    0 |  b = c-SC-m
  0    0  135    0    0    0    0    0 |  c = c-CS-s
  0    0    0  135    0    0    0    0 |  d = c-SC-s
  0    0    0    0  135    0    0    0 |  e = t-CS-m
  0    0    0    0    0  135    0    0 |  f = t-SC-m
  0    0    0    0    0    0  105    0 |  g = t-CS-s
  0    0    0    0    0    0    0  135 |  h = t-SC-s
```

4) Confusion Matrix for classification using Random Forest with 50–50% train-test data partition

```
  a  b  c  d  e  f  g  h  <-- classified as
 70  0  0  0  0  0  0  0 | a = c-CS-m
  0 77  0  0  0  0  0  0 | b = c-SC-m
  0  0 72  0  0  0  0  0 | c = c-CS-s
  0  0  0 68  0  0  0  0 | d = c-SC-s
  0  0  0  0 62 0  0  0 | e = t-CS-m
  0  0  0  0  0 65  0  0 | f = t-SC-m
  0  0  0  0  0  0 51  0 | g = t-CS-s
 0   0  0  0  0  0   0 75 | h = t-SC-s
```

5) Confusion Matrix for classification using SVM with 10 fold cross validation

```
  a    b    c    d    e    f    g    h   <-- classified as
 150   0    0    0    0    0    0    0 |  a = c-CS-m
  0  150    0    0    0    0    0    0 |  b = c-SC-m
  0    0  135    0    0    0    0    0 |  c = c-CS-s
  0    0    0  135    0    0    0    0 |  d = c-SC-s
  0    0    0    0  135    0    0    0 |  e = t-CS-m
  0    0    0    0    0  135    0    0 |  f = t-SC-m
  0    0    0    0    0    0  105    0 |  g = t-CS-s
  0    0    0    0    0    0    0  135 |  h = t-SC-s
```

6) Confusion Matrix for classification using SVM with 50–50% train-test data partition

```
  a  b  c  d  e  f  g  h  <-- classified as
 70  0  0  0  0  0  0  0 | a = c-CS-m
  0 77  0  0  0  0  0  0 | b = c-SC-m
  0  0 72  0  0  0  0  0 | c = c-CS-s
  0  0  0 68  0  0  0  0 | d = c-SC-s
  0  0  0  0 62  0  0  0 | e = t-CS-m
  0  0  0  0  0 65  0  0 | f = t-SC-m
  0  0  0  0  0  0 51  0 | g = t-CS-s
  0  0  0  0  0  0  0 75 | h = t-SC-s
```

TABLE IV.
TP AND FP RATES OBTAINED FROM CONFUSION MATRIXES

c-CS-m class	Bayesian Network		Random Forest		SVM	
	10 fold cross validation	50–50% train-test data partition	10 fold cross validation	50–50% train-test data partition	10 fold cross validation	50–50% train-test data partition
TP rate	0.87333	0.95714	1	1	1	1
FP rate	0.09655	0.17283	0	0	0	0

IV. CONCLUSION AND FUTURE WORK

In the literature, mice protein expression data set has not been classified using five different classification algorithms.

In this study, the mice protein expression data set stored on MongoDB database is classified with 94.3519% accuracy, with 99.2593% accuracy, with 95.4630% accuracy, with 100% accuracy and with 100% accuracy, using Bayesian Network, KNN, Decision Table, Random Forest and SVM on WEKA tool by 10 fold cross validation, respectively.

On the other hand, in this study, the mice protein expression data set stored on MongoDB database is classified with 95.3704% accuracy, with 98.3333% accuracy, with 98.3333% accuracy, with 100% accuracy and with 100% accuracy, using Bayesian Network, KNN, Decision Table, Random Forest and SVM on WEKA tool by equally train-test data partition, respectively.

In the future works, classification algorithms for mice protein expression data set will be proposed with feature selection methods.

REFERENCES

[1] Han, Jiawei, Jian Pei, and Micheline Kamber. *Data mining: concepts and techniques*. Elsevier, 2011.

[2] Győrödi, C., Győrödi, R., Pecherle, G., & Olah, A. (2015). *A comparative study: MongoDB vs. MySQL. In Engineering of Modern Electric Systems (EMES)* 2015 13th International Conference on (pp. 1-6). IEEE.

[3] Nayak, A., Poriya, A., & Poojary, D. (2013). *Type of NOSQL databases and its comparison with relational databases*. International Journal of Applied Information Systems, 5(4), 16-19.

[4] Othman, Mohd Fauzi, and Thomas Moh Shan Yau. *Comparison of different classification techniques using WEKA for breast cancer.* 3rd Kuala Lumpur International Conference on Biomedical Engineering. Springer, 2007.

[5] Kumar, Ajay, and Indranath Chatterjee. *Data Mining: An experimental approach with WEKA on UCI Dataset*. International Journal of Computer Applications 138.13 (2016).

[6] Kulkarni, Priti, and Haridas Acharya. *Comparative analysis of classifiers for header based emails classification using supervised learning*. International Research Journal of Engineering and Technology, 03 (03), 1939- 1944 (2016).

[7] Modi, Ms Urvashi, and Anurag Jain. *A survey of IDS classification using KDD CUP 99 dataset in WEKA.* (2016).

[8] Sarunyoo Boriratrit, Sirapat Chiewchanwattana, Khamron Sunat, Pakarat Musikawan and Punyaphol Horata. *Harmonic extreme learning machine for data clustering*. Computer Science and Software Engineering (JCSSE), 13th International Joint Conference on. IEEE, 2016.

[9] Zhonghuan Tian, Raymond Wong, Richard Millham. *Elephant search algorithm on data clustering*. Natural Computation, Fuzzy Systems and Knowledge Discovery (ICNC-FSKD), 12th International Conference on. IEEE, 2016.

[10] Raikwal, J. S., and Kanak Saxena. "Performance evaluation of SVM and k-nearest neighbor algorithm over medical data set." *International Journal of Computer Applications* 50.14 (2012).

[11] Deekshatulu, B. L., and Priti Chandra. "Classification of heart disease using k-nearest neighbor and genetic algorithm." *Procedia Technology* 10 (2013): 85-94.

[12] Khalilia, Mohammed, Sounak Chakraborty, and Mihail Popescu. "Predicting disease risks from highly imbalanced data using random forest." *BMC medical informatics and decision making* 11.1 (2011): 51.

[13] Blake, C. & Merz, C. (1998). *UCI repository of machine learning databases*. University of California, Irvine, Dept. of Inf. and Computer Science.

[14] Higuera C, Gardiner KJ, Cios KJ. (2015) *Self-Organizing Feature Maps Identify Proteins Critical to Learning in a Mouse Model of Down Syndrome*. PLoS ONE 10(6): e0129126.

[15] Heckerman, David. *A tutorial on learning with Bayesian networks*. Innovations in Bayesian networks. Springer, 33-82, 2008.

[16] Buntine, W. (1991). *Theory refinement on Bayesian networks*. In B. D. D'Ambrosio, P. Smets, & P.P. Bonissone (Eds.), Proceedings of the Seventh Annual Conference on Uncertainty Artificial Intelligent pp. 52-60. San Francisco, CA.

[17] Daniel Grossman and Pedro Domingos (2004). *Learning Bayesian Network Classifiers by Maximizing Conditional Likelihood*. In Press of Proceedings of the 21st International Conference on Machine Learning, Banff, Canada.

[18] Bhatia, Nitin. "Survey of nearest neighbor techniques." *arXiv preprint arXiv:1007.0085* (2010).

[19] T.M. Mitchell, Machine Learning, The McGraw-Hill Companies Press, 1997.

[20] Mahajan, Aditi, and Anita Ganpati. "Performance evaluation of rule based classification algorithms." *International Journal of Advanced Research in Computer Engineering & Technology (IJARCET) Vol* 3 (2014): 3546-3550.

[21] Vapnik, V. (1998). Statistical Learning Theory. New York: Wiley.

[22] Kumari, V. Anuja, and R. Chitra. "Classification of diabetes disease using support vector machine." *International Journal of Engineering Research and Applications* 3.2 (2013): 1797-1801.

[23] Cortes, C., Vapnik, V., "Support-vector networks", Machine Learning, 20(2), pp. 273-297, 1995.

Vapnik, V. (1998). Statistical Learning Theory. New York: Wiley.

[24] WEKA at http://www.cs.waikato.ac.nz/~ml/weka. (last accessed:15 September 2018)

[25] Mark Hall, Eibe Frank, Geoffrey Holmes, Bernhard Pfahringer Peter Reutemann, Ian H. Witten. *The WEKA data mining software: an update.* ACM SIGKDD explorations newsletter 11.1 (2009): 10-18.

[26] Eibe Frank, Mark A. Hall, and Ian H. Witten (2016). *The WEKA Workbench. Online Appendix for "Data Mining: Practical Machine Learning Tools and Techniques"*, Morgan Kaufmann.

Fahriye Gemci Furat was born in Kahramanmaraş, Turkey in 1986. She was graduated from department of Computer Engineering in 2010. She received the MSc. degree in Electronics and Computer Engineering in 2015. She is currently pursuing a doctoral degree at Çukurova University. Her research interests are artificial intelligence, data mining, bioinformatics and social media.

Turgay Ibrikci received his BS degree in physics (Cukurova University, Adana, Turkey), MSc in computer science (Nova Southeastern University, Fort Laudardale, Florida, USA), and PhD in Electrical and Electronics Engineering Department (Cukurova University). Currently, he is an associate professor at Electrical-Electronics Engineering Department, Cukurova University. He had international experiences as a visiting researcher at Computational Neuro Engineering Lab (CNEL), University of Florida (1999), at the Neurosignal Analysis Lab (NAL), University of Texas, Health Science Center (2001 and 2004) and at the Institute of Bioinformatics, University of Georgia (2011). His research interests include machine learning, bioinformatics, biomedical data, protein structures, and medical image processing.