



# Balkan Journal of Electrical & Computer Engineering

An International Peer Reviewed, Referred, Indexed and Open Access Journal

[www.bajece.com](http://www.bajece.com)

Vol : 6  
No : 2  
Year : 2018  
ISSN : 2147 - 284X



This journal is supported by the Istanbul Technical University. It is abstracted and indexed in, Index Google Scholarship, the PSCR, Cross ref, DOAJ, Research Bible, Indian Open Access Journals (OAJ), Institutional Repositories (IR), J-Gate (Informatics India), Ulrich's, International Society of Universal Research in Sciences, DRJI, EyeSource, Cosmos Impact Factor, Cite Factor, SIS SIS Scientific Indexing Service, IJIF, iiiFactor. ULAKBİM-TR Dizin.



**General Publication Director & Editor-in-Chief**  
Ş.Serhat Seker, Istanbul Technical University, Turkey

#### Scientific Committee

Abhishek Shukla (India)  
Abraham Lomi (Indonesia)  
Aleksandar Georgiev (Bulgaria)  
Arunas Lipnickas (Lithuania)  
Audrius Senulis (Lithuania)  
Belle R. Upadhyaya (USA)  
Brijender Kahanwal (India)  
Chandar Kumar Chanda (India)  
Daniela Dzhonova-Atanasova (Bulgaria)  
Deris Stiawan (Indonesia)  
Emel Onal (Turkey)  
Emine Ayaz (Turkey)  
Enver Hatimi (Kosovo)  
Ferhat Sahin (USA)  
Gursel Alici (Australia)  
Hakan Temeltaş (Turkey)  
Ibrahim Akduman (Turkey)  
Jan Izykowski (Poland)  
Javier Bilbao Landatxe (Spain)  
Jelena Dikun (Lithuania)  
Karol Kyslan (Slovakia)  
Kunihiko Nabeshima (Japan)  
Lambros Ekonomou (Greece)  
Lazhar Rahmani (Algerie)  
Marcel Istrate (Romania)  
Marija Eidukeviciute (Lithuania)  
Milena Lazarova (Bulgaria)  
Muhammad Hadi (Australia)  
Muhamed Turkanović (Slovenia)  
Mourad Houabes (Algerie)  
Murari Mohan Saha (Sweden)  
Nick Papanikolaou (Greece)  
Okyay Kaynak (Turkey)  
Osman Nuri Ucan (Turkey)  
Ozgur E. Mustecaplioglu (Turkey)  
Padmanaban Sanjeevikumar (India)  
Ramazan Caglar (Turkey)  
Rumen Popov (Bulgaria)  
Tarek Bouktir (Algeria)  
Sead Berberovic (Croatia)  
Seta Bogosyan (USA)  
Savvas G. Vassiliadis (Greece)  
Suwarno (Indonesia)  
Tulay Adali (USA)  
Yogeshwarsing Calleecharan (Mauritius)  
YangQuan Chen (USA)  
Youcef Soufi (Algeria)

#### Aim & Scope

The journal publishes original papers in the extensive field of Electrical-Electronics and Computer engineering. It accepts contributions which are fundamental for the development of electrical engineering, computer engineering and its applications, including overlaps to physics. Manuscripts on both theoretical and experimental work are welcome. Review articles and letters to the editors are also included.

Application areas include (but are not limited to): Electrical & Electronics Engineering, Computer Engineering, Software Engineering, Biomedical Engineering, Electrical Power Engineering, Control Engineering, Signal and Image Processing, Communications & Networking, Sensors, Actuators, Remote Sensing, Consumer Electronics, Fiber-Optics, Radar and Sonar Systems, Artificial Intelligence and its applications, Expert Systems, Medical Imaging, Biomedical Analysis and its applications, Computer Vision, Pattern Recognition, Robotics, Industrial Automation.



ISSN: 2147- 284X

Vol: 6

No : 2

Year: April 2018

#### CONTENTS

- A. Onan;** Sentiment Analysis on Twitter Based on Ensemble of Psychological and Linguistic Feature Sets, .....**69-77**
- Y. Kaya;** Classification of PVC Beat in ECG Using Basic Temporal Features, .....**78-82**
- B. Karakaya, T. Kaya, A. Gulten;** FPGA-based ANN Design for Detecting Epileptic Seizure in EEG Signal, .....**83-87**
- M. Kara, M. Furat,** Client-Server Based Authentication Against MITM Attack via Fast Communication for IIoT Devices, .....**88-93**
- H. Karayığit, Ç. Aci, A. Akdağlı;** A Review of Turkish Sentiment Analysis and Opinion Mining, .....**94-98**
- S. G. Eraldemir, M. T. Arslan, E. Yildirim;** Investigation Of Feature Selection Algorithms On A Cognitive Task Classification: A Comparison Study, .....**99-104**
- T. Tulgar, A. Haydar, İ. Erşan;** A Distributed K Nearest Neighbor Classifier for Big Data, .....**105-111**
- F. G.Furat, T. Ibriki;** Classification of Down Syndrome of Mice Protein Dataset on MongoDB Database, .....**112-117**
- F. Karaomerlioglu;** Analysis of Photonic Crystal Tuned by Nematic Liquid Crystals, .....**118-121**
- C. Bakir, M. Yuzkat;** Speech Emotion Classification and Recognition with different methods for Turkish Language, .....**122-128**
- J. Dikun, L. Urmoniene, D. Stanelyte;** Spectral Ratio Method for Fault Detection in Rotating Machines, .....**129-131**
- V. Garousi, A. Tarhan;** Investigating the Impact of Team Formation by Introversion/Extraversion in Software Projects, .....**132-140**
- M. Yılmaz;** Real Measure of a Transmission Line Data with Load Forecast Model for The Future, .....**141-145**

#### BALKAN JOURNAL OF ELECTRICAL & COMPUTER ENGINEERING

(An International Peer Reviewed, Indexed and Open Access Journal)

#### Contact

Istanbul Technical University  
Department of Electrical Engineering  
Ayazaga Campus, Maslak, Istanbul-Turkey

**Web:** <https://www.bajece.com>  
<http://dergipark.gov.tr/bajece>  
**e-mail:** [editor@bajece.com](mailto:editor@bajece.com)

# Sentiment Analysis on Twitter Based on Ensemble of Psychological and Linguistic Feature Sets

A. Onan

**Abstract**—With the advances in information and communication technologies, social media and microblogging platforms serve as an important source of information. In microblogging platforms, people can share their opinions, complaints, sentiments and attitudes towards topics, current issues and products. Sentiment analysis is an important research direction in natural language processing, which aims to identify the sentiment orientation of source materials. Twitter is a popular microblogging platform, where people all over the world can interact by user-generated text messages. Information obtained from Twitter can serve as an essential source for several applications, including event detection, news recommendation and crisis management. In sentiment classification, the identification of an appropriate feature subset plays an important role. LIWC (Linguistic Inquiry and Word Count) is an exploratory text analysis software to extract psycholinguistic features from text documents. In this paper, we present a psycholinguistic approach to sentiment analysis on Twitter. In this scheme, we utilized five main LIWC categories (namely, linguistic processes, psychological processes, personal concerns, spoken categories and punctuation) as feature sets. In the experimental analysis, five LIWC categories and their ensemble combinations are taken into consideration. To explore the predictive performance of different feature engineering schemes, four supervised learning algorithms (namely, Naïve Bayes, support vector machines, k-nearest neighbor algorithm and logistic regression) and three ensemble learning methods (namely, AdaBoost, Bagging and Random Subspace) are utilized. The experimental results indicate that ensemble feature sets yield higher predictive performance compared to the individual feature sets.

**Index Terms**— Machine learning, psychological feature sets, sentiment analysis, Twitter.

## I. INTRODUCTION

THE IMMENSE QUANTITY OF INFORMATION available with the remarkable growth of social media and microblogging platforms can serve as an essential source for decision making about products, services and policies [1-2].

Twitter is a popular and fast growing microblogging platform, where people can send short messages (referred as tweets) within a character limit of 140.

A. ONAN, is with Department of Software Engineering, Celal Bayar University, Manisa, Turkey, (e-mail: [aytug.onan@cbu.edu.tr](mailto:aytug.onan@cbu.edu.tr)). Manuscript received August 12, 2017; accepted Nov 16, 2017. DOI: [10.17694/bajece.419538](https://doi.org/10.17694/bajece.419538)

Twitter enables users to communicate in an efficient way. The user generated content on Twitter provide a useful source of information for researchers and practitioners [3]. Information obtained from Twitter can serve as an essential source of information for several applications, including event detection, epidemic dispersion, news recommendation and crisis management [4-6]. Sentiment analysis (also known as opinion mining) is an important research direction in natural language processing, which aims to identify the sentiment orientation of source materials. Sentiment analysis can be utilized for obtaining information regarding new products and services. It can be further applied to identify positive and negative aspects of a particular product or service [7].

The methods of sentiment analysis can be mainly divided into two groups as lexicon-based approaches and machine-learning based approaches. In addition, sentiment analysis can be conducted at different granularities based on the levels of details. Based on the levels of details, sentiment analysis methods are grouped into three categories as: document-level, sentence-level and aspect-level sentiment analysis [8].

Sentiment analysis can be modelled as a text classification problem. In machine learning based sentiment analysis, supervised classification algorithms (such as Naïve Bayes algorithm, support vector machines, k-nearest neighbor algorithm and logistic regression) can be utilized to identify sentiment orientation. Machine learning based sentiment analysis schemes involve data preprocessing, feature extraction and selection and training supervised classification algorithms with labelled data set.

In order to obtain a classification scheme with high predictive performance, feature extraction is an essential task [9]. LIWC (Linguistic Inquiry and Word Count) is an exploratory text analysis software to extract psycholinguistic features from text documents. Features related to psychological, linguistic, social and cultural aspects can be important for sentiment analysis [10]. For this purpose, we present a psycholinguistic approach to sentiment analysis on Twitter. In this paper, we utilized five main LIWC categories (namely, linguistic processes, psychological processes, personal concerns, spoken categories and punctuation) as feature sets. In the experimental analysis, five LIWC categories and their ensemble combinations are taken into consideration. To explore the predictive performance of

different feature engineering schemes, Naïve Bayes, support vector machines, k-nearest neighbor algorithm and logistic regression are utilized. In addition, ensemble methods (namely, Bagging, AdaBoost and Random Forest algorithms) are also considered to examine the predictive performance of supervised learning algorithms in conjunction with ensemble methods for sentiment analysis.

The rest of the paper is organized as follows: In Section 2, related work on sentiment analysis is presented. Section 3 presents the methodology of the study and Section 4 presents the experimental procedure and empirical results. Section 5 describes the concluding remarks.

## II. RELATED WORK

Sentiment analysis on Twitter data has attracted research attention. This section briefly reviews the existing works on sentiment analysis on Twitter data. Sentiment analysis on Twitter poses several challenges, due to the short length of messages and unstructured, informal and irregular nature of content. Hence, identification of an appropriate feature set is an important research direction. For instance, Go et al. [11] examined the usage of unigrams, bigrams, unigrams and bigrams and part of speech tags as features. In the classification phase, Naïve Bayes, maximum entropy and support vector machine classifiers were utilized. The empirical analysis indicated augmenting unigrams and bigrams yields better predictive performance compared to the other feature engineering schemes. Part of speech tags were not useful features and the highest classification performance was achieved by maximum entropy learner. In another study, Barbosa and Feng [12] explored the predictive performance of n-grams and tweet syntax features (such as retweets, hashtags, replies, links, punctuation, emoticons and upper cases). The empirical analysis on support vector machines indicated that tweet syntax features enhance the predictive performance of sentiment classification schemes on Twitter and n-grams cannot completely reveal the text messages. Similarly, Pak and Paroubek [13] examined the usage of n-grams and part of speech tags as features. In the empirical analysis, multinomial Naïve Bayes, support vector machines and conditional random field classifiers were utilized. The empirical analysis indicated that the utilization of part of speech tags in conjunction to n-grams yields better predictive performance on sentiment analysis of Twitter data. In another study, Koulumpis et al. [14] explored the usage of n-gram features, lexicon features, part of speech tags and microblogging features (such as the presence of positive, negative and neutral emoticons and the presence of intensifiers) on sentiment analysis of Twitter data. The experimental analysis indicated that the highest predictive performance among different feature engineering schemes was obtained by n-gram features in conjunction to lexicon features and microblogging features. In addition, the results indicated that integrating parts of speech features dropped the predictive performance of sentiment classification. Similarly, Agarwal et al. [15] examined the usage of part of speech features, lexicon features and microblogging features for sentiment analysis of Twitter data. In addition, they introduced a tree based

representation to augment different feature engineering schemes in an efficient way. The experimental analysis indicated that the usage of prior polarity of words in conjunction with their part of speech tags yields the highest classification accuracy.

In another study, Saif et al. [16] examined the usage of unigram features, part of speech features and sentiment topic features for sentiment analysis on Twitter. The experimental analysis indicated that semantic feature set based approach yield better predictive performance compared to other feature engineering schemes.

Onan [1] examined the predictive performance of different n-gram models (namely, unigram, bigram and trigram) and their combinations on sentiment analysis of Turkish Twitter messages. In the empirical analysis, the highest predictive performance is achieved by the combination of unigram and bigram features. In another study, Salas-Zarate et al. [10] examined the performance of psycholinguistic feature sets in sentiment analysis of product reviews. In this study, linguistic processes, psychological processes, personal concerns, spoken categories and punctuation are taken into consideration. While existing work on sentiment analysis of Twitter data concentrates on n-grams, part of speech tags and microblogging based features, this study aims to examine the predictive performance of psychological and linguistic features obtained by LIWC on sentiment analysis on Twitter.

## III. METHODOLOGY

This section describes dataset collection process, data processing, feature engineering schemes to represent the dataset, classification algorithms and ensemble learning methods utilized in the experimental analysis.

### A. Dataset Collection

To evaluate the predictive performance of psychological and linguistic features on sentiment analysis, we have carried out an analysis on English messages on Twitter that contain positive, negative and neutral sentiments. In the dataset collection, we adopted the framework presented in [17]. We utilized Twitter4J, an open-source Java library for utilizing Twitter Streaming API, to collect tweets. Each tweet is labelled by a single class label, either as positive, negative or neutral. After collecting the tweets, automatic filtering was applied to remove irrelevant and redundant tweets (retweets and duplicates). In this way, we obtained a collection of 6218 negative, 4891 positive and 4252 neutral tweets. In order to obtain a balanced corpus, our final dataset contains a collection of 4200 negative, 4200 positive and 4200 neutral tweets.

### B. Data Preprocessing

Due to irregular and informal nature of Twitter messages, it is essential to preprocess the tweets so that particular problems (such as initialisms, unnecessary repetitions and misuse of letters) can be eliminated [18]. In the preprocessing stage, we adopted the framework presented in [19]. The preprocessing stage mainly seeks to remove unnecessary characters or

sequences, which have no value to the sentiment classification. For this purpose, the following tasks were performed on each tweet [19]:

- Remove mentions and replies to other users' tweets, which are represented by strings starting with "@".
- Remove URLs (namely, strings starting with "http://").
- Remove "#" character.

### C. Feature Engineering

In this section, we examine different psycholinguistic feature sets on sentiment analysis. In this scheme, we utilized LIWC (Linguistic Inquiry and Word Count) to extract psycholinguistic features from the dataset. LIWC categories have been successfully utilized in several fields of computational linguistics, including sarcasm identification and satire detection [20, 21].

LIWC is a text analysis application to identify emotional, cognitive and structural aspects of verbal and written speech samples. The first version of LIWC application was developed in 1993 and the most recent version was released on 2015 [22]. LIWC contains dictionaries on several languages, including English, Spanish, German, Dutch, Norwegian, Italian and Portuguese.

LIWC Dictionary contains approximately 6400 words, word stems and emoticons. Each entry of the dictionary contains one or several word categories or sub-dictionaries. For a particular word encountered in the text, scores for the corresponding categories or dictionaries are incremented. The categories can be further classified into five main sets as linguistic processes, psychological processes, personal concerns, spoken categories and punctuation. In Table 1, main LIWC sets and categories are listed.

TABLE I  
MAIN LIWC SETS AND CATEGORIES

Feature Set	Categories
Linguistic Processes	Word count, total pronouns, personal pronouns, articles, prepositions, auxiliary verbs, adverbs, conjunctions
Psychological Processes	Affective processes, positive emotion, negative emotion, social processes, cognitive processes, perceptual processes
Personal Concerns	Work, leisure, home, money
Spoken Categories	Assent, Non-fluencies, fillers
Punctuation	Total punctuation, periods, commas, colons, semicolons, question marks, exclamation marks, dashes

As it can be observed from the categories listed in Table 1, linguistic processes contains grammatical information, such as word count, total number of pronouns, personal pronouns, articles, prepositions and auxiliary verbs. Psychological processes involves psychological information, such as affective processes, positive emotion, and negative emotion and so on. Personal concerns contains information, such as work, leisure, home and money. Spoken categories involves information regarding the spoken language. Finally, punctuation set involves punctuation marks, such as punctuation, periods, commas, colons, semicolons, question

marks, exclamation marks, dashes.

Based on the aforementioned five main LIWC feature sets, target words or word stems are searched through the LIWC dictionary. Each word is assigned to one or more sub-dictionaries.

### D. Classification Algorithms

To evaluate the predictive performance of different feature engineering schemes, Naïve Bayes, support vector machines, k-nearest neighbor algorithm and logistic regression algorithm are utilized.

Naïve Bayes algorithm (NB) is a probabilistic classification algorithm based on Bayes' theorem. It has a simple structure due to the assumption of conditional independence. Though its simple structure, it can be effectively utilized in text and web mining applications [23].

Support vector machines (SVM) are supervised learning algorithms that can be utilized to solve classification and regression problems. They can be applied effectively to classify both linear and non-linear data [24]. Support vector machines build a hyperplane in a higher dimensional space to solve classification or regression problem. The hyperplane aims to make a good separation by achieving the largest distance to the nearest training data points of classes (known as functional margin).

K-nearest neighbor algorithm (KNN) is an instance-based classifier. In KNN algorithm, the class label of each instance is determined based on the k-nearest neighbors of the instance. Based on the predictions of the neighbor instances, a majority voting scheme is utilized to determine the class label.

Logistic regression (LR) is a linear classification algorithm, which uses a linear function of a set of predictor variables to model the probability of some event's occurring [25]. Linear regression can yield good results. However, the membership values generated by linear regression cannot be always in [0-1] range, which is not an appropriate range for probabilities. In logistic regression, a linear model is constructed on the transformed target variable whilst eliminating the mentioned problems.

### E. Ensemble Learning Methods

This section briefly describes the ensemble learning algorithms utilized in the empirical analysis.

Bagging (Bootstrap aggregating) is a popular ensemble learning method, which aims to obtain a single prediction with higher predictive performance by combining weak learning algorithms trained on different training sets [26]. In this scheme, different training sets are obtained by simple random sampling with replacement. The predictions of weak learning algorithms are combined by majority voting or weighted voting.

AdaBoost algorithm is another popular ensemble learning method, which aims to obtain a robust classification scheme by focusing on the data points that are difficult to classify [27]. In this scheme, the weight values assigned to the instances of the training set are adjusted so that the weight values of misclassified instances are increased, whereas the

weight values of correctly classified instances are decreased. In this way, the learning algorithms focus on classifying the difficult instances.

Random subspace algorithm is an ensemble learning algorithm which combines multiple classifiers trained on the randomly selected feature subspaces [28]. The algorithm aims to avoid over-fitting, while providing high predictive performance by training the weak learning algorithms on different samples of the feature space.

#### IV. EXPERIMENTAL ANALYSIS AND RESULTS

This section presents the evaluation measures, experimental procedure and the experimental results of the study.

##### A. Evaluation Measures

In order to evaluate the performance of classification algorithms, two different evaluation measures, namely, classification accuracy and F-measure.

Classification accuracy (ACC) is the proportion of true positives and true negatives obtained by the classification algorithm over the total number of instances as given by Equation 1 [29]:

$$ACC = \frac{TN + TP}{TP + FP + FN + TN} \quad (1)$$

where TN denotes number of true negatives, TP denotes number of true positives, FP denotes number of false positives and FN denotes number of false negatives.

Precision (PRE) is the proportion of the true positives against the true positives and false positives as given by Equation 2:

$$PRE = \frac{TP}{TP + FP} \quad (2)$$

Recall (REC) is the proportion of the true positives against the true positives and false negatives as given by Equation 3:

$$REC = \frac{TP}{TP + FN} \quad (3)$$

F-measure takes values between 0 and 1. It is the harmonic mean of precision and recall as determined by Equation 4:

$$F - measure = \frac{2 * PRE * REC}{PRE + REC} \quad (4)$$

##### B. Experimental Procedure

In the experimental analysis, 10-fold cross validation method is employed. In this scheme, the original dataset is randomly divided into ten mutually exclusive folds. Training and testing process is repeated ten times and each part is tested and trained ten times and the average results for 10-fold are reported. The experimental analysis is performed with the machine learning toolkit WEKA (Waikato Environment for Knowledge Analysis) version 3.9, which is an open-source platform that contains many machine learning algorithms implemented in JAVA.

##### C. Experimental Results

In Tables 2-3, classification accuracies and F-measure results obtained by psycholinguistic feature sets and the four base learning algorithms are presented, respectively.

TABLE II  
CLASSIFICATION RESULTS OBTAINED BY SUPERVISED LEARNING METHODS ON PSYCHOLINGUISTIC FEATURE SETS

Feature set	NB	SVM	KNN	LR
LP	77.35	76.73	72.30	72.80
PP	77.28	76.57	72.17	72.06
PC	76.40	75.77	71.97	71.76
SC	74.57	75.66	70.88	71.55
PU	74.25	75.60	70.56	71.54
LP+PP	79.41	78.06	73.86	76.52
LP+PC	79.35	78.06	73.85	76.43
LP+SC	79.25	78.04	73.84	76.32
LP+PU	79.17	77.95	73.80	76.06
PP+PC	79.14	77.92	73.51	76.05
PP+SC	77.98	77.33	72.86	74.62
PP+PU	77.71	77.10	72.82	74.59
PC+SC	77.63	76.98	72.78	74.55
PC+PU	77.50	76.91	72.78	73.95
SC+PU	77.36	76.82	72.50	72.99
LP+PP+PC	<b>86.94</b>	<b>82.50</b>	<b>79.39</b>	<b>81.09</b>
LP+PP+SC	80.49	79.12	75.70	78.04
LP+PP+PU	80.44	79.10	75.53	78.01
LP+PC+SC	80.29	79.05	75.45	77.71
LP+PC+PU	80.24	78.94	75.28	77.69
LP+SC+PU	80.04	78.81	75.09	77.56
PP+PC+SC	79.90	78.61	74.99	77.52
PP+PC+PU	79.81	78.55	74.78	77.46
PP+SC+PU	78.75	77.82	73.27	75.92
PC+SC+PU	78.69	77.60	73.25	75.85
LP+PP+PC+SC	85.24	82.02	78.33	80.81
LP+PP+PC+PU	84.56	81.96	77.98	80.70
LP+PP+SC+PU	83.88	81.75	77.59	80.61
LP+PC+SC+PU	83.81	81.55	77.27	80.09
PP+PC+SC+PU	83.38	81.34	77.24	80.01
LP+PP+PC+SC+PU	83.20	81.08	77.14	79.27

NB: Naïve Bayes algorithm, SVM: support vector machines, KNN: K-nearest neighbor algorithm, LR: logistic regression, LP: linguistic processes, PP: psychological processes, PC: personal concerns, SC: Spoken categories, PU: punctuation.

In the first column, the different dimensions of LIWC used for training a particular classifier are reported. For instance, LP+PP indicates that linguistic processes and psychological processes are taken into account in the empirical analysis.

TABLE III  
F-MEASURE VALUES OBTAINED BY SUPERVISED LEARNING  
METHODS ON PSYCHOLINGUISTIC FEATURE SETS

Feature set	NB	SVM	KNN	LR
LP	0.78	0.77	0.73	0.73
PP	0.78	0.77	0.72	0.72
PC	0.77	0.76	0.72	0.72
SC	0.75	0.76	0.71	0.72
PU	0.75	0.76	0.71	0.72
LP+PP	0.80	0.78	0.74	0.77
LP+PC	0.80	0.78	0.74	0.77
LP+SC	0.80	0.78	0.74	0.77
LP+PU	0.79	0.78	0.74	0.76
PP+PC	0.79	0.78	0.74	0.76
PP+SC	0.78	0.78	0.73	0.75
PP+PU	0.78	0.77	0.73	0.75
PC+SC	0.78	0.77	0.73	0.75
PC+PU	0.78	0.77	0.73	0.74
SC+PU	0.78	0.77	0.73	0.73
LP+PP+PC	<b>0.87</b>	<b>0.83</b>	<b>0.80</b>	<b>0.81</b>
LP+PP+SC	0.81	0.79	0.76	0.78
LP+PP+PU	0.81	0.79	0.76	0.78
LP+PC+SC	0.81	0.79	0.76	0.78
LP+PC+PU	0.81	0.79	0.76	0.78
LP+SC+PU	0.80	0.79	0.75	0.78
PP+PC+SC	0.80	0.79	0.75	0.78
PP+PC+PU	0.80	0.79	0.75	0.78
PP+SC+PU	0.79	0.78	0.74	0.76
PC+SC+PU	0.79	0.78	0.74	0.76
LP+PP+PC+SC	0.86	0.82	0.79	0.81
LP+PP+PC+PU	0.85	0.82	0.78	0.81
LP+PP+SC+PU	0.84	0.82	0.78	0.81
LP+PC+SC+PU	0.84	0.82	0.78	0.80
PP+PC+SC+PU	0.84	0.82	0.78	0.80
LP+PP+PC+SC+PU	0.84	0.81	0.77	0.80

NB: Naïve Bayes algorithm, SVM: support vector machines, KNN: K-nearest neighbor algorithm, LR: logistic regression, LP: linguistic processes, PP: psychological processes, PC: personal concerns, SC: Spoken categories, PU: punctuation.

Considering the experimental results (in terms of classification accuracies and F-measure values) presented in Tables 2-3, different classification algorithms have different predictive performance. The highest predictive performance among the compared classifiers is achieved by Naïve Bayes algorithm and the second highest predictive performance is obtained by support vector machines. Logistic regression classifier and K-nearest neighbour algorithm generally yield

similar predictive performance.

The study seeks to examine the predictive performance of different LIWC categories (namely, linguistic processes, psychological processes, personal concerns, spoken categories and punctuation) and their subsets as feature sets.

Regarding the predictive performance of individual feature sets, the highest predictive performance is achieved by using linguistic processes (denoted as LP). The second highest predictive performance is obtained by psychological processes, the third highest predictive performance is obtained by personal concerns and the lowest predictive performance is obtained by punctuation. Hence, linguistic processes, psychological processes and personal concerns provide clues to better analysis sentiment on Twitter. The highest predictive performance (77.35%) by the individual feature sets is achieved by linguistic processes and Naïve Bayes algorithm.

As it can be observed from the results listed in Tables 2-3, ensemble feature sets (combining different LIWC dimensions) yield higher predictive performance compared to the individual feature sets. The highest predictive performance among the ensemble feature sets is obtained by combining linguistic processes, psychological processes and personal concerns. The highest predictive performance achieved by this configuration is 86.94%, it is utilized in conjunction to Naïve Bayes classifier.

Ensemble learning methods can be utilized to further enhance the predictive performance of supervised learning algorithms. In the empirical analysis, we have also considered ensemble of supervised learning algorithms in conjunction with psycholinguistic feature sets. In this regard, twelve ensemble schemes (AdaBoost, Bagging and Random Subspace ensembles of four supervised learning algorithms) are considered.

In Table 4, classification accuracies obtained by ensembles of feature sets and classifiers are presented. As it can be observed from the results listed in Table 4, the predictive performances of supervised learning algorithms are generally improved by using ensemble learning methods. Regarding the predictive performance of different ensemble learning methods, Random Subspace method generally yield better results than other ensemble learning methods. The highest predictive performance on the ensemble learning methods is achieved by the ensemble feature sets combining linguistic processes, psychological processes and personal concerns. For this configuration, Random Subspace ensemble of Naïve Bayes ensemble is utilized as the classifier. This configuration achieves a classification accuracy of 89.10%.

In Table 5, F-measure values obtained by ensembles of feature sets and classifiers are presented. The predictive performance patterns obtained in terms of classification accuracies are also valid for F-measure values listed in Table 5.

To summarize the main findings of the study, Figure 1 and Figure 2 depict the main effect plot for classification accuracy and the main effect plot for F-measure values, respectively.

TABLE IV  
CLASSIFICATION RESULTS OBTAINED BY ENSEMBLE LEARNING METHODS ON PSYCHOLINGUISTIC FEATURE SETS

Ensemble Method	Bagging	Random Subspace	Ada-Boost	Bagging	Random Subspace	Ada-Boost	Bagging	Random Subspace	Ada-Boost	Bagging	Random Subspace	Ada-Boost
Base Learner	NB	NB	NB	SVM	SVM	SVM	KNN	KNN	KNN	LR	LR	LR
LP	78.10	79.51	78.20	77.46	77.63	77.49	73.11	73.13	73.08	73.64	73.71	73.64
PP	78.04	79.42	78.16	77.25	77.48	77.35	72.99	73.00	72.98	72.91	73.01	72.93
PC	77.14	78.54	77.29	76.47	76.69	76.58	72.82	72.78	72.76	72.59	72.67	72.67
SC	75.34	76.74	75.48	76.40	76.53	76.49	71.71	71.71	71.70	72.40	72.47	72.44
PU	75.00	76.40	75.18	76.26	76.47	76.39	71.39	71.39	71.37	72.38	72.43	72.42
LP+PP	80.16	81.59	80.31	78.79	78.97	78.87	74.64	74.71	74.65	77.41	77.41	77.40
LP+PC	80.11	81.53	80.25	78.76	78.97	78.85	74.64	74.66	74.65	77.26	77.32	77.30
LP+SC	80.02	81.44	80.15	78.75	78.96	78.84	74.64	74.67	74.61	77.19	77.22	77.19
LP+PU	79.92	81.35	80.15	78.65	78.88	78.75	74.65	74.68	74.60	76.91	77.00	76.94
PP+PC	79.91	81.31	80.01	78.60	78.81	78.75	74.32	74.34	74.29	76.89	76.98	76.95
PP+SC	78.76	80.17	78.88	78.05	78.25	78.12	73.68	73.70	73.67	75.49	75.51	75.48
PP+PU	78.46	79.91	78.60	77.77	78.00	77.90	73.61	73.66	73.65	75.42	75.52	75.47
PC+SC	78.34	79.81	78.54	77.71	77.87	77.76	73.59	73.64	73.59	75.40	75.46	75.41
PC+PU	78.27	79.67	78.44	77.64	77.79	77.71	73.63	73.64	73.55	74.83	74.87	74.80
SC+PU	78.10	79.52	78.22	77.56	77.75	77.59	73.35	73.33	73.28	73.82	73.90	73.88
LP+PP+PC	<b>87.68</b>	<b>89.10</b>	<b>87.82</b>	<b>83.19</b>	<b>83.38</b>	<b>83.33</b>	<b>80.19</b>	<b>80.22</b>	<b>80.17</b>	<b>81.92</b>	<b>82.03</b>	<b>81.94</b>
LP+PP+SC	81.25	82.68	81.40	79.80	79.98	79.94	76.53	76.50	76.52	78.86	78.92	78.91
LP+PP+PU	81.18	82.61	81.37	79.78	79.95	79.90	76.33	76.34	76.33	78.86	78.93	78.89
LP+PC+SC	81.03	82.42	81.24	79.69	79.95	79.85	76.28	76.29	76.26	78.57	78.62	78.59
LP+PC+PU	80.99	82.45	81.13	79.65	79.82	79.68	76.11	76.13	76.07	78.54	78.57	78.55
LP+SC+PU	80.82	82.25	80.96	79.57	79.72	79.56	75.93	75.90	75.88	78.40	78.47	78.43
PP+PC+SC	80.65	82.07	80.80	79.34	79.49	79.41	75.83	75.81	75.83	78.36	78.47	78.40
PP+PC+PU	80.56	81.98	80.67	79.26	79.48	79.37	75.58	75.62	75.59	78.27	78.34	78.35
PP+SC+PU	79.51	80.97	79.65	78.56	78.69	78.60	74.10	74.11	74.05	76.77	76.81	76.76
PC+SC+PU	79.43	80.90	79.62	78.32	78.50	78.39	74.08	74.08	74.03	76.70	76.77	76.74
LP+PP+PC+SC	85.99	87.41	86.12	82.72	82.95	82.79	79.14	79.15	79.13	81.65	81.73	81.71
LP+PP+PC+PU	85.32	86.74	85.46	82.61	82.87	82.79	78.78	78.81	78.83	81.54	81.63	81.53
LP+PP+SC+PU	84.61	86.02	84.77	82.48	82.67	82.58	78.44	78.39	78.44	81.46	81.53	81.44
LP+PC+SC+PU	84.55	85.96	84.73	82.22	82.44	82.34	78.09	78.06	78.06	80.97	81.03	80.96
PP+PC+SC+PU	84.11	85.56	84.32	82.02	82.23	82.15	78.06	78.06	78.01	80.90	80.93	80.86
LP+PP+PC+SC+PU	83.94	85.39	84.06	81.80	82.02	81.89	77.96	77.99	77.99	80.13	80.20	80.15



TABLE V  
F-MEASURE VALUES OBTAINED BY ENSEMBLE LEARNING METHODS ON PSYCHOLINGUISTIC FEATURE SETS

Ensemble Method	Bagging	Random Subspace	Ada-Boost	Bagging	Random Subspace	Ada-Boost	Bagging	Random Subspace	Ada-Boost	Bagging	Random Subspace	Ada-Boost
Base Learner	NB	NB	NB	SVM	SVM	SVM	KNN	KNN	KNN	LR	LR	LR
LP	0.80	0.81	0.80	0.79	0.79	0.79	0.75	0.75	0.75	0.75	0.75	0.75
PP	0.80	0.81	0.80	0.79	0.79	0.79	0.74	0.74	0.74	0.74	0.74	0.74
PC	0.79	0.80	0.79	0.78	0.78	0.78	0.74	0.74	0.74	0.74	0.74	0.74
SC	0.77	0.78	0.77	0.78	0.78	0.78	0.73	0.73	0.73	0.74	0.74	0.74
PU	0.77	0.78	0.77	0.78	0.78	0.78	0.73	0.73	0.73	0.74	0.74	0.74
LP+PP	0.82	0.83	0.82	0.80	0.81	0.80	0.76	0.76	0.76	0.79	0.79	0.79
LP+PC	0.82	0.83	0.82	0.80	0.81	0.80	0.76	0.76	0.76	0.79	0.79	0.79
LP+SC	0.82	0.83	0.82	0.80	0.81	0.80	0.76	0.76	0.76	0.79	0.79	0.79
LP+PU	0.82	0.83	0.82	0.80	0.80	0.80	0.76	0.76	0.76	0.78	0.79	0.79
PP+PC	0.82	0.83	0.82	0.80	0.80	0.80	0.76	0.76	0.76	0.78	0.79	0.79
PP+SC	0.80	0.82	0.80	0.80	0.80	0.80	0.75	0.75	0.75	0.77	0.77	0.77
PP+PU	0.80	0.82	0.80	0.79	0.80	0.79	0.75	0.75	0.75	0.77	0.77	0.77
PC+SC	0.80	0.81	0.80	0.79	0.79	0.79	0.75	0.75	0.75	0.77	0.77	0.77
PC+PU	0.80	0.81	0.80	0.79	0.79	0.79	0.75	0.75	0.75	0.76	0.76	0.76
SC+PU	0.80	0.81	0.80	0.79	0.79	0.79	0.75	0.75	0.75	0.75	0.75	0.75
LP+PP+PC	<b>0.89</b>	<b>0.91</b>	<b>0.90</b>	<b>0.85</b>	<b>0.85</b>	<b>0.85</b>	<b>0.82</b>	<b>0.82</b>	<b>0.82</b>	<b>0.84</b>	<b>0.84</b>	<b>0.84</b>
LP+PP+SC	0.83	0.84	0.83	0.81	0.82	0.82	0.78	0.78	0.78	0.80	0.81	0.81
LP+PP+PU	0.83	0.84	0.83	0.81	0.82	0.82	0.78	0.78	0.78	0.80	0.81	0.81
LP+PC+SC	0.83	0.84	0.83	0.81	0.82	0.81	0.78	0.78	0.78	0.80	0.80	0.80
LP+PC+PU	0.83	0.84	0.83	0.81	0.81	0.81	0.78	0.78	0.78	0.80	0.80	0.80
LP+SC+PU	0.82	0.84	0.83	0.81	0.81	0.81	0.77	0.77	0.77	0.80	0.80	0.80
PP+PC+SC	0.82	0.84	0.82	0.81	0.81	0.81	0.77	0.77	0.77	0.80	0.80	0.80
PP+PC+PU	0.82	0.84	0.82	0.81	0.81	0.81	0.77	0.77	0.77	0.80	0.80	0.80
PP+SC+PU	0.81	0.83	0.81	0.80	0.80	0.80	0.76	0.76	0.76	0.78	0.78	0.78
PC+SC+PU	0.81	0.83	0.81	0.80	0.80	0.80	0.76	0.76	0.76	0.78	0.78	0.78
LP+PP+PC+SC	0.88	0.89	0.88	0.84	0.85	0.84	0.81	0.81	0.81	0.83	0.83	0.83
LP+PP+PC+PU	0.87	0.89	0.87	0.84	0.85	0.84	0.80	0.80	0.80	0.83	0.83	0.83
LP+PP+SC+PU	0.86	0.88	0.87	0.84	0.84	0.84	0.80	0.80	0.80	0.83	0.83	0.83
LP+PC+SC+PU	0.86	0.88	0.86	0.84	0.84	0.84	0.80	0.80	0.80	0.83	0.83	0.83
PP+PC+SC+PU	0.86	0.87	0.86	0.84	0.84	0.84	0.80	0.80	0.80	0.83	0.83	0.83
LP+PP+PC+SC+PU	0.86	0.87	0.86	0.83	0.84	0.84	0.80	0.80	0.80	0.82	0.82	0.82

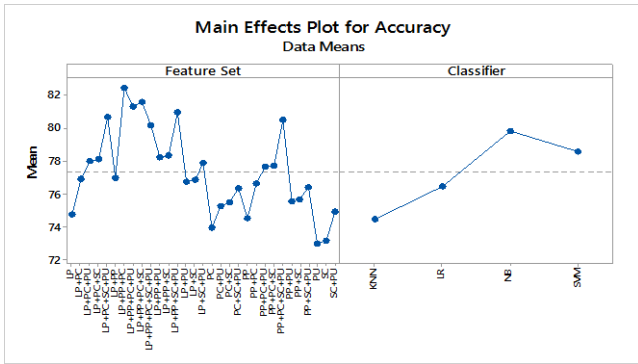


Fig.1. The main effects plot for accuracy

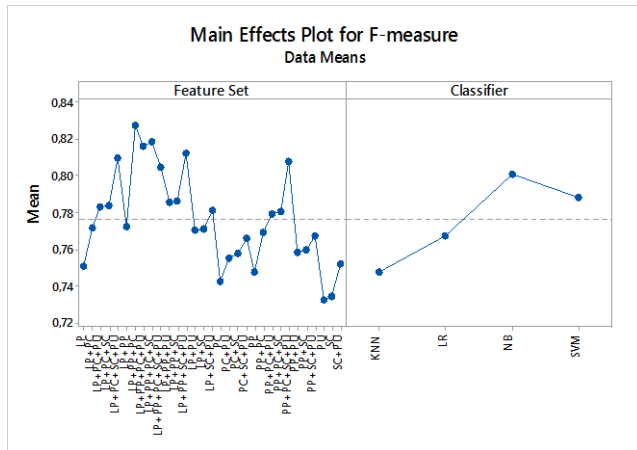


Fig.2. The main effects plot for F-measure

V. CONCLUSION

Social media and microblogging platforms serve as an essential source of information. Sentiment analysis on Twitter is a promising research direction. Sentiment analysis on Twitter is a challenging problem, where unstructured, informal and irregular content should be properly handled. The identification of an appropriate feature set is important to build classification schemes with high predictive performance. In the earlier work on sentiment analysis of Twitter data, n-grams, part of speech tags and microblogging based features are considered.

In this paper, we examined the predictive performance of psychological and linguistic features obtained by LIWC on sentiment analysis on Twitter. For this purpose, five main LIWC categories (namely, linguistic processes, psychological processes, personal concerns, spoken categories and punctuation) and their combinations are taken as feature sets. The experimental analysis with classification algorithms indicate that psycholinguistic feature sets can yield encouraging results on sentiment analysis of Twitter data. The experimental analysis indicates that ensemble feature sets outperforms the individual feature sets. For sentiment analysis on Twitter, the highest predictive performance (89.10%) is achieved by by combining linguistic processes, psychological processes and personal concerns with Random Subspace ensemble of Naïve Bayes.

REFERENCES

- [1] A. Onan, "Twitter mesajları üzerinde makine öğrenmesi yöntemlerine dayalı duygu analizi", *Yönetim Bilişim Sistemleri Dergisi*, Vol. 3, No. 2, 2017, pp. 1-14.
- [2] A. Onan, S. Korukoğlu, and H. Bulut, "A multiobjective weighted voting ensemble classifier based on differential evolution algorithm for text sentiment classification", *Expert Systems with Applications*, Vol.62, 2016, pp.1-16.
- [3] A.Onan, "A machine learning based approach to identify geo-location of Twitter users", in *Proceedings of the ICC 2017*, UK, 2017, pp.1-7.
- [4] J. Mahmud, J. Nichols, and C. Drews, "Home location identification of twitter users", *ACM Transactions on Intelligent Systems and Technology*, Vol. 5, No.3, 2014, pp.47.
- [5] Z. Cheng, J. Caverlee, and K.Lee, "You are where you tweet: a content-based approach to geo-location twitter users", in *Proceedings of the 19th ACM International Conference on Information and Knowledge Management*, USA, 2010, pp.759-768.
- [6] B.Hecht, L.Hong, B. Suh and E.D.Chi, "Tweets from Justin Bieber's heart: the dynamics of the location field in user profiles", in *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, USA, 2011, pp.237-246.
- [7] A. Onan and S. Korukoğlu, "Makine öğrenmesi yöntemlerinin görüş madenciliğinde kullanılması üzerine bir literatür araştırması", *Pamukkale Üniversitesi Mühendislik Bilimleri Dergisi*, Vol. 22, No. 2, 2016, pp. 111-122.
- [8] W. Medhat, A. Hassan and H. Korashy, "Sentiment analysis algorithms and applications: a survey", *Ain Shams Engineering Journal*, Vol. 5, No. 4, 2014, pp. 1093-1113.
- [9] A. Onan and S. Korukoğlu, "A feature selection model based on genetic rank aggregation for text sentiment classification", *Journal of Information Science*, Vol. 43, No.1, 2017, pp.25-38.
- [10] M.P. Salas-Zarate, E.Lopez-Lopez, R.Valencia-Garcia, N. Gilles, A.Almela and G.Alor-Hernandez, "A study on LIWC categories for opinion mining in Spanish reviews", *Journal of Information Science*, Vol.40, No.6, 2014, pp.749-760.
- [11] A.Go, R. Bhayani, and L. Huang, "Twitter sentiment classification using distant supervision", *CS224N Project Report*, 2009.
- [12] L. Barbosa and J. Feng, "Robust sentiment detection on twitter from biased and noisy data", in *Proceedings of ACL*, USA, 2010, pp. 36-44.
- [13] A.Pak and P.Paroubek, "Twitter as a corpus for sentiment analysis and opinion mining", in *Proceedings of LREC 2010*, USA, 2010, pp. 1320-1326.
- [14] E. Kouloumpis, T.Wilson and J.D.Moore, "Twitter sentiment analysis: the good, the bad and the omg!", in *Proceedings of ICWSM 2011*, USA, 2011, pp. 538-541.
- [15] A.Agarwal, B.Xie, I.Vovsha, O.Rambow and R. Passonneau, "Sentiment analysis of twitter data", in *Proceedings of ACL 2011*, USA, 2011, pp. 30-38.
- [16] H.Saif, Y.He and H.Alani, "Semantic sentiment analysis of twitter", in *Proceedings of ISWC 2012*, USA, 2012, pp.508-524.
- [17] M.Salas-Zarate, M.A. Paredes-Valverde, M.A.Rodriguez-Garcia, R.Valencia-Garcia and G.Alor-Hernandez, "Automatic detection of satire in Twitter: a psycholinguistic-based approach", *Knowledge-Based Systems*, Vol.128, 2017, pp.20-33.
- [18] J.M.Cotelo, F.L.Cruz, J.A.Troyano and F.J.Ortega, "A modular approach for lexical normalization applied to Spanish tweets", *Expert Systems with Applications*, Vol. 42, No.10, 2015,pp. 4743-4754.
- [19] E.Kontopoulos, C.Berberidis, T.Dergiades and N.Bassiliades, "Ontolog-based sentiment analysis of twitter posts", *Expert Systems with Applications*, Vol.40, No.10, 2013, pp.4065-4074.
- [20] R.Justo, T.Corcoran, S.M.Lukin, M.Walker and M.I.Torres, "Extracting relevant knowledge for the detection of sarcasm and nastiness in the social web", *Knowledge-Based Systems*, Vol. 69, 2014, pp.124-133.

- [21] S.Skalicky and S.Crossley, "A statistical analysis of satirical Amazon.com product reviews", *European Journal of Humour Research*, Vol.2, 2015, pp.66-85.
- [22] J.W.Pennebaker, R.L.Boyd, K.Jordan and K.Blackburn, "The development and psychometric properties of LIWC 2015".
- [23] A.Onan, "Classifier and feature set ensembles for web page classification", *Journal of Information Science*, Vol. 42, No.2, pp.150-165.
- [24] A.Onan, "Sarcasm identification on twitter: a machine learning approach", in *Proceedings of CSOC 2017*, Germany, 2017, pp.374-383.
- [25] M.Kantardzic, *Data mining: concepts, models, methods and algorithms*, John Wiley & Sons, 2011, p.552.
- [26] L.Breiman, "Bagging predictors", *Machine Learning*, Vol.4, No.2, pp.123-140.
- [27] Y.Freund and R.E.Schapire, "Experiments with a new boosting algorithm", in *Proceedings of the Thirteenth International Conference on Machine Learning*, Italy, 1996, pp.148-156.
- [28] T.K. Ho, "The random subspace method for constructing decision forests", *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol. 20, No.8, pp.832-844.
- [29] A.Onan, "Artificial immune system based web page classification", in *Proceedings of CSOC 2015*, Germany, 2015, pp.189-199.

### BIOGRAPHIES



**AYTUĞ ONAN** was born in 1987. Dr. Aytuğ Onan received his BS. in Computer Engineering from Izmir University of Economics (Turkey) in 2010. He earned his MS in Computer Engineering and PhD in Computer Engineering from Ege University (Turkey) in 2013 and 2016, respectively. He has been working as "assistant professor" since January 2017 at the Department of Software Engineering of Celal Bayar University (Turkey). He has been reviewing for several international journals, including *Expert Systems with Applications*, *Plos One*, *International Journal of Machine Learning and Cybernetics*, *Journal of Information Science*. He has published several journal papers on machine learning and computational linguistics.

# Classification of PVC Beat in ECG Using Basic Temporal Features

Y. Kaya


**Abstract**—Premature ventricular contraction (PVC) is one of the most important arrhythmias among the various hearth abnormalities. Premature depolarization of the myocardium in the ventricular region causes PVC and it is usually associated with structural heart conditions. Arrhythmias can be detected by examining the ECG signal and this review requires large-size data to be examined by physicians. The time spent by the physician in examining the signal can be reduced using CAD systems. In this study, we propose a high performance PVC detection system using the feature extraction and classification scheme bringing low computational burden. The test set consisting of 81844 beats from the MIT-BIH arrhythmia database was used for the experimental results. We compared the performances of the various classifiers using proposed feature set in the experiments and obtained classification accuracy of 98.71% using NN classifier.

**Index Terms**—Arrhythmia, Classification, Decision Tree, Heartbeat, k-Nearest Neighbor, k-NN, Neural Network, Premature Ventricular Contraction, PVC, Support Vector Machine.

## I. INTRODUCTION

AN ECG IS A SIGNAL that can be easily obtained with electrodes placed on the human body, containing important information indicating the abnormal state of the cardiovascular system. Detection of different types of heartbeats is vital to identify cardiac disorders. Premature ventricular contraction (PVC) is one of the most important arrhythmias among the various anomalies related to cardiac rhythms [1], [2]. Premature depolarization of the myocardium in the ventricular region causes PVC and it is a common arrhythmia usually found in adults. It is estimated to have a prevalence of between 1% and 4% of the general population and usually associated with structural heart conditions and increases the risk of sudden death [2].

ECG signals need to be analyzed to detect arrhythmias, and this analysis is a time-consuming process that requires cardiologists to examine large-scale data. The accurate and rapid detection of PVC in the ECG signal is closely related to the correct identification of the features to represent a heartbeat. When PVC beat is examined, it is quite easy to distinguish it from normal sinus rhythm.

Y. KAYA, is with Department of Computer Engineering of Karadeniz Technical University, Trabzon, Turkey, (e-mail: [yasın@ktu.edu.tr](mailto:yasın@ktu.edu.tr)).   
Manuscript received August 9, 2017; accepted Nov 16, 2017.  
DOI: [10.17694/bajece.419541](https://doi.org/10.17694/bajece.419541)

The signals can accurately and quickly be distinguished between large amounts of data by using computer aided diagnosis (CAD) systems and the time spent by the physician in examining the signal can be reduced. At this point, it is important to develop methods to distinguish between PVC and normal beats.

Researchers have done a lot of work for detecting of the PVC. The authors have proposed different models for feature extraction, noise elimination, feature size reduction and classification schemes for classification of arrhythmias in the following articles.

Liu et al. proposed a deep learning method for the recognition of PVC in children [3]. Kaya and Pehlivan compared various methods to classify PVC and investigated the model that gives the best result [2], [4]. Zhou et al. used deep neural network (NN) and rule inference to detect PVC [5]. Xiuling et al. proposed a model containing Lyapunov exponents and LVQ neural network to classify PVC beats [6].

Bortolan et al. compared the classification capabilities and the learning capacities of the k-nearest neighbor (k-NN), NN, fuzzy logic (FL) and, discriminant analysis (DA) classifiers for distinguish normal and PVC beats. The authors used 26 shape features in their work, which consisted of amplitude information, area, special interval times and QRS metrics. k-NN classifier reported to be more effective than other classifiers [7]. Ebrahimzadeh and Khazaei proposed a method to distinguish the PVC beat from normal and other beats. The authors used wavelet transform to eliminate noise in the ECG signal. They used one temporal and 10 morphologic features [8]. In another study, Christov et al. used both leads from the ECG signals from the MIT-BIH arrhythmia database to extract the feature for classification of PVC beats. They used k-NN as the classifier in the study and achieved the classification accuracy of 96.7% [9]. Jenny et al. used discrete wavelet transform to reduce noise in the signal, independent component analysis for dimension reduction and k-means and fuzzy c-means to classify for PVC beats [10].

In recent years, researchers have proposed new studies on wearable ECG analysis systems and mobile ECG analysis systems. One of the most important constraints for the systems is the calculation load. For this reason, it is very important for new methods developed to bring a low calculation burden [11].

In this study, we propose an approach for the classification of normal (N) and PVC beats. The main purpose of the work is to realize a high performance PVC detection system using

the feature extraction and classification scheme bringing low computational burden. Unlike other studies in the literature, we achieved high classification performance for the classification of PVC beats using three basic features. Extraction of the features from the signal did not cause calculation loads on the system. For the experimental tests, the test set consisting of 81844 beats from the MIT-BIH arrhythmia database [12] was used in the study. We used k-NN, NN, support vector machine (SVM) and, decision trees (DT) classifiers to calculate and compare experimental test results. We distinguished between PVC and N beats with 98.71% classification accuracy using the NN classifier.

## II. MATERIAL AND METHOD

### A. ECG Database

MIT-BIH arrhythmia database consisted of 48 ECG records with two-channel. Each record contained about 30 minutes ECG data. All of these records were obtained from 47 patients examined by the BIH Arrhythmia laboratory between 1975 and 1979. The first section of the database, numbered 100-124, was generated from 23 randomly selected 24-hour records. The second part of the database was carefully selected by cardiologists and numbered 200-234 for important clinical events [12], [13]. Fig.1. shows the normal beat and three types of PVC beat.

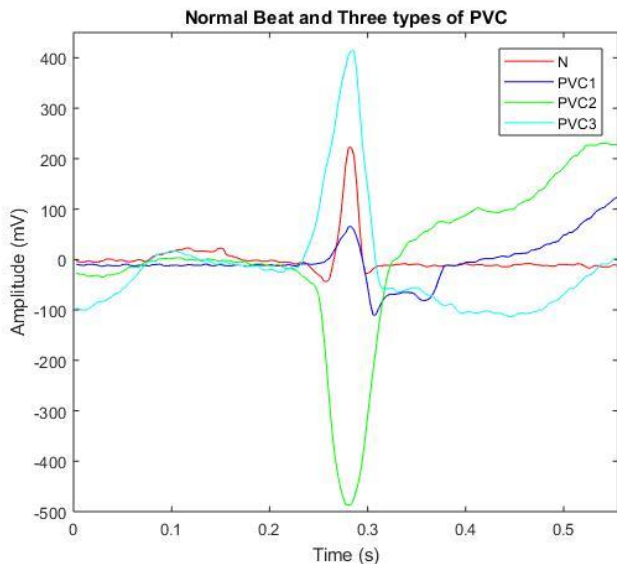


Fig.1. Normal beat and three types of PVC.

In the study, we used 46 ECG records containing MLII lead from the MIT-BIH arrhythmia database but two records without MLII were not preferred.

### B. Classification of PVC beats

Decision-making in the classification of ECG beats is a three-step process similar to other machine learning methods.

**Signal preprocessing:** The fluctuations in the ECG signal are removed. In this step, the noise-free signal is split into beats and ready for feature extraction.

**Feature extraction:** The features to represent a beat are

calculated using specified mathematical and statistical calculation methods. The method used at this step affects the accuracy of classification. The identification of better representing features of the beats provides the classification algorithm to learn better the data during the training step and gives better results in the test step.

**Classification:** In the classification step, tests are carried out with the proposed classification scheme. At this stage, a classification model must be determined according to the problem.

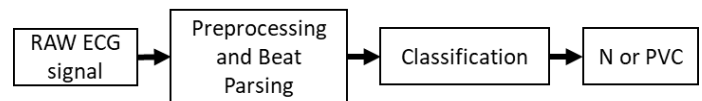


Fig.2. Proposed PVC detection approach

Fig.2. demonstrates the proposed approach for the detection of PVC beats. We determined three attributes in the feature calculation step in the study. These were the previous R-R interval (RRP), the next R-R interval (RRN), and arithmetic mean values of 50 amplitude values (MEAN) centered active R peak. The detection process successfully was performed with three features that can be obtained simply from the signal.

## III. EXPERIMENTAL RESULTS

We calculated the experimental results using the records from the MIT-BIH arrhythmia database in the study. Since the previous and next R-R values were calculated, the first and last beat in the signal files were not included in the analysis. Except for these beats, all PVC and N beats in the specified 46 signal files were included in the experiment set. These consisted of 7122 PVC beats and 74722 normal beats.

We used Matlab software to remove noise from the signal, beat parsing, and feature calculation steps. A beat parsing step was performed to obtain the signal values indicating a heart cycle. We calculated three features from these values, used as input vector in the classifiers, and evaluated classification performances.

### A. Preprocessing and beat parsing

The signal was passed through various filters and the fluctuations found in the signal were reduced [2], [14], [15]. The most important of these fluctuations is the baseline wander that occurs with daily movements such of the patient as breathing, swelling, coughing, etc. during the ECG recording [15]. We removed the frequency components below 2 Hz from the signal using a high-pass filter to eliminate baseline wander [2].

### B. Feature Calculation

We calculated the RRP, RRN and MEAN values used for the classification step for each beat in the feature calculation step. Fig.3. shows the calculation of RRP and RRN features. These values are the difference operation on the time axis.

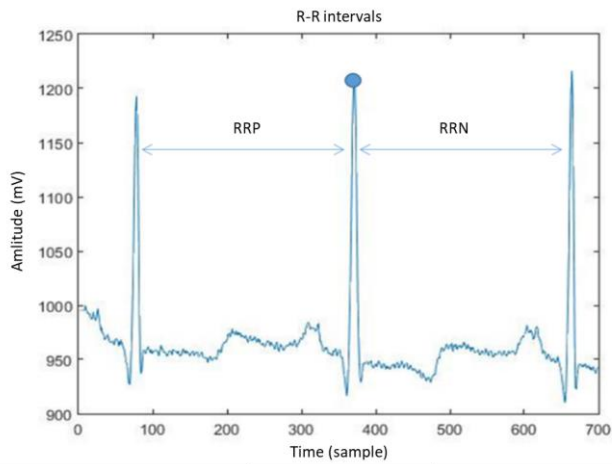


Fig.3. RRP and RRN features

Fig.4. shows the calculation of the MEAN feature. The window width was set to 50 for the calculation of the MEAN feature. We constructed a series of 50 sample amplitude values centered on R peak. Using these values, we calculated MEAN attribute. The notes in the database were used to determine the R peak. The arithmetic mean was calculated according to Equation (1) using the 50 amplitude values shown in Fig 4.

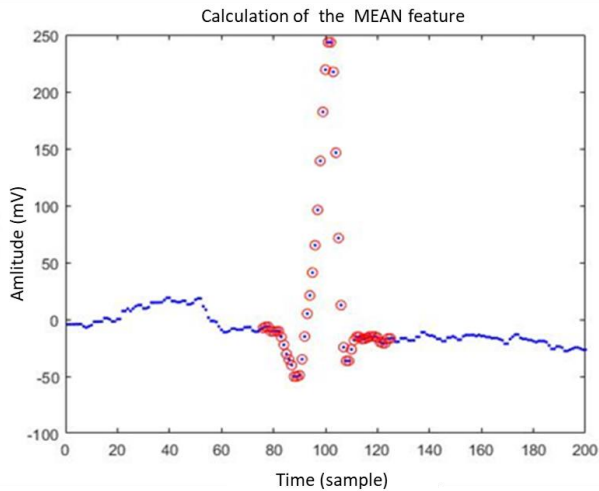


Fig.4. Calculation of MEAN feature

$$MEAN = \frac{\sum_{i=1}^n x_i}{n} \tag{1}$$

where  $x_i (\forall i \in 1..n)$  is the R peak centered amplitude values, n is the window width, which is 50.

C. Classification

We calculated three features in the previous step. In this step, the classification accuracy was calculated by classifying the features using k-NN, NN, SVM, and DT classifiers. The heartbeats shown in Table 1 were gathered from the MIT-BIH arrhythmia database. Normal beat was selected from 38

subjects and PVC beat from 35 subjects. We used 10-fold cross validation to evaluate results in the classification step. In 10-fold cross validation, the test data were divided into 10 separate pieces of equal width. One section used to test and the remaining nine were used for training the system at each step. After 10 repetitions, the overall performance of the system was calculated by evaluating the average of the classification performance achieved at each step.

We used classification accuracy, specificity, and sensitivity performance metrics to evaluate the results. Accuracy is defined as the ratio of the number of correctly classified samples to the total number of samples. Sensitivity is the ratio of the number of correctly classified positive samples to the total number of positive samples. Specificity is the ratio of the number of samples belonging to a correctly classified class to the total number of samples estimated for that class.

The accuracy, sensitivity and specificity measures are calculated from the confusion matrix using equations (2)-(4).

$$Accuracy = \frac{TP + TN}{TP + FP + FN + TN} \tag{2}$$

$$Sensitivity = \frac{TP}{TP + FN} \tag{3}$$

$$Specificity = \frac{TN}{TN + FP} \tag{4}$$

where TP is correctly classified normal beat, TN is classified correctly PVC beat, FP is misclassified normal beat, FN is misclassified PVC beat.

TABLE I  
HEART BEATS USED IN EXPERIMENTS AND RELATED RECORDS

Signal Files	N	PVC	Total
100	2237	1	2238
101	1858		1858
103	2080		2080
105	2524	41	2565
106	1505	520	2025
107		59	59
108	1737	17	1754
109		38	38
111		1	1
112	2535		2535
113	1787		1787
114	1818	43	1861
115	1951		1951
116	2300	109	2409
117	1532		1532
118		16	16
119	1541	444	1985
121	1859	1	1860
122	2474		2474
123	1513	3	1516
124		47	47
200	1742	825	2567
201	1623	198	1821
202	2059	19	2078
203	2527	444	2971

Signal Files	N	PVC	Total
205	2569	71	2640
207		105	105
208	1585	992	2577
209	2619	1	2620
210	2421	194	2615
212	922		922
213	2639	220	2859
214		256	256
215	3193	164	3357
217	244	162	406
219	2080	64	2144
220	1952		1952
221	2029	396	2425
222	2060		2060
223	2027	473	2500
228	1686	362	2048
230	2253	1	2254
231	314	2	316
233	2229	830	3059
234	2698	3	2701
<b>Total</b>	<b>74722</b>	<b>7122</b>	<b>81844</b>

#### IV. DISCUSSIONS

In the study, 81844 samples of two classes from the MIT-BIH arrhythmia database were analyzed. In the experimental tests, almost all of the PVC and normal beats in the database were used. The arithmetic mean was calculated from 50 amplitude values of a beat and RRP and RRN temporal difference features were calculated for the classification process. These three features were fed to the k-NN, NN, SVM, and DT classifiers and the results were compared.

Table 2 shows the classification results. In the classification step, the k parameter for the k-NN classifier was set to one. A forward feed NN trained by the backpropagation algorithm was used. There was one hidden layer in the used architecture and the number of nodes in the hidden layer was set to 20 in the experimental tests. Another classifier used in experiments was SVM. In SVM, kernel type was defined as the radial based function (RBF) and gamma and C parameters were used as zero. DT was the last classifier used in experiments. We determined the maximum depth as 20, the minimum gain as 0.2, the minimum branch size as two, and minimum division number as four in the classification step with the KA. Experiments demonstrated that the best result was obtained with NN.

TABLE II  
CLASSIFICATION RESULTS (%)

Method	Accuracy	Sensitivity <sup>a</sup>	Specificity <sup>a</sup>
k-NN	98.58	91.52	99.26
NN	98.71	90.80	99.46
SVM	98.55	87.67	99.58
DT	98.30	83.60	99.70

a. Specificity and Sensitivity values were calculated for the case where the PVC beat was positive class.

Table 3 shows the confusion matrix for the test performed with NN. In the study, high performance was achieved by using three simple time domain attributes. The fact that all PVC beats in the database were used in the experiments confirms the validity of the results obtained in the study.

TABLE III  
CONFUSION MATRIX OF NN CLASSIFICATION

Class	True N	True PVC	Class Precision
<b>Prediction N</b>	74321	655	99.13%
<b>Prediction PVC</b>	401	6467	94.16%
<b>Class Recall</b>	99.46%	90.80%	

The proposed method shows that high performance can be achieved by using only three features when compared with other studies in this topic. Christov et al. classified PVC beats and reported the classification performance of sensitivity of 96.9% and specificity of 96.7% [9]. Jenny et al. used an unsupervised learning method and therefore achieved a lower performance than the other recommended methods [10]. Similarly, in other study, the authors classified PVC beats using k-NN and obtained specificity of 98.7% and sensitivity of 91.3% [7].

Table 4 summarizes the methods, the number of features, and the classification performance achieved in the proposed approach and the related studies. In our previous work in the same topic, we used 200 amplitude values for the classification of PVC beats. These 200 data were reduced to lower numbers using mathematical models [2].

TABLE IV  
COMPARISON WITH OTHER STUDIES

The Authors	Method	Feature Size	Classification Accuracy
Bortolan et al. [7]	k-NN	26	98.7% spe. 91.3% sen.*
Liu et al. [3]	1D CNN	486	83%
Zhou et al. [5]	CDNN	150	99.41%
Christov et al. [9]	k-NN	26	96.7% spe. 96.9% sen.*
Ebrahimzadeh et al. [8]	NN	11	95.37%
Jenny et al. [10]	Fuzzy C-means	-	80.94%
Kaya et al. [2]	k-NN	17	99.63%
<b>Proposed Approach</b>	<b>NN</b>	<b>3</b>	<b>98.71%</b>

\* Some authors did not specified classification accuracy. For this reason, specificity (spe.) and sensitivity (sen.) metrics are shown in the table.

In a more recent study, Liu et al. proposed a deep learning based method for perceiving PVC beats in children. The authors recorded the test data themselves and obtained the correct classification accuracy of 83% using the recommended method [3]. In a similar study using combined deep neural networks and rules inference, the authors achieved a classification success of 99.41%. The authors used the experimental set consisting of 3194 PVC beats and 46329 normal beats to achieve this success [5].

Researchers have worked extensively on these issues and have proposed complex models. The proposed models bring

the computational burden. Used methods for feature extraction in the studies summarized in Table 4 were complex and computationally expensive.

## V. CONCLUSIONS

In this study, we propose a classification scheme based on NN classifier in order to obtain the highest performance with minimum account load. The model uses three attributes that are easy to calculate in time domain. In this respect, the study uses fewer features than other studies in the literature. The use of almost all N and PVC beats in the database confirms the validity of the results obtained by the proposed method.

Due to the small number of attributes and simple calculation, and the simplicity of the classification scheme used, the proposed method can be integrated with the mobile ECG recording and analysis applications to be developed. The proposed model can be used in wearable ECG systems because it can be classified with low complexity and high accuracy.

## REFERENCES

- [1] G. K. Lee, K. W. Klarich, M. Grogan, and Y.-M. Cha, "Premature ventricular contraction-induced cardiomyopathy: a treatable condition.," *Circ. Arrhythm. Electrophysiol.*, vol. 5, no. 1, pp. 229–36, Feb. 2012.
- [2] Y. Kaya and H. Pehlivan, "Classification of Premature Ventricular Contraction in ECG," *Int. J. Adv. Comput. Sci. Appl.*, vol. 6, no. 7, pp. 34–40, 2015.
- [3] Y. Liu, Y. Huang, J. Wang, L. Liu, and J. Luo, "Detecting Premature Ventricular Contraction in Children with Deep Learning," *J. Shanghai Jiaotong Univ.*, vol. 23, no. 1, pp. 66–73, Feb. 2018.
- [4] Y. Kaya and H. Pehlivan, "Classification of Premature Ventricular Contraction Beat Using Basic Temporal Features," in *International Advanced Researches & Engineering Congress-2017*, 2017, pp. 1313–1318.
- [5] F. Zhou, L. Jin, and J. Dong, "Premature ventricular contraction detection combining deep neural networks and rules inference," *Artif. Intell. Med.*, vol. 79, pp. 42–51, Jun. 2017.
- [6] X. Liu, H. Du, G. Wang, S. Zhou, and H. Zhang, "Automatic diagnosis of premature ventricular contraction based on Lyapunov exponents and LVQ neural network.," *Comput. Methods Programs Biomed.*, vol. 122, no. 1, pp. 47–55, Oct. 2015.
- [7] G. Bortolan, I. Jekova, and I. Christov, "Comparison of four methods for premature ventricular contraction and normal beat clustering," in *Computers in Cardiology*, 2005, vol. 32, pp. 921–924.
- [8] A. Ebrahimzadeh and A. Khazaei, "Detection of premature ventricular contractions using MLP neural networks: A comparative study," *Meas. J. Int. Meas. Confed.*, vol. 43, pp. 103–112, 2010.
- [9] I. Christov, I. Jekova, and G. Bortolan, "Premature ventricular contraction classification by the K th nearest-neighbours rule," *Physiol. Meas.*, vol. 26, no. 1, pp. 123–130, Feb. 2005.
- [10] N. Z. N. Jenny, O. Faust, and W. Yu, "Automated Classification of Normal and Premature Ventricular Contractions in Electrocardiogram Signals," *J. Med. Imaging Heal. Informatics*, vol. 4, no. 6, pp. 886–892, Dec. 2014.
- [11] M. M. Baig, H. Gholamhosseini, and M. J. Connolly, "A comprehensive survey of wearable and wireless ECG monitoring systems for older adults," *Med. Biol. Eng. Comput.*, vol. 51, no. 5, pp. 485–495, May 2013.
- [12] G. B. Moody and R. G. Mark, "The impact of the MIT-BIH arrhythmia database.," *IEEE Eng. Med. Biol. Mag.*, vol. 20, no. 3, pp. 45–50, 2001.
- [13] G. Moody and R. Mark, "The MIT-BIH Arrhythmia Database on CD-ROM and software for use with it," in *[1990] Proceedings Computers in Cardiology*, 1990, pp. 185–188.

- [14] Y. Kaya and H. Pehlivan, "Comparison of classification algorithms in classification of ECG beats by time series," in *23rd Signal Processing and Communications Applications Conference (SIU)*, 2015, pp. 407–410.
- [15] Y. Kaya, H. Pehlivan, and M. E. Tenekeci, "Effective ECG beat classification using higher order statistic features and genetic feature selection," *Biomed. Res.*, vol. 28, no. 17, pp. 7594–7603, 2017.

## BIOGRAPHIES



**YASIN KAYA** was born in Adana, in 1979. He received the B.S. degree in statistics and computer science from the Karadeniz Technical University, in 1999. He received M.S. and Ph. D. degrees in Computer Engineering from the Karadeniz Technical University, Trabzon, in 2006 and 2017, respectively.

From 1999 to 2015, he was a lecturer with the Department of Informatics. Since 2015, he has been a Lecturer with the Distance Learning Application and Research Center, Karadeniz Technical University. He is the author of four books, three international indexed (SCIE, ESCI) articles, and eight international conference proceedings. His research interests include biomedical signal processing, arrhythmia detection, pattern recognition, and image processing.



# FPGA-based ANN Design for Detecting Epileptic Seizure in EEG Signal

B. Karakaya, T. Kaya and A. Gulden

**Abstract**—This study aims to represent an FPGA (Field Programmable Gate Array) design of Artificial Neural Network (ANN) for Electroencephalography (EEG) signal processing in order to detect epileptic seizure. For analyzing brain's electrical activity, feedforward ANN model is used for classification of EEG signals. The designed ANN output layer makes a decision whether the person has epilepsy or not. In the proposed system, the ANN model is programmed and simulated on Xilinx ISE editor via computer and then, EEG signal data are transferred to FPGA-based ANN emulator core. The Core is trained on data which are patient's data and healthy person's data. After training, test data is loaded to ANN Emulator Core to detect any epileptic seizure of person's EEG signal. The main advantage of FPGA in the system is to improve speed and accuracy for epileptic seizure detection.

**Index Terms**—ANN, EEG, FPGA, Epilepsy.

## I. INTRODUCTION

ELECTROENCEPHALOGRAM (EEG) which is obtained from recording of brain's electrical activity is important data to analyze brain's normal and abnormal activities. Epilepsy that is significant disease of brain is a chronic disease which causes sensory loss, unbalanced deictic gesture or muscular contraction comprised by abnormal activity of a group of neuron in brain. On the recognition of this disease, analysis of EEG has great importance [1].

In the analysis of EEG signal, many methods are used. In [2], high frequency and low frequency noise were suppressed by moving average and derivative-based filter. This method was used to classify normal or epileptic EEG signals. In [3], the user interface program was generated in Laboratory Virtual Instrument Engineering Workbench (LabVIEW) that has visual programming language in order to analyze EEG signals in determination of sleep stages. EEG signals can be analyzed in two domains. Due to the characteristic of signal in frequency domain, signal differs from before, during and after attack. Analyzing the characteristic of signal in time domain gives better result.

**B. KARAKAYA**, works at Department of Electrical-Electronics Engineering, Faculty of Engineering, Firat University, Elazig, Turkey, (e-mail: bkarakaya@firat.edu.tr).

**T. KAYA**, works at Department of Electrical-Electronics Engineering, Faculty of Engineering, Firat University, Elazig, Turkey, (e-mail: tkaya@firat.edu.tr).

**A. GULTEN**, works at Department of Electrical-Electronics Engineering, Faculty of Engineering, Firat University, Elazig, Turkey, (e-mail: agulden@firat.edu.tr).

Manuscript received August 9, 2017; accepted Nov 16, 2017.

DOI: [10.17694/bajece.419544](https://doi.org/10.17694/bajece.419544)

In literature, there are many studies which are based time domain but the best one is Wavelet Transform Technique on epileptic seizure detection [4], [5]. There are also classification studies of EEG signal by using ANN model. In [6], the design of a new window function that has side-lobe roll-off ratio characteristic of ultra-spherical window and Kaiser Window's main-lobe width and ripple ratio was obtained by helping ANN. ANN model can be used for pattern recognition as well as EEG signal processing and classification [7], [8].

In literature, there are few studies on classification of EEG signals based on FPGA using neural network algorithms. One of them has two neural network algorithms that are implemented with the best accuracies into FPGA which achieves on 68% accuracy for MIT-BIH data and 70% accuracy for Mitra data [9]. In [10], simulation platform is introduced and starting from simulation in the learning phase with fixed-point operators, a methodology has been developed that is able to realize EEG signal processing with ANN model. The aim of this paper is to process EEG signal that is filtered in time domain by using ANN architecture, increase operating frequency and parallel processing ability of design. Furthermore, designed ANN model is programmed on FPGA and then signal is classified. ANN is preferred for its speed and parallel processing ability. Also, ANN can solve complex mathematical problems in real-time based on observations.

## II. MATERIALS AND METHODS

### A. The Study Area

Nowadays, new disease recognition implementations attract researcher's attention who work in hospitals and biomedical device industry because of faster and more accurate results requirement.

In this study, three groups of data [11] are formed from pre-processed EEG signal. These are patient's data, healthy person's data and test data. ANN is trained by transferring patient's data and healthy person's data to FPGA-based ANN Emulator Core and processing respectively. After training, test data is loaded to ANN Emulator Core on Xilinx ISE simulator to detect any epileptic seizure of person's EEG signal.

### B. The Experimental Design

This simulation study is named as FPGA-based ANN Emulator Core and illustrated in Figure 1. ANN Emulator Core consists of two parts. They are Adder and Multiplier

circuits. EEG data that are obtained from patient and healthy person, are filtered and normalized -1 to 1. These data must be transformed to a binary number format which has to be in fractional mode. Therefore, fixed-point number format is chosen. Numbers are arranged as 16 bit in width and quantized in the range of -1 to 1.

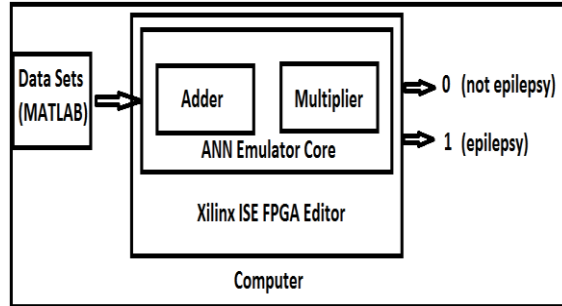


Fig. 1. A scheme of FPGA-based ANN Emulator Core Design.

Fixed-point number format is symbolized as Qm.n where m and n stand for integer part and fractional part of number respectively. All numbers for this design are used as signed form. Therefore, 1 bit is reserved for sign bit. Finally, number representation format is organized as Q2.13 16 bit signed fixed-point format for MATLAB. In this case, the precision of numbers is obtained as 0,122. 10<sup>-4</sup>.

Network that is programmed in the study is shown in Figure 2. As shown in Figure 2, ANN model has 4 inputs, 1 hidden layer, 1 output layer and 2 activation functions with input, bias and coefficient weightiness. ANN model is trained toward given inputs and weightiness by solving these equations below.

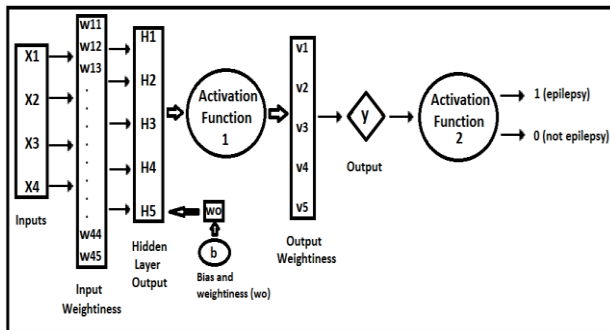


Fig. 2. Artificial Neural Network Processor Design.

$$\begin{aligned}
 H_1 &= x_1 * w_{11} + x_2 * w_{21} + x_3 * w_{31} + x_4 * w_{41} + b * wo_1 \\
 H_2 &= x_1 * w_{12} + x_2 * w_{22} + x_3 * w_{32} + x_4 * w_{42} + b * wo_2 \\
 H_3 &= x_1 * w_{13} + x_2 * w_{23} + x_3 * w_{33} + x_4 * w_{43} + b * wo_3 \\
 H_4 &= x_1 * w_{14} + x_2 * w_{24} + x_3 * w_{34} + x_4 * w_{44} + b * wo_4 \\
 H_5 &= x_1 * w_{15} + x_2 * w_{25} + x_3 * w_{35} + x_4 * w_{45} + b * wo_5
 \end{aligned}
 \tag{1}$$

Activation function 1 is given as below

$$H_{(i)} = \begin{cases} -1, \dots, H_{(i)} < -1 \\ H_{(i)}, \dots, |H_{(i)}| \leq 1 \\ 1, \dots, H_{(i)} > 1 \end{cases}
 \tag{2}$$

and concluding calculation is  $y_{net}$ .

$$y_{net} = H_1 * v_1 + H_2 * v_2 + H_3 * v_3 + H_4 * v_4 + H_5 * v_5
 \tag{3}$$

Activation function 2 is given as below

$$y = \begin{cases} 1, \dots, y_{net} > 0 \\ 0, \dots, y_{net} < 0 \end{cases}
 \tag{4}$$

and decision is determined as,

$$error = y_{desired} - y
 \tag{5}$$

if error is greater than required error limit, all weightiness are updated as below equations where  $\alpha$  is training parameter. Then, ANN model is reworked with new weightiness. If not, y is the correct output.

$$\begin{aligned}
 \Delta w &= error * input(j) * w(i, j) \\
 w &= w + \alpha * \Delta w \\
 \Delta v &= error * H(i) * v(i) \\
 v &= v + \alpha * \Delta v \\
 \Delta wo &= error * wo(i) \\
 wo &= wo + \alpha * \Delta wo
 \end{aligned}
 \tag{6}$$

Arithmetic circuits in the design that are Adder and Multiplier Circuits are coded and synthesized on Xilinx ISE FPGA Editor. These two arithmetic circuits give the response in 2 clock cycles. Figure 3 illustrates port information of arithmetic circuits [12].

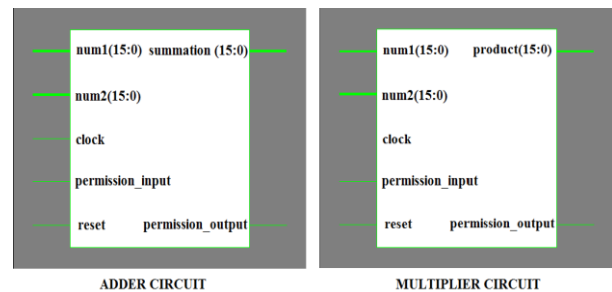


Fig. 3. Adder & Multiplier Circuit of ANN Emulator Core [12].

Arithmetic circuits begin to work with *permission\_input* signal. After addition or multiplication operation, arithmetic circuits send *permission\_output* signal with result in order to indicate that circuits are ready for new calculation.

Table 1. Resource utilization and latencies of Adder and Multiplier circuits [13].

	Adder	Multiplier
Used Logic Slices	9	35
Used Flip-Flops	1	20
Used LUTs	18	61
Latency	2	2

Table 1 represents resource utilization and latencies of Adder and Multiplier circuits which are used for ANN Core design. Table 2 represents resource utilization of ANN Core design overall.

Table 2. Resource utilizations of ANN Core design.

Used Logic Slices	9255
Used Flip-Flops	11246
Used LUTs	17149

### III. RESULTS

In the simulator screen, clock cycle is selected as 100 Mhz. Total clock cycle that is needed to complete emulation of signal is 136. It means total time that is needed for emulation is 1360 ns. After training completed by using ANN model above, patient’s and healthy person’s data are implemented to the emulator core respectively. It is required that when a group of patient’s data are applied to the core, it must give a result of 1 while in the case of healthy person’s data, it is 0. Figure 4

shows the outcome of ANN Core when a group of healthy person’s data are applied to the core. Figure 5 shows the outcome of ANN Core when a group of patient’s data are applied to the core. Figure 6 shows the outcome of ANN Core when a group of healthy person’s data are applied to the core, but in this case ydesired is selected as the person has epilepsy. Therefore, well-trained ANN gives an error as 1. Because the person is healthy.

In outcomes of ANN Emulator Core, y[15:0] represents output of the ANN model that is used for detection of seizure on EEG signal. e\_hata[15:0] represents error of the ANN model calculation that is used for recalculation of weightiness if it is greater than 0. ydesired[15:0] is used for if it is 1, it means the signal is belong to epilepsy patient. If it is 0, it means the signal is belong to healthy person.

The accuracy of the study can be determined by using Mean Square Error (MSE) algorithm after signal progressed. A MATLAB function is created and the design achieves on 86.7% accuracy of detection epilepsy event.

### IV. CONCLUSION

As a result of this study, it is proven that EEG signal can be classified as normal or epileptic by using ANN Core design on FPGA platform with more accuracy. Furthermore, the total required time to classify EEG signal is 1360 ns. Maximum operating frequency is obtained as 82 MHz from Xilinx Synthesis Tool. The speed of real-time implementation may change respect to design.

ANN Core design can be updated and implemented in real-time on FPGA. Furthermore, fixed-point arithmetic can be arranged as obtaining better precision.

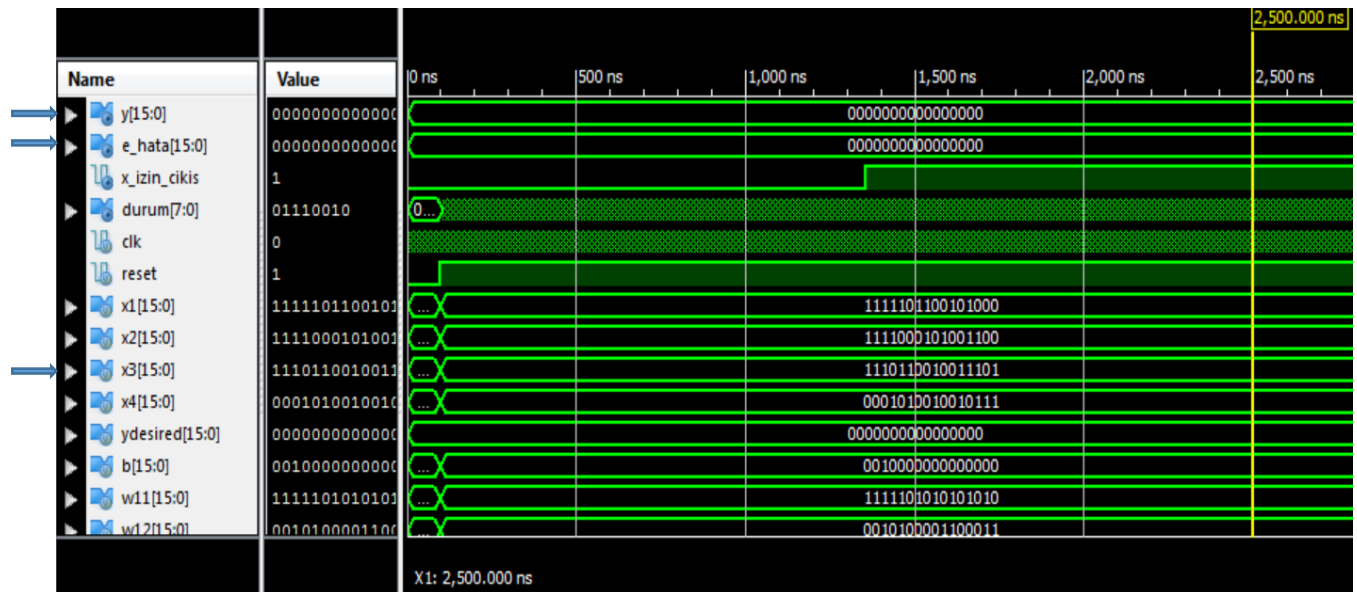


Fig. 4. Outcome of the Core when ydesired= 0, y= 0 and error=0.

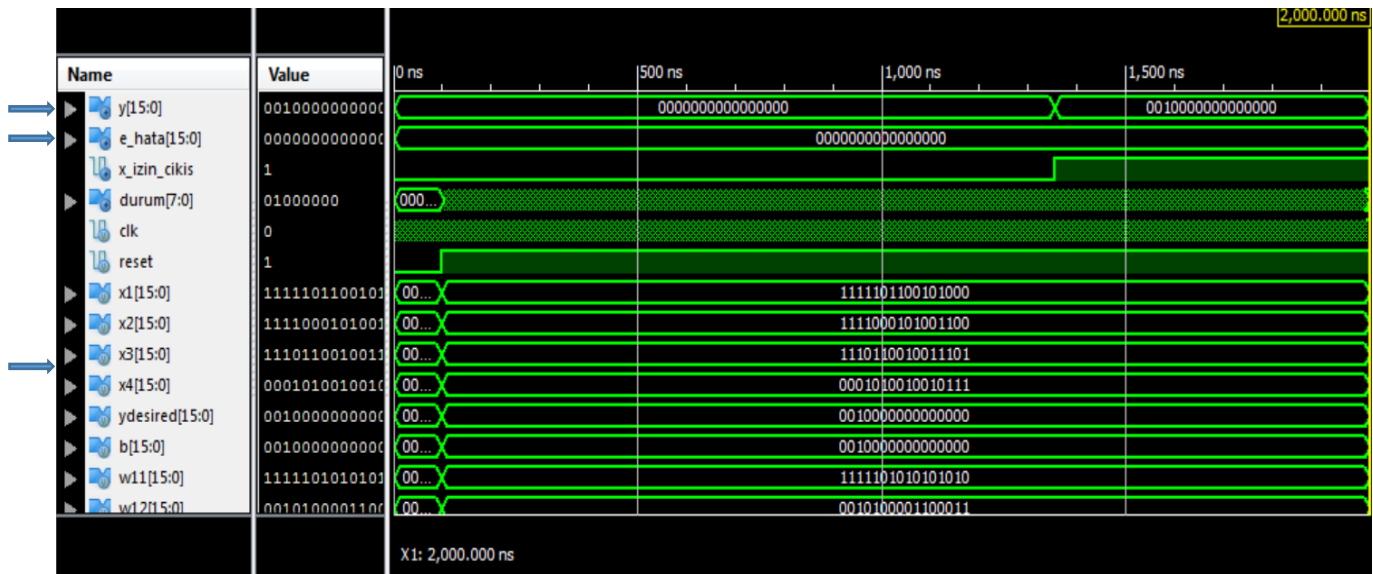


Fig. 5. Outcome of the Core when ydesired= 1, y= 1 and error=0.

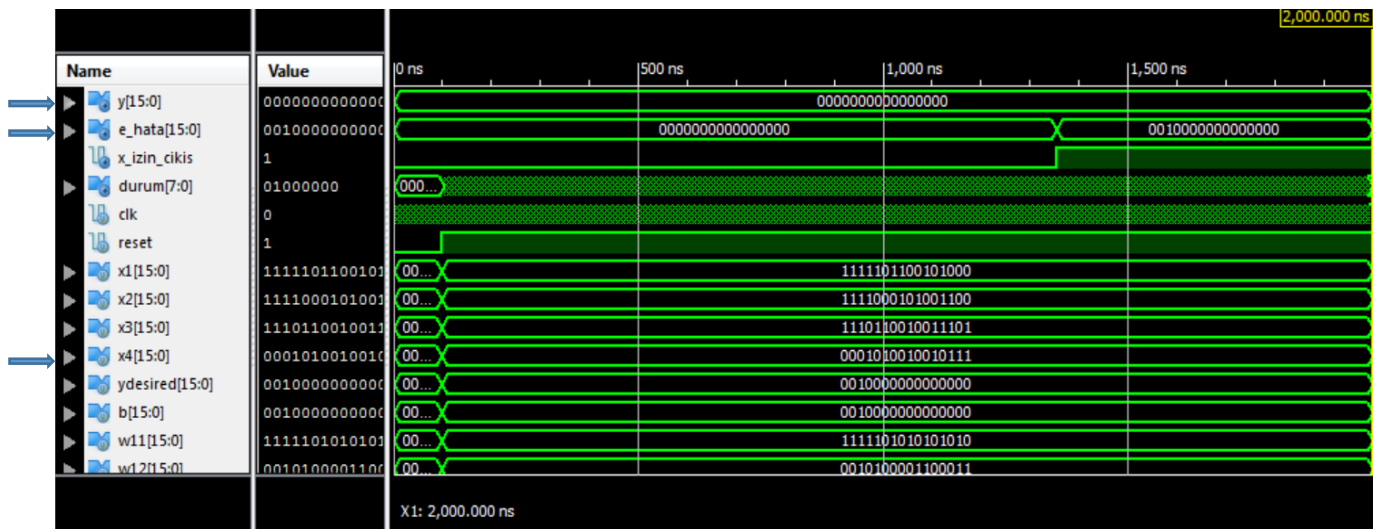


Fig. 6. Outcome of the Core when ydesired= 1, y= 0 and error=1

REFERENCES

- [1] J. Gotman, "Automatic recognition of epileptic seizures in the EEG", *Electroencephalography and Clinical Neurophysiology*, Vol. 54, No. 5, 1982, pp. 530-540.
- [2] D., Gür, T. Kaya, M. Türk, "Analysis of Normal and Epileptic EEG Signals with Filtering Methods", 2014 IEEE 22nd Signal Processing and Communications Applications Conference, SIU 2014, pp. 1877-1880.
- [3] S. Güzel, T. Kaya, H. Güler, "LabVIEW-Based Analysis of EEG Signals in Determination of Sleep Stages", 2015 IEEE 23rd Signal Processing and Communications Applications Conference, SIU 2015, 2015.
- [4] H. Adeli, S. Ghosh-Dastidar, N. Dadmehr, "A wavelet-chaos methodology for analysis of EEGs and EEG subbands to detect seizure and epilepsy", *IEEE Trans. Biomed. Eng.*, Vol. 54, 2007, pp. 205-211.
- [5] M.M. Shaker, "EEG Waves Classifier using Wavelet Transform and Fourier Transform", *Journal of Medical, Pharmaceutical Science and Engineering*, Vol. 1, No. 3, 2007.
- [6] T. Kaya, M.C. Ince, "Design of FIR Filter Using Modeled Window Function with Helping of Artificial Neural Networks", *Journal of The Faculty of Engineering and Architecture of Gazi University*, Vol. 27, No. 3, 2012, 599-606.
- [7] R. Schuyler, A. White, K. Staley, K.J. Cios, "Epileptic seizure detection", *IEEE Eng. Med. Biol. Mag.*, 2007, pp. 74-81.
- [8] A. Ersöz, S. Özşen, "Uyku EEG Sinyalinin Yapay Sinir Ağ Modeli ile Sınıflandırılması", *Elektrik Elektronik Bilgisayar Sempozyumu*, Elazığ, 2011.
- [9] S.M. Eka, M. Fajar, T. Iqbal, W. Jatmiko, I.M. Agus, "FNGLVQ FPGA design for sleep stages classification based on electrocardiogram signal", 2012 IEEE International Conference on Systems, Man, and Cybernetics (SMC), 2012, pp. 2711-2716.
- [10] A.G. Blaiech, K. Ben Khalifa, M. Boubaker, M.H. Bedoui, "Multi-width fixed-point coding based on reprogrammable hardware implementation of a multi-layer perceptron neural network for alertness classification", 2010 10th International Conference on Intelligent Systems Design and Applications (ISDA), 2010, pp. 610-614.
- [11] American Clinical Neurophysiology Society, "Guideline 8: guidelines for recording clinical EEG on digital media", *J Clin Neurophysiol*, Vol. 23, No. 2, 2006, pp. 122-124.
- [12] B. Karakaya, R. Yeniceri, M.E. Yalcin, "Wave computer core using fixed-point arithmetic", 2015 IEEE International Symposium on Circuits and Systems (ISCAS), 2015, pp. 1514-1517.

- [13] E. Özpolat, B. Karakaya, T. Kaya, A. Gülten, "FPGA-based digital Filter Design for Biomedical Signal", 2016 XII International Conference on Perspective Technologies and Methods in MEMs Design (MEMSTECH), 2016, pp. 70-73.

### BIOGRAPHIES



**BARIS KARAKAYA** was born in Elazig, Turkey 1990. He received the B.S. degree in electrical-electronics engineering from the Firat University in Elazig 2012 and M.S. degree in electronics and communication engineering from the Istanbul Technical University in Istanbul 2014. Since 2015, he is Ph.D. student at department of electrical-electronics engineering, Firat University, Elazig,

Turkey.

Since 2012, he is a Research Assistant at the Department of Electrical-Electronics Engineering, Faculty of Engineering, Firat University in Elazig, Turkey.

His research interests include embedded systems, FPGA programming, high-speed filtering processes and cryptography.



**TURGAY KAYA** was born in Elazig/Turkey, in 1982. He received the B.S., M.S. and Ph.D. degree in electrical-electronics engineering from the Firat University in 2003, 2006 and 2011, respectively. From 2004 to 20013, he was a Research Assistant at department of electrical-electronics engineering, Firat University, Elazig, Turkey. Since 2013,

he has been an Assistant Professor same department. His research interests include digital and analog filter design, biomedical signal processing, signal and image processing, micro-processing, artificial intelligence, heuristic optimization.



**ARIF GULTEN** was born in Elazig, Turkey 1970. He received the B.S. degree in electrical-electronics engineering from the Gazi University in Ankara 1993, M.S. degree from the Firat University in Elazig 1996 and PhD degree from the Firat University in Elazig 2003.

Since 2012, he is a Associate Professor at the Department of Electrical-Electronics Engineering, Faculty of Engineering, Firat University in Elazig, Turkey.

His research interests include Chaos, digital filters, bond graph and control systems.

# Client-Server Based Authentication Against MITM Attack via Fast Communication for IIoT Devices


M. Kara, and M. Furat


**Abstract**—Security is an important issue that should be taken care of by every system. In recent years, however, attackers are constantly developing themselves with new techniques to obtain personal information on the network. As systems evolve, the data that needs to be protected is increasingly appreciated and carries a higher risk of attack. As with many attacks such as MITM, there are solutions against this attacks. Nevertheless, these safety measures must be developed continuously. For this reason, we have developed a new system architecture with user-defined authentication against the intruders for the systems having large amount of data transmission rate. To maintain integrity of data, over a reliable system is that all incoming data are authenticated when data send to the server, on the other hand, in this system, the user-defined authentication can provide fast communication and it can decrease authentication time. The proposed system introduced in the present study checks for any changes in our instantaneous data. Moreover, we control the data integrity on simple devices such as sensors and motors or other industrial devices. Instead of using encryption, basically client-server based authentication system is used to avoid complex operations and protect the big data.

**Index Terms**—Client-server based authentication, IIoT, MITM, Security

## I. INTRODUCTION

TODAY'S Internet of Things (IoT) technology world, we are trying to store and qualify the data that millions of sensors have produced. This situation occurs big data. Big data comes from the combination of thousands of sensors creates security risks for large scale network. We must take measures to protect big data and take precautions for the safety of these critical system clearly.

**M. KARA**, is with Department of Electrical Department, Mustafa Kemal University, Hatay, Turkey, (e-mail: [mustafakara@mku.edu.tr](mailto:mustafakara@mku.edu.tr)) 

**M. FURAT**, is with Department of Electrical and Electronics Engineering Iskenderun University, Iskenderun, Turkey, (e-mail: [murat.furat@iste.edu.tr](mailto:murat.furat@iste.edu.tr)) 

Manuscript received August 11, 2017; accepted Nov 16, 2017.  
DOI: [10.17694/bajece.419546](https://doi.org/10.17694/bajece.419546)

Along with the developing IoT device technology such as wireless sensors [1], attacks on the network system are increasing in recent years. It is vital to detect these attacks and protect our network system, especially in the industrial fields. IIoT means Industrial Internet of Things (IIoT) in the industrial systems. If we examine them in detail, IIoT and IoT are absolutely different. IIoT technology includes networked smart power, factory and manufacturing. For this reason, IIoT devices require rapid communication between themselves and must be protected against harmful attack.

Intrusion Detection Systems (IDS) on a computer network is a first step of preventing the system to malicious use [2]. Sending to the main servers to analyze the information from various sensors is vital to large areas such as industrial areas. Therefore, intrusion detection systems have been developed. IDS may prevent many important damages for our system. IDS can be basically divided into two broad categories with respect to its architecture. [3]. Even then, a hybrid third one was developed by combining these two architectures. However, these architectures are basically similar in appearance to similar technologies in their application areas. Intrusion detection systems detect attacks made by instant analysis. The central server analyzes the events at different times with a base detection system.

We propose a new system architecture with user-defined authentication that simultaneously put on authentication and industrial field control at the same time. Besides we present a fast and secure client-server based system instead of encryption algorithms. In this study, the system architecture with user-defined authentications explained and the performance is discussed with popular encryption algorithms.

The IDS, which basically work with Host-based Intrusion Detection Systems (HIDS) concept, have turned into use Network-based Intrusion Detection Systems (NIDS) technology over time with different needs. HIDS are slightly different from the NIDS in that it is the technology in the protected computer. HIDS will not receive untrained traffic to the main computer under protection by itself. Instead, the HIDS tool monitors critical system file packages or files on the machine [4] and disconnects the network when there is an attack. It notifies a central management console. NIDS is deployed at strategic locations in the network system infrastructure (outside the firewall, especially in areas like the Demilitarized Zone, DMZ) to control traffic flow and compare known attack types against a database [5].

The intrusion detection systems, main server-based, are the focus of this article. In this paper, implemented with appropriate network connection is proposed to deal with intrusion detection problems using HIDS technology. Authentication system is designed to trigger instant attacks by performing key matching with specific functions. Thus, information obtained from instantaneous data generating devices such as sensors will be determined to have not undergone any changes during transmission.

The present paper is organized as follows. Section 2 describes the need for security, section 3 compares the methods of encrypting with the intrusion detection system, and explains why the IDS are preferred. Potential cyber threats are given in section 4 and the proposed system structure is presented in section 5. The conclusion can be found in the last section.

## II. BIG DATA NEEDS BIG SECURITY

It's no secret that encryption is large proponents of data security. Attacks such as network misuse on the security side, data modification, data theft and unauthorized access to IoT devices should be avoided by taking security precautions. Data security is the most important factor in network construction. Because of this, industrial companies are trying to provide data security by making a monetary investment in a large amount.

Large scale systems depend on computers or servers to control field devices. By the nature of computer systems, important amount of data that is controlled and processed is very important and sensitive. Because of this reason, big data requires protection against intruders. In the light of this information, Big Data needs Big Security. For example, one the most important device of a plant could be seized and a different value is sent to the authentication server by changing the temperature, humidity or pressure sensor values and this will be a problem for large fields. Security for data integrity is vital for cyber physical system [6].

## III. COMPARISON

### A. Intrusion Detection Systems

Intrusion Detection System [3, 4] is a network security system designed to detect security vulnerabilities against applications that are likely to be attacked or computer systems. An IDS technology is used to detect explicit attacks and is out of bandwidth in the networking system; It finds no real-time communication between the server and the client as illustrated in Fig.1. Obviously, the IDS technology to be described here is just a listening device. The IDS monitors the network traffic instantly and reports its results to a system administrator, but cannot automatically take action to prevent a detected exploit from taking over the system.

To summarize, the three main functions of the IDS include controlling (evaluating), examining (detecting), and reacting (reporting) the attacking eyebrows in software systems and networks [7].

Intrusion detection system can be divided into three categories with respect to their architectures [3, 7] as shown in Table 1.

### 1) Host-based Intrusion Detection System

The host-based intrusion detection system allows critical incidents to be seen in transmission systems. One can also detect and respond to malicious attacks or unusual movements discovered in the network system. It checks data integrity and traces the network system.

### 2) Network-based Intrusion Detection System

A network-based intrusion detection system monitors the network traffic to protect a system from threats and analyzes it according to the information in its hand [4]. It reads all packages and examines the packages which detect as dangerous. The system categorizes dangerous packages and notifies IT (Information Technology) staff. It can also block the packets according to IP address.

### 3) Distributed Intrusion Detection System

Distributed Intrusion Detection Systems (DIDS) over a huge network [8], all of which communicate with each other, or with a central server that simplify developed network monitoring, event analysis, and instant attack data [9].

TABLE I  
A COMPARISON OF DIFFERENT IDSS BASED ON THEIR ARCHITECTURE [7].

IDS Type	Deployment Location	Information source	Control domain
HIDS	Under-control system, software process	Local traffic (on OS level) and Log files	Local hosting system
NIDS	Isolated system on network traffic route, software process	Network traffic (raw data packets of the network)	Local segment or whole network
DIDS	Distributed and heterogeneous (host, network and central management system)	Host traffic and network traffic	Network wide (all hosts and different network segments)

### B. Encryption

Encryption is a method that transforms the information on the computer into an unintelligible form. Hence, even if someone can access a network system with important data [10], it will not be able to do anything unless it is some special software or the original data key.

The main function of the encryption is to convert the not only a normal text into an encrypted text but also a binary data into an encrypted binary data. Encryption helps to ensure that the data is not readable by the unauthorized people.

There are three different basic encryption methods that have different advantages for themselves. Their properties are tabulated in Table 2. It is clearly seen from Table 2, encryption methods notice integrity, authentication and non-reputation.

1) *Hashing Functions*

The way of summarization functions work is to show a shorter area by taking a long input. The goal is to reflect on the exit when there is a change in the ground. An encryption hash function [11] is a type of algorithm that can be applied on a piece of data, such as a single file or password, and produces a result called a checksum. The logic underlying the encryption hash function is to verify the authenticity of the data packet. For example, two files can be the same, but only if the checksums generated from each file use the same encryption burst function. Some of the summarization functions are MD5 (Message-Digest algorithm 5) and SHA1 (Secure Hash Algorithm 1).

2) *Symmetric Methods*

Symmetric methods care about integrity and authentication with a symmetric key. Essentially, symmetric algorithm means hash function with a key. Encrypt the whole data. AES (Advanced Encryption Standart) is a kind of symmetric key encryption [12].

3) *Asymmetric Methods*

Asymmetric encryption uses two keys for encryption or decryption. This method cares about integrity, authentication and non-reputation with the asymmetric key [13].

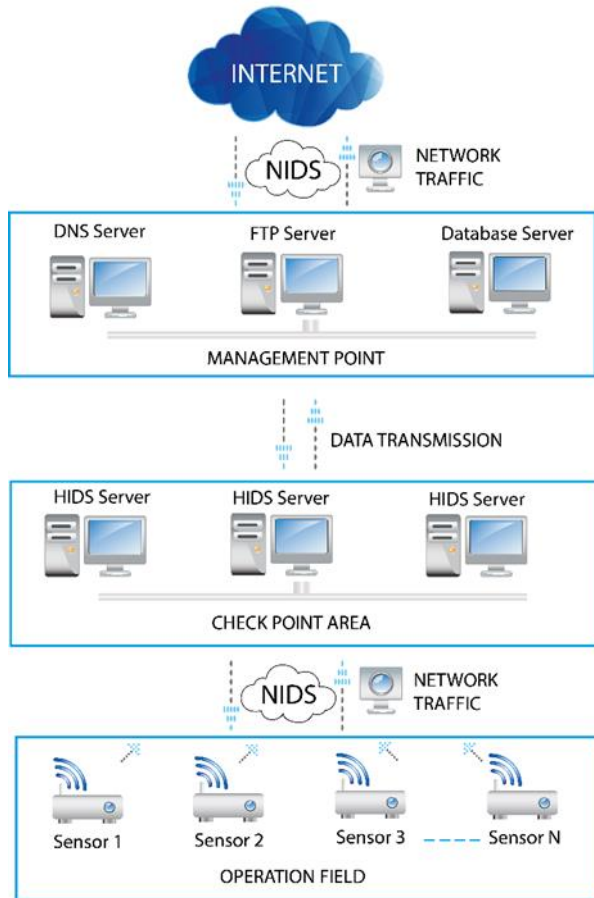


Fig 1. HIDS and NIDS are illustrated basically in the same picture in terms of how they work

TABLE II  
CRYPTOGRAPHIC TECHNICNS WITH KEY USAGE

Property	Hashing Functions	Symmetric Methods	Asymmetric Methods
Integrity	Yes	Yes	Yes
Authentication	No	Yes	Yes
Non-reputation	No	No	Yes
Key Usage	None	Symmetric Key(One key)	Assymmetric Key(two key)

C. *Reason for IDS Selection*

Some systems, particularly networks that care about data coming from sensors, want to analyze whether only incoming information has been changed. Thus, they achieve both a faster system and a lower cost. Such systems in an industry can need to control the accuracy of instantaneous transmission by considering data integrity directly rather than encryption. Because symmetric [10] cryptosystems involve significant communication problems. The secret key is forwarded to the receiving system before sending the actual data. The system with the Internet connection is 99% safe so there is no guarantee that the hacker will not attack. For this reason, the only safe way to change the keys is to exchange them personally. There is no need encryption in some systems which care about data integrity to make process go faster. In this paper, we developed a network system that allows verifying the identity of the sender. For this reason, we only care about data integrity during data transmission. We propose to select a user-defined authentication system in this network to determine if there is a change in the data package.

IV. CYBERSECURITY THREATS

The ever-evolving world of the internet is becoming the focus of hackers. As IoT technology evolves, attacks are increasing in species. In particular, aggressive tools such as botnets take advantage of the exploits in the system to capture the network. Security systems often catch many attackers on the network. However, imagine that an attacker has captured a system such as an industry or a large hospital. Let's consider a production site that works with very sensitive sensors. The risk of mortal accidents is very high when cameras and sensor devices are seized. As a real example, we have come to the conclusion that by the end of 2015, cyber-attacks [14] were a serious threat when Ukraine seized large portions of the electricity grid and proved to be dark in the middle of winter [15]. Currently, IoT technology brings a great deal of connectivity and convenience to modern day-like. However, the benefits created by IoT technology require manufacturers and users to be alert throughout the product life cycle. Protection against such danger must be provided by improved algorithms. Today, the types of attacks are increasing. Some attacks can get some information between two computers that are in direct communication. The type of attack we are handling in this study is the man-in-the-middle attack (MITM). The MITM



intercepts communication between the two network devices such as router, key server and switch etc. and this attack can change, modify or filter data [16].

The purpose of network security is to provide information security. This information security is based on 3 basic reasons: confidentiality, integrity and availability.

In some cases, the network's data may change because the data is gone through the intruder's computer. In this cases, we need to know if there is an attack for changing the data. Otherwise, we cannot secure communication. So, IoT technology can never be passed in industry or any other huge systems.

Such attack as man-in-the-middle attack on confidentiality and integrity or Denial of Service Attack [17] on availability [18]. Compromised credentials, cross-site scripting, man-in-the-middle attack, data breach, denial-of-service attack, malicious insiders, arp-poisoning, malware, ransomware and spear phishing etc. There are many kinds of attacks that can be performed by intruders for capturing. There is too much danger in the transmission phase from device to device. That's why, some attack detection algorithms should be developed to take precautions while transmitting the data.

Intrusion detection systems such as Firewalls, Network IDS, SNORT, Firestorm, Host IDS are developed to prevent the attacks [19]. However, these systems have the advantages of intrusion detection but have some disadvantages also. To explain some of them; Firewalls have legitimate user restriction, diminished performance, vulnerabilities, internal attack and high cost. For this reason, we use the firewall together with IDS to avoid many attacks. SNORT is an important system which rules define new attacks can easily be written and added. On the other hand, a kind of system which encrypted communications (VPNs, SSL, SSH, etc.) cannot be monitored, it cannot keep up with high volume traffic, network infrastructure requires change/editing and produce false alarms very intensely. Firestorm logs the data regularly every day using the administration console. But this brings overload to the system on the network. Our system removes the overload of computers. Because the network system is very simple. It just cares about data integrity and detects the intruders.

## V. CLIENT-SERVER BASED AUTHENTICATION SYSTEM

The most important purpose of our system is to evaluate this big data obtained by IIoT devices with sensors in the industrial field. As the amount of data increases, attackers are regularly trying to change this files and logges. For this reason, we propose a system via HIDS. This intrusion system is a technology that works like a central server and scans its own systems for activities. Typically, HIDS scans daily or weekly files in the operating system, application log files, or files in the database to find attack traces. For this reason, HIDS only works depending on the daily or weekly received file. As a result of this dependent situation, HIDS cannot detect the occurrence of an attack on the network by itself if the data of the weekly database files are bad or the information is changed by the attacker.

When controlling the data coming from the sensors, we must move with an advanced algorithm by controlling the server

entrance, taking filtering precautions, monitoring the events instantly and activating the warning system.

The result of the control scan performed by the host-based intrusion detection system is filed and logged securely and compared with logs to detect any malicious attempt. As a result of this situation, we propose a new system architecture with user-defined authentication. This is a kind of instant detection system for intruders.

### A. Proposed System Architecture

Reliability of data integrity always needs security via observation. For this reason, the system must be installed with some rules on network system with actuator and sensors as follows:

- Monitor and check access to the variables instantaneously.
- Abnormal sensor data should be detected and attempted attack should be prevented.
- Real-time intrusion must be detected and alerted.
- After the attack, check the system and analyze the output event.
- Give a report to IT staff after alarm against the possibility of changing big data environments.

The proposed client-server based authentication system in an industrial plant network is illustrated in Fig. 2. In this network system, large amount of data produced instantly by thousands of sensors in an industrial field are supposed. In the context of IIoT, owing to the instant transmission in the proposed system is carried out through smart devices, the emerged large data is naturally in the network environment. This system requires the data transmission between industrial devices to be controlled and secure. For this reason, we propose an architecture for this system consisting of 3 layers. These are Sensor and Actuator Layer, Transmission Layer and Administrator and Management Layer.

First, the data produced by the field sensors data is sent from Sensor and Actuator Layer to Transmission Layer. The data reaching the Transmission Layer is collected by the Sensor Unit. The Sensor Unit transmits this data to the Authentication Server. The Authentication Server updates the sensor data depending on the functions defined by the system user. For example, these functions can be adapted to be able to authenticate within an algorithmic rule. After this process, the data is sent to the Administration and Management Layer. Within the Administration and Management Layer, Management Points or IT Staff departments make the necessary assessments. The evaluated data is sent to the Authentication Client Server in the Transmission Layer again. The Authentication Client Server controls the authenticity of these data with respect to the previously defined authentication rules. After control by the Authentication Client Server, if there is no attack trace in the coming data, it sends these new inputs to the controller unit. The Actuator Unit sends these control inputs to actuators in the industrial field. As a result, the proposed system works in a secure structure.

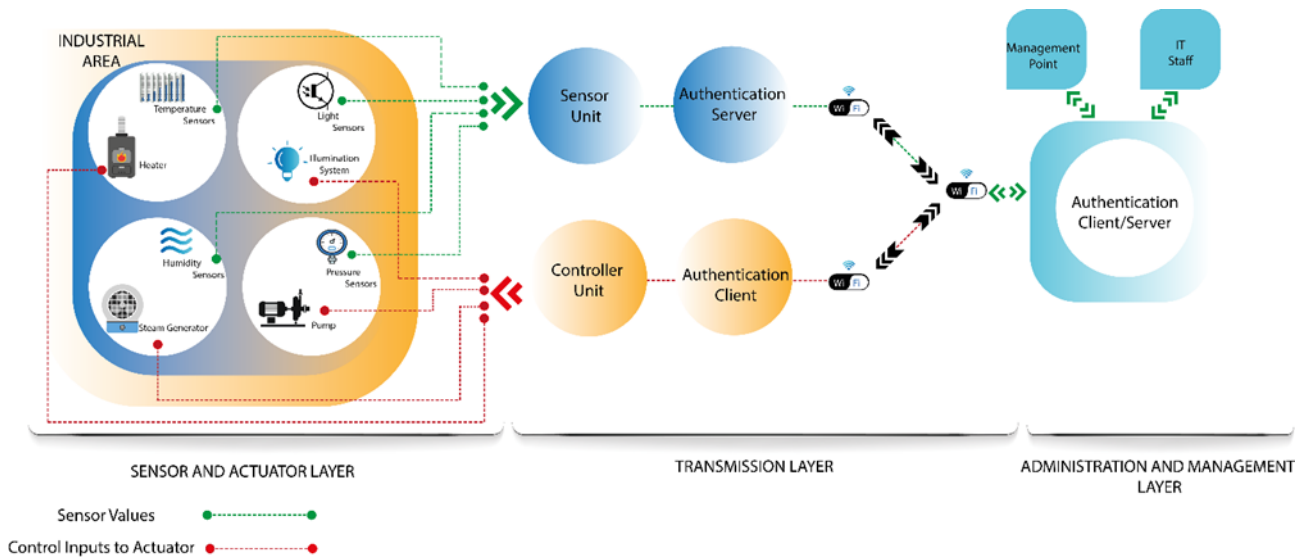


Fig.2. Proposed client-server based authentication system in an industrial plant network.

**B. Proposed System Algorithm**

The intrusion detection algorithm works within the proposed system can be summarized step by step as follows:

- Step 1:** The measurements of the sensors are collected in the sensor unit at each sampling time.
- Step 2:** The authentication server directly connected to the sensor unit adds a specific key to the sensors' data and sends the corresponding clients.
- Step 3:** The clients validates the data at first and then transmits the unit connected to itself.
- Step 4:** The tasks defined in Step 2 and 3 are valid for the Authentication Server/Client in Administration and Management Layer.
- Step 5:** If the validation is completed with success, then the data is proceeded.
- Step 6:** If the validation fails, then alarm is set to the IT staff and the data is logged.
- Step 7:** Stop

The flowchart of the algorithm is illustrated in Fig. 3.

**C. User Defined Authentication**

Authentication mechanism is one of the most effective ways to determination attacks. The proposed user-defined authentication technique has a lot of advantages over traditional encryption schemes.

First, the user-defined authentication procedures can be determined much simpler, and consequently. Then, it is much faster as compared to the encryption and decryption systems.

Second, the data traffic level can be reduced to very low levels in the industrial fields while sending data due to the authentication control based method here, together with the instantaneous comparison operations performed in the algorithm.

Third, the computation time is potentially reduced depending on the user-defined authentication function.

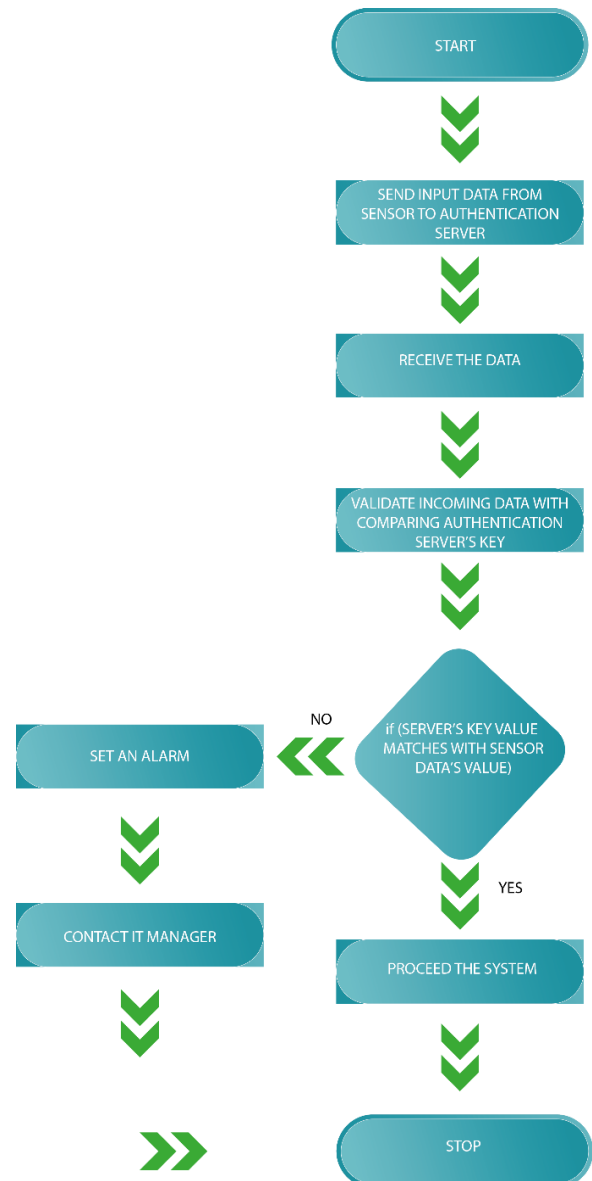


Fig 3. Flowchart of the proposed system algorithm.

## VI. CONCLUSION

In recent years, IT staff in industrial systems want to detect the changes from the sensors during the instantaneous data transmission process. Taking this into account, the proposed system architecture with user-defined authentication works to ensure the confidentiality, accessibility and security integrity of the data transmission system. Our presented system, especially for IIoT technology, detects the modification of the packet at the moment of data transmission. This system provides high performance and efficient technology when compared with existing encryption techniques. This is very important if large amount of data transmissions exists in a system. Transmitted and stored correct data directly effects the reliability of the large systems, i.e. IIoT based large networks.

## REFERENCES

- [1] K. Hewage, S. Raza, and T. Voigt, "An experimental study of attacks on the availability of glossy". *Computers & Electrical Engineering*, V.41, 2015, pp. 115-125.
- [2] B. Daya, "Network security: History importance and future", <http://web.mit.edu/~bdaya/www/Network%20Security.pdf>
- [3] S. Parveen and C. Sharma. "A Survey: Intelligent Intrusion Detection System in Computer Security", *International Journal of Computer Applications*, Vol.151, No.3, 2016, pp.18-22.
- [4] K. Nesreen, N. Hamdy and S. H. Ahmed, "A Proposed Intrusion Detection System for Encrypted Computer Networks", *Third International Conference on Informatics and Systems*, Giza, Egypt, 2005.
- [5] K. A. Varunkumar, M. Prabhakaran, A. Kaurav, S. S. Chakkaravarthy, S. Thiagarajan, P. Venkatesh. "Various Database Attacks and its Prevention Techniques", *International Journal of Engineering Trends and Technology*, Vol. 9, No. 11, 2014, pp. 532-536.
- [6] C. Shire. "Advanced mobile security in silicon". *Secure Mobile Communications Forum: Exploring the Technical Challenges in Secure GSM and WLAN*. The 2nd IEE, London, UK , 2004.
- [7] H. Jadidoleslami. "Weaknesses, Vulnerabilities and Elusion Strategies Against Intrusion Detection Systems", *International Journal of Computer Science and Engineering Survey*, Vol. 3, No. 4, 2012, pp. 15-25.
- [8] R. Robbins. "Distributed intrusion detection systems: An introduction and review", SANS Reading Room, GSEC Practical Assignment, 2002.
- [9] N. Einwechter. "An Introduction to Distributed Intrusion Detection Systems", 2002, <https://www.symantec.com/connect/articles/introduction-distributed-intrusion-detection-systems>
- [10] K. T. Nguyen, M. Laurent, and N. Oualha. "Survey on secure communication protocols for the Internet of Things", *Ad Hoc Networks* Vol. 32, 2015, pp. 17-31.
- [11] D. Wang, Y. Jiang, H. Song, F. He, M. Gu and J. Sun. "Verification of implementations of cryptographic hash functions", *IEEE Access*, V. 5, 2017, pp. 7816 - 7825.
- [12] Advantages and Disadvantages of Asymmetric and Symmetric Cryptosystems, [www.uobabylon.edu.iq/eprints/paper\\_1\\_2264\\_649.pdf](http://www.uobabylon.edu.iq/eprints/paper_1_2264_649.pdf)
- [13] M. Vigil, J. Buchmann, D. Cabarcas, C. Weinert and A. Wiesmaier. "Integrity, authenticity, non-repudiation, and proof of existence for long-term archiving: a survey". *Computers & Security*, Vol. 50, 2015, pp. 16-32.

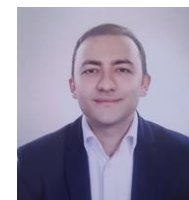
- [14] J. Jang-Jaccard and S. Nepal. "A survey of emerging threats in cybersecurity". *Journal of Computer and System Sciences*, Vol. 80, Issue. 5, 2014, pp. 973-993.
- [15] K. Zetter. "Inside the cunning, unprecedented hack of Ukraine's power grid", 2016, <https://www.wired.com/2016/03/inside-cunning-unprecedented-hack-ukraines-power-grid/>.
- [16] Thuc, N. D., Phu N.C., Bao T.N., Hai V.T. "A Software Solution for Defending Against Man-in-the-Middle Attacks on Wlan". *Department of Electronic Engineering and Information Sciences, RuhrUniversity Bochum, Germany*, 2015.
- [17] A. Mittal, A. K. Shrivastava and M. Manoria. "A review of DDoS attack and its countermeasures in TCP based networks", *International Journal of Computer Science & Engineering Survey (IJCSSES)*, Vol. 2, No. 4, 2011, pp. 177-187.
- [18] TAN, Shuaishuai; LI, Xiaoping; DONG, Qingkuan. TrustR. "An integrated router security framework for protecting computer networks". *IEEE Communications Letters*, 2016, 20.2: 376-379, 2016.
- [19] S. R. Borhade and S. A. Kahate. "Intrusion Detection System based on Hashing Technique". *Global Journal of Engineering Science and Researches*, Vol. 3, No. 6, 2016, pp. 31-34.

## BIOGRAPHIES

**MUSTAFA KARA** (1989) Mustafa Kara received the BSc in Computer Engineering (CE), Beykent University/Turkey in 2013. He started working as software engineering about computer and system security at some private companies about 3 years. After that, he started working as Lecturer at Mustafa Kemal University/Turkey in 2016 and started his



MSc İskenderun Technical University in 2016. He started his Ph.D at Hezarfen Aeronautics and Space Technologies Institute, National Defence University in 2018. His research interests are information, computer and system security, software engineering, industrial robots, electronic and mobile electronic signature and public key infrastructure, malware and spyware, personal and corporate information security and related fields.



**MURAT FURAT** (1977) received the B.Sc. in Electrical and Electronic Engineering (EEE), Gaziantep University/Turkey in 2002. He started to work as research assistant at Mustafa Kemal University/Turkey in 2002 and completed his M.Sc. at the same university in 2006. He completed his Ph.D at Çukurova University in 2014. He currently works as assistant professor doctor at Iskenderun Technical University. His research interests are process control, sliding-mode control, IoT, IIoT, metaheuristic algorithms and their applications to real systems.

# A Review of Turkish Sentiment Analysis and Opinion Mining

H. Karayığit, Ç. Acı and A. Akdağlı

**Abstract**—Social Media is one of the most frequently used platforms today. Users can easily share their views, ideas, and thoughts on this platform. The data shared on social media platforms is actually a great deal that can be transformed into meaningful information. The obtained big data can be analyzed and evaluated by various data analysis methods. Whether or not the data contain a feeling, if it is included; the type of the feeling (i.e. positive, negative or neutral) can be determined by emotion analysis methods. Sentiment Analysis studies in later times began to turn to analysis indicating different sentiments. Thus the foundations of Opinion Mining were laid. When ideas conveyed by social media information are presented semantically, they are expressed by Opinion Mining. The purpose of this paper is to explain the relationship between the concepts of Sentiment Analysis and Opinion Mining. The terms used in Sentiment Analysis and Opinion Mining are explained and examples of Turkish Sentiment Analysis are given. It has been tried to suggest solutions for the problems encountered in Turkish studies.

**Index Terms**— Turkish, Sentiment Analysis, social media, Opinion Mining.

## I. INTRODUCTION

**S**ENTIMENT analysis and Opinion Mining are the fields of study that analyzes people's views, evaluations, attitudes, and emotions from a written platform. Sentiment Analysis is used to find out what an existing text expresses emotionally. In some studies, the concept of Sentiment Analysis is also referred to by names such as thought mining, thought extraction. In industrial applications, Sentiment Analysis is more commonly used. Sentiment Analysis and Opinion Mining nomenclature are used in academic studies [1].

Opinion Mining is trying to show the meaning of the present text scientifically. Opinion Mining is the concept of Sentiment Analysis as a title [2]. The first study on Sentiment Analysis was conducted by Pang, Lee and Vaithyanatham in 2002 and movie comments existing in Internet Movie Database were taken and classified by various machine learning algorithms [3].

**H. KARAYIĞIT**, Department of Electrical and Electronics Engineering, Faculty of Engineering, Mersin University, Mersin, Turkey, (e-mail: [d2014242@mersin.edu.tr](mailto:d2014242@mersin.edu.tr)).

**Ç. ACI**, Department of Computer Engineering, Faculty of Engineering, Mersin University, Mersin, Turkey, (e-mail: [caci@mersin.edu.tr](mailto:caci@mersin.edu.tr)).

**A. AKDAĞLI**, Department of Electrical and Electronics Engineering, Faculty of Engineering, Mersin University, Mersin, Turkey, (e-mail: [akdagli@mersin.edu.tr](mailto:akdagli@mersin.edu.tr)).

Manuscript received August 19, 2017; accepted Nov 16, 2017.

DOI: [10.17694/bajece.419547](https://doi.org/10.17694/bajece.419547)

The first study in the concept of Sentiment Analysis and Opinion Mining are used together, semantic relations were established between emotional expressions and the subject rather than being classified as positive or negative [4].

In the first work under the name of Opinion Mining, a list of both product qualities (quality, characteristics) was created using the Opinion Mining tools, and opinions (weak, mixed, good) were collected about each one [5].

Although they look like different concepts, Sentiment Analysis, and Opinion Mining cannot be considered separately. Very similar methods are used while the significance of the dataset is more important for the Opinion Mining, the aim is to make emotions meaningful in Sentiment Analysis.

The remainder of the paper is organized as follows: In the first part of the second section, the concepts of Sentiment Analysis and Opinion Mining and their relation are mentioned. In the second phase of the second section, the Turkish Sentiment Analysis studies are briefly explained. In the conclusion section, evaluation was made for Sentiment Analysis and Opinion Mining relationship and it has been tried to suggest solutions for the problems encountered in studies conducted in Turkish. In the last section, references are included.

## II. TURKISH SENTIMENT ANALYSIS

### II. I. TURKISH SENTIMENT ANALYSIS AND OPINION MINING

Grammar rules vary for each language. Verb conjugation in a sentence differs between languages. The Turkish language is a structurally agglutinative language. Structural processing is more difficult than English. In Turkish, verb conjugation and lexical items are different from other languages as shown in Figure 1.

Gender discrimination in languages such as Arabic, English, German does not exist in Turkish. A sentence spoken by multiple words in English can be explained in Turkish by a word. For example; 'Gidemedik' is expressed in English 'We were not going'. There are 8 vowels (a, e, i, o, ö, u, ü) and 21 consonants (b, c, ç, d, f, g, ğ, h, j, k, l, m, n, p, r, s, ş, t, v, y, z) in Turkish. The seven letters are unique to the Turkish language alphabet (ç, ı, ş, ö, ü, ğ, İ).

When doing Sentiment Analysis in Turkish, you cannot use language models like Sentiment Analysis in English but all of

the methods used in the Sentiment Analysis literature can be used for the Turkish Sentiment Analysis.

TURKISH	ENGLISH
Okula gidiyorum.	I'm going to school.

Fig.1. Sentence structures in Turkish and English

In Sentiment Analysis studies positive, negative or neutral expressions are searched and analyzed in the dataset. The results of the analysis determine the attitudes of the persons or groups in the study. However, if the dataset is large, the classification of opinions cannot be done individually. The most commonly used method for Sentiment Analysis and Opinion Mining analysis is machine learning with text mining.

Preparing the data set for text classification is the most important step in the pre-processing step. In text mining, it is always necessary to deal with words containing noisy and insignificant information. Pre-processing step is made to solve the problems on the dataset, to be able to make more meaningful and quality analyzes by learning the natural structure of the dataset and to generate more meaningful information from the dataset [6].

Data Pre-processing Stage

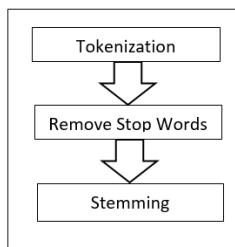


Fig.2. Data pre-processing stage

Pre-processing step is usually carried out in 3 steps as shown in Figure 2.

- Tokenization
- Stop-Word
- Stemming

In the next stage of the qualitative inference and selection process; add-subtract feature, creating a vector space model and attribute selection were made as shown Figure 3.

Feature Extraction

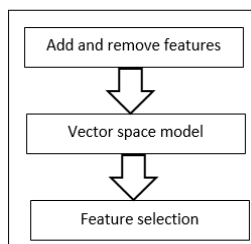


Fig.3. Feature Extraction and Feature Selection Stage

Classification methods are applied for data passing through the text mining stage. The concept of classification is simply to distribute data among the various classes defined on a data set. Classification algorithms learn this distribution form from given training set and then try to classify correctly when the test data arrives that the class is not specified.

The values that specify these classes on the dataset are given a label name and are used to determine the data class during the test, if necessary. Machine learning algorithms are used for classification [7].

Machine learning methods extract features from texts. Some of the methods used in Sentiment Analysis; Naive Bayes (NB), k-NN, Center Based Classifier, Artificial Neural Networks (ANN), Support Vector Machines (SVM).

There are two different approaches for Sentiment Analysis except text mining with machine learning approaches. These are natural language processing (NLP) methods and hybrid approaches.

Using the NLP methods, Sentiment Analysis is performed according to the results obtained after the analysis developed specifically for the language. The hybrid method is a hybrid approach and involves both methods.

When Sentiment Analysis studies go beyond the text-to-label linkage, the studies have turned into open-ended results. This is how the idea of Opinion Mining was formed.

As a result, both Sentiment Analysis and Opinion Mining are research topics for the following common materials;

- Classify personal opinions
- Sentiment classification of texts
- Summarizing thoughts in texts
- Inference of ideas from texts
- Classification of negative, sarcasm or irony-containing statements

## II. II. THE STUDIES ON TURKISH SENTIMENT ANALYSIS

Parlar [8] investigated various data preprocessing combinations in Sentiment Analysis of Turkish interpretations and investigated which qualification selection methods were effective results. Sentiment Analysis has proposed a new method of feature selection in selecting the most valuable features. It was tried to improve the accuracy and efficiency characteristics of the Sentiment Analysis process by using qualitative selection methods such as Chi-square, Information Gain, Document Frequency Difference and Optimal Orthogonal Centroid. The proposed method of selection of qualities is compared with these methods. In analyzing Turkish interpretations, it has been concluded that holding certain punctuation marks and ineffective words qualitatively contributes positively and contributes to better results with the quality selection methods used.

Evirgen [9] has proposed a general framework in terms of the fact that Turkish sentences, which express emotions in a

short and concise manner, are not properly formatted so that words can be directly processed and processed by R programming language, and that this is the starting point. It has been stated that there is not yet a robust and usable web or client application on Turkish Sentiment Analysis and that it is a starting point for one step forward in the study.

Kama [10] has implemented Turkish application for blogs and forums, comments and reviews collected by the collection of data created by previously developed for English.

Demirci [11] focused on emotional analysis on microblogs. Automatically detecting emotions in micro-blogs is a new research area which gains importance with the rapid growth of the micro-blogs in the last few years. Mining emotions in micro-blogs have some practical uses which can improve human-computer interaction. As opposed to regular text used in text mining studies, micro-blog entries are short and not well-formed enough to process directly. He has proposed a general framework that special uses, symbols, and facilities that can be used in micro-blogs can greatly affect the emotion of the text. He has created new Turkish Twitter dataset for Sentiment Analysis.

Çelik [12] developed A custom data preparation algorithm for the Turkish language in order to maximize the accuracy. He applied Naïve Bayes, Support Vector Machines, Logistic Regression and Decision Tree machine learning algorithms to the data sets. Naïve Bayes has worked 20 or 30 times faster than his nearest competitor, the Support Vector Machines algorithm. Besides the accuracy, the execution time of Naïve Bayes algorithm has been much faster than the others. Positive comments about everything have been 20-25 times more than negative comments. Accuracy has to decrease while n-gram size increases. Thresholding has generally little positive effect on accuracy but has improved execution time gradually on algorithms other than Naïve Bayes.

Boynukalın [13] has focused on emotion analysis of Turkish texts and it has been shown that using ML methods for Turkish texts on analysis of emotions is feasible and gives promising results. Several methods have been applied to get the best results with Turkish language and experimental results have been evaluated. She has defined a new set of data with using two different sources and this set was used in the study. Different automatic learning techniques have been tried and results are compared. Additions to the methods used due to the features that are separated between Turkish and English were made and the success of the analysis in Turkish texts has been tried to be increased.

Yelmen [14] focused on the selection of attributes from the Turkish texts written in daily conversation language in his work and used support vector machines, artificial neural networks, and centroid based classification algorithms on the detailed pre-processing data. SVM, ANN, and Centroid Based Classification Algorithms have been used on the data that is

processed through detailed pre-processing. Gini Index, Data Gain and Genetic Algorithm have been used as hybrids together with 3 different classification algorithms on Twitter messages belonging to followers of three different GSM operators.

Çoban [15] has studied two categorized sentimental analysis on Turkish messages from Twitter. Sentiment Analysis was considered as a text classification problem; Sentiment Analysis techniques, as well as classical text classification techniques, have been used. Machine learning methods are used to automatically detect the dominant sensation in Twitter messages. In this study, in which both text classification and emotion analysis experiments were performed, the main goal was to increase the success of emotional analysis. For this purpose, the effects of different preprocessing, labeling, classification and similarity methods on the Twitter emotional analysis in Turkish have been examined. In addition, a labeling method based on topical information was proposed and the highest success rate of 92.50% was achieved. Thus, the emotion analysis success can be achieved higher than previous studies.

Kaplan [16] has analyzed Twitter messages of Twitter users who are social media networks. The emotions expressed by shared Twitter messages are classified into four different classes. These classes are; 'Happiness', 'Anger', 'Sadness' and 'Surprise' groups. All typo mistakes of tweets were proofread with the help of 'Zemberek' library for classifying these accurately. Proofread tweets were labeled on these four categories of volunteers. In the study, messages collected from Twitter were analyzed with a decision tree and fuzzy learning techniques. Despite the fact that the fuzzy rules and decision tree methods have very close success rates, it has been found that both methods of emotion categorization give results that cannot give confidence in different categories. Nevertheless, it is suggested that the fuzzy decision method is a more valid method when closely examined in the categories.

Yurt [17] has tested the achievements of emotion analysis using pre-designed NLP algorithms in Turkish texts. Turkish texts were taken from the network environment, pre-processed, analyzed with tools that help in data mining and machine learning and the results are discussed. The study done using other machine learning algorithms has resulted in better results than Naïve Bayes.

Türkmenoğlu [18] has performed Two separate methods of sentimental analysis for Turkish, from machine learning methods and dictionary-based Sentiment Analysis, previously studied for English and Turkish. These methods were applied to two different sets of Turkish data, short and long texts, and their performance was measured. They have carried out preliminary operations considering the structural characteristics of the Turkish language.

Akba [19] has examined emotional analysis methods by

comparing success rates. According to some experimental results, he has tried to establish a system which can respond in a short time and needs less human power. Using data (Turkish movie comments) are interpreted and scored by the users on the website. Furthermore, in the evaluation of the experiment, it was tried to determine the number of the most appropriate words to be used in determining the feelings in the sentences written in Turkish. It was observed that the most successful results were obtained with 375 terms.

Nangir [20] has tried and succeeded in a multi-classifier approach which is an unexamined approach for the Turkish language. He has shown that the performance of the classifiers can be improved by parameter optimization operation and the accuracy of the classification operation can be increased. This success has shown that the multiple classifier approaches is a successful machine learning approach and can be used in many studies. The multiple classifier approaches have been used with three singular classifiers and a majority voting method.

### III. CONCLUSIONS

Turkish Sentiment Analysis and Opinion Mining are almost identical concepts. Sentiment Analysis, when looking for emotional words/expressions in a medium, the Opinion Mining outlines and analyzes the opinions of people about an entity.

While Sentiment Analysis is mostly used in industry, Opinion Mining is more frequently used in the academic field. Both concepts work together on a common research theme.

The following solutions are proposed for the problems observed in the Turkish emotional analysis thesis studies;

The lack of any open data sets in Turkish that can be used in the field of emotion analysis can be overcome by the data obtained from social media by various methods.

The difficulties related to the inadequacy of the studies done in the field of emotion analysis in Turkish language, Researchers can be overcome by providing project support from the economic side.

Although it is a Turkish WordNet database, the number of terms in this database is low and the result is that the opinion dictionary to be formed is composed of few terms. This problem can be solved if each researcher who works on emotion analysis increases the number of terms in the database.

In document-based classifications, there is a lack of work that has been reduced to sentence and sub-sentence structures and even subject-based classification. Studies on these issues can be increased.

Correct adjustment of the quality and categorical distribution of the training set has a direct impact on the analysis. The fact that the training data to be used for emotion analysis is sector-based will increase the success of the results of emotional analysis and quality extraction to be achieved.

### REFERENCES

- [1] B. Liu, *Sentiment Analysis and Opinion Mining*, Morgan & Claypool Publishers, 2012.
- [2] S.E. Şeker, "Duygu Analizi (Sentimental Analysis)", YBS Ansiklopedi, 2016, pp.21-36.
- [3] B. Pang, L. Lee, S. Vaithyanathan, "Thumbs up? Sentiment Classification Using Machine Learning Techniques", Proceedings of EMNLP 2002, pp. 79-86, 2002.
- [4] T. Nasukawa and J. Yi, "Sentiment Analysis: Capturing favorability using natural language processing", Proceedings of the 2nd international conference on Knowledge capture, Sanibel Island, FL, USA, 2003.
- [5] K. Dave, S. Lawrence, and M. P. David, "Mining the peanut gallery: Opinion extraction and semantic classification of product reviews", Proceedings of the 12th international conference on World Wide Web, ACM, 2003.
- [6] T. Kaşıkçı, H. Gökçen, "Metin Madenciliği İle E-Ticaret Sitelerinin Belirlenmesi, Bilişim Teknolojileri Dergisi", 7(1), 2014, pp. 25-32.
- [7] S.E. Şeker, *Sınıflandırma*, 08 December 2017. [Online]. Available: <http://bilgisayarkavramlari.sadievrenseker.com>
- [8] T. Parlar, *Feature Selection for Sentiment Analysis in Turkish Texts*, Published Ph.D. Thesis, Adana, 2016.
- [9] E. Evirgen, *Sentiment Analysis of Turkish Tweets*, Published Master's Thesis, İstanbul, 2016.
- [10] B. Kama, *Feature Based Sentiment Analysis on Informal Turkish Texts*, Published Master's Thesis, Ankara, 2016.
- [11] S. Demirci, *Sentiment Analysis on Turkish tweets*, Published Master's Thesis, Ankara, 2014.
- [12] H. Çelik, *Sentiment Analysis for Turkish Language*, Published Master's Thesis, İstanbul, 2013.
- [13] Z. Boynukalın, *Sentiment Analysis of Turkish Texts by Using Machine Learning Methods*, Published Master's Thesis, Ankara, 2012.
- [14] İ. Yelmen, *Sentiment Analysis on Turkish Social Media Data with Natural Language Processing Methods*, Published Master's Thesis, İstanbul, 2016.
- [15] Ö. Çoban, *Turkish Twitter Feelings Analysis with Text Classification Techniques*, Published Master's Thesis, Erzurum, 2016.
- [16] B. A. Kaplan, *Sentiment Analysis on Turkish Twitter Messages by Using Data Mining*, Published Master's Thesis, İstanbul, 2016.
- [17] E. A. Yurt, *Sentiment Analysis in Turkish Documents*, Published Master's Thesis, İstanbul, 2015.
- [18] C. TÜRKMENÖĞLU, *Sentiment Analysis in Turkish Texts*, Published Master's Thesis, İstanbul, 2015.
- [19] F. Akba, *Assesment of Feature Selection Metrics for Sentiment Analysis: Turkish Movie Reviews*, Published Master's Thesis, Ankara, 2014.
- [20] M. Nağır, *Classification of Emotions with Multiple Classifier Methods for Turkish Language*, Published Master's Thesis, İstanbul, 2013.

### BIOGRAPHIES



**HABİBE KARAYIĞIT** received the B.S. degree from the Electronic-Computer Technical Education department, Fırat University, Elazığ, Turkey, in 2002, and the M.S. degrees from the Electrical and Electronics Engineering department, Mustafa Kemal University, Hatay, Turkey in 2014. She is now a Ph.D.

student at the Mersin University. She is currently technical computer teacher and undergraduate student in computer engineering at the İSTE (İskenderun Technical University, İskenderun, Turkey). Her research interests Social Media Analysis, Text Mining, Machine Learning, Web Programming.



**ÇİĞDEM ACI** received her B.S. degree in Computer Engineering from Fırat University, Elazığ, in 2007. M.S. and Ph.D. degrees in Computer Engineering from the University of Çukurova, Adana, in 2009 and 2013. From 2008 to 2013, she was a Research Assistant with the Department of Computer Engineering at Çukurova University. Between 2013-2015 years, she worked as an Assistant Professor in the Computer Engineering Department, Çukurova University. Since 2015, she has been an Assistant Professor and Deputy Dean in the Faculty of Engineering, Mersin University. Her research areas are Machine Learning, Data Mining, and High-Performance Computing.



**ALİ AKDAĞLI** received his B.S. degree in Electrical and Electronics Engineering from Erciyes University, Kayseri, in 1995. M.S. and Ph.D. degrees in Electrical and Electronics Engineering from the University of Erciyes, Kayseri, in 1997 and 2002. From 2006 to 2008, he was Assistant Professor in the Department of Electrical and Electronics Engineering at Mersin University. Between 2008-2013 years, he worked as an Associate professor in the Electrical and Electronics Engineering Department, Mersin University. Since 2013, he has been a Professor and Dean in the Faculty of Engineering, Mersin University. His research areas are Embedded Systems, Wireless Communication Systems, Computer Networks and Security, Intelligent Antennas, Microwave Technology, RFID.



# Investigation Of Feature Selection Algorithms On A Cognitive Task Classification: A Comparison Study

S. G. Eraldemir, M.T. Arslan, and E. Yildirim

**Abstract**—In this study, the effects of feature selection on classification of the electrical signals generated in the brain during numerical and verbal operations are investigated. 18 healthy university/college students were chosen for the experimental study. EEG signals were recorded during silent reading and mental arithmetic operations without using any pen and paper. A total of 60 slides, 30 of which contained reading passages and the rest contained arithmetic operations, were presented in the experiment. EEG signals recorded from 26 channels during the slide show. The recorded EEG signals were analyzed by Hilbert Huang Transform (HHT), and then features were extracted. 312 features were classified by Bayesian Network algorithm without applying feature selection with 92.60% average accuracy. Consistency measures and Correlation based Feature Selection methods were, then, used for feature selection and the numbers of selected features are 8 and 39 on average, respectively. Classification accuracies by using these feature selection algorithms were obtained as 93.98% and 95.58%, respectively. The results showed that feature selection algorithms contribute positively to the classification performance.

**Index Terms**— Hilbert Huang Transform, Consistency Measures, Correlation based Feature Selection, EEG Classification.

## I. INTRODUCTION

Brain-computer interfaces (BCI) are systems that allow to communicate the brain with external devices via a computer. These systems are based on the principle of real-time control of various systems to ease the lives of people who have suffered cognitive, sensory and motor functions such as paralysis or muscular dysfunction [1, 2, 3, 4]. EEG based BCI systems are used intensively for reasons such as high time resolution, low cost, and portability. Recorded EEG signals need to be analyzed, classified and transmitted to the system very quickly [5, 6]. In BCI systems, in general, the following steps are followed: Recording EEG signals, pre-processing, feature extracting and classifying. Finally, In the BCI system, a command is sent to the relevant system considering the classification results.

**S. G. ERALDEMIR**, is with Dep. of Computer Programing, Iskenderun Technical University, Hatay, Turkey, (e-mail: sgoksel.eraldemir@iste.edu.tr)

**M. T. ARSLAN**, is with Department of Computer Technology, Mustafa Kemal University, Hatay, Turkey, (e-mail: [mtarslan@mku.edu.tr](mailto:mtarslan@mku.edu.tr))

**E. YILDIRIM**, is with Department of Electrical-Electronics Engineering, Adana Science and Technology University, Adana, Turkey, (e-mail: [esenyildirim@gmail.com](mailto:esenyildirim@gmail.com))

Manuscript received August 15, 2017; accepted Nov 16, 2017.

DOI: [10.17694/bajece.419549](https://doi.org/10.17694/bajece.419549)

In the preprocessing process, the EEG signals are cleaned out noises resulting from the network or subject by various methods. Feature extraction process is an important field which many operations are performed on the signal processing. The most important point in signal processing is whether the signal is stationary or not. Fourier transform is a suitable signal processing method if the signal is a stationary signal whose frequency value does not change by time. Wavelet Transform and Hilbert Huang Transform (HHT) are shown to be more efficient methods to analyze non-stationary signals [7,8]. The Hilbert Huang Transformation, which is preferred in this study, has been used in many studies in the analysis of EEG signals in the literature [9,10,11,12,].

R. Wang et al. showed that HHT gives better results than Fast Fourier Transform and Continuous Wavelet Transform (CWT) in sleepy EEG signals [10]. While sleeping and waking, K. Rai et al. analyzed EEG signals using HHT based features and then they have classified EEG signals by Fuzzy Logic method [11]. Swarnalatha. R. and Prasad D.V in 2015 analyzed the recorded EEG data using HHT method for early diagnosis of bruxism disease, called “interdental rubbing disorder” [12]. In another study, J. Kortelainen et al. have developed a novel approach to early diagnosis of hypoxic ischemic encephalopathy (HIE), permanent illness in infants who do not have adequate blood in the brain, using HHD [9].

In BCI systems, features obtained by various signal processing methods can be used directly for classification process. Furthermore, it is possible to eliminate some of those features that may adversely affect the classification result and thus improve classifier performance. Feature selection methods reduced features without changing the signal's characteristic. Feature selection eliminates irrelevant and unnecessary features and allows selection of features that will provide the actual contribution in the classification process so that both training and test times are shortened during classification. Thus, problems that can arise from the delays in real-time BCI systems are reduced. Feature selection techniques have been used many studies in classification of EEG signals [13, 14, 15, 16, 17]. Feature selection approaches are known as filter, wrapper and embedded systems. Filter-based methods provide fast results because they perform selection based on statistical data in the analysis of large data groups. Wrapper methods choose features that performs well for the chosen classifier. Embedded systems combine the operating logic of the two systems mentioned above.

In this study, we used correlation-based feature selection (CFS) and consistency measures, which are filter based feature selection methods which achieve faster results specifically for high dimensional data.

Ji et al. [14] analyzed EEG signals recorded during sleep from 16 volunteers. In the study, they acquired features by applying the Fourier Transform and the short-time Fourier Transform to the EEG data for the detection of different states of sleeping. Afterwards, they carry out CFS and the genetic algorithm-based feature selection methods to reduce EEG features. As a result, the best results were found by CFS. Hu et al. [15] collected EEG signals to investigate the effects of attention on distance education methods. They reported that 63.90% was achieved for EEG data without feature selection while 80.84% was achieved by applying CFS for feature selection [15]. Onan A. and Korukoglu S. preferred filter based feature selection and wrapper based feature selection methods to reduce the number of features in text classification [17]. The highest result was yielded by filter based feature selection method with a classification accuracy of %89.72 [17].

In this experimental study, EEG data collected from 18 subjects were used, and then the signals were analyzed by HHT. Then HHT based features are classified by a probability based approach, Bayesian Network algorithm. To reduce the cost of operation and achieve better classification rates, consistency measures and correlation-based feature selection methods are applied and the classification performances are compared to the results obtained without feature selection.

## II. MATERIALS AND METHOD

EEG data were obtained from 18 volunteer all-male university/college students. EEG data were collected from 32 electrodes with a sampling frequency of 1000Hz. Feature extraction was applied on EEG data by means of HHT, and then consistency-measures and CFS were applied for feature selection.

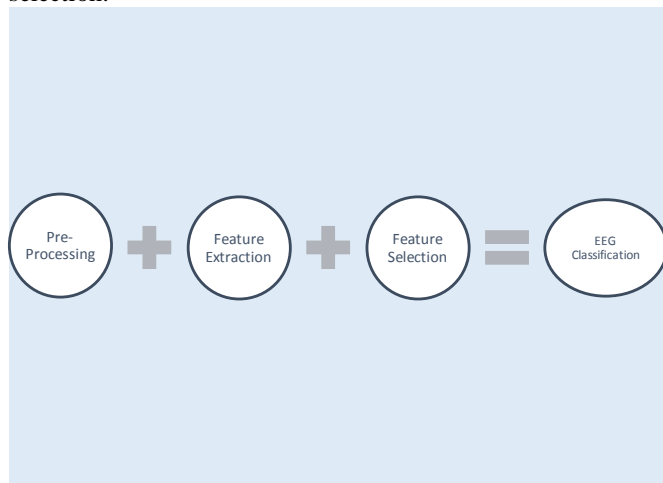


Fig. 1 The approach followed in the study

### A. Collection of EEG data

The EEG markers were collected from volunteers using electrodes placed on the scalp as shown in Figure 2 in accordance with the international 10-20 system.

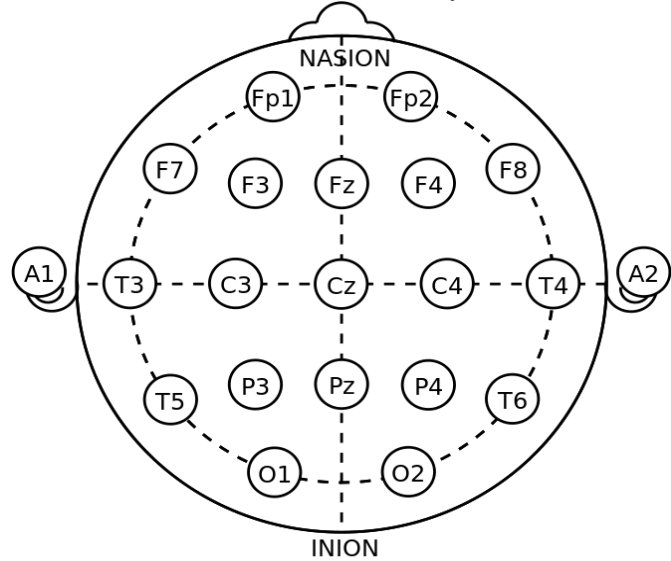


Fig. 2. The position of the electrodes relative to the 10-20 system for EEG recording (Top View)

The important considerations before and during the recording of EEG data are expressed as follows:

- Subjects are warned about their hair being clean and short, and not to use any hair styling products.
- Subjects were told not to take any medication prior to EEG recording.
- It has been stated that subjects should focus only on numerical slides or verbal slides on the screen.
- Subjects were told not to move their body parts such as hands, arms, head, legs and eyes during recording as shown in Figure 3. In addition, they were initially seated in a comfortable position.

A total of 60 slides, 30 numeric and 30 verbal texts, were shown during the recording. The duration of each slide is set to 13.25 seconds.

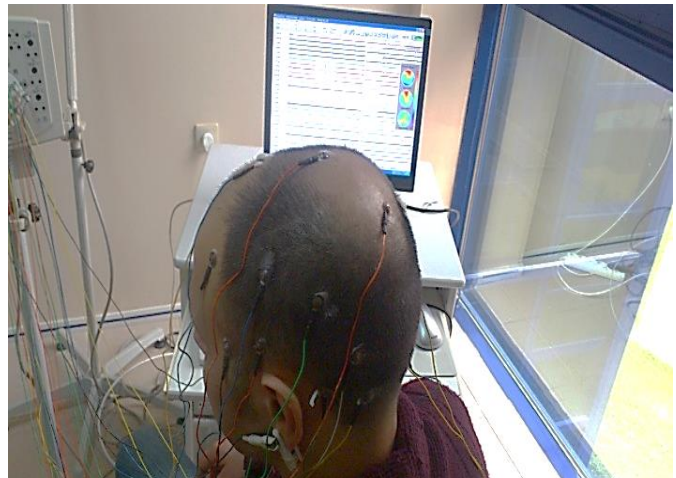


Fig. 3 EEG recording environment

Figure 4 is an example of a slide containing numerical operations.

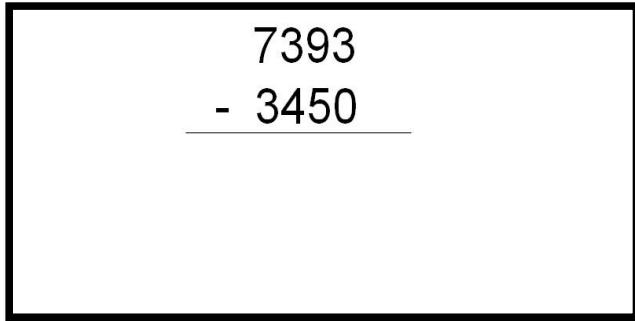


Fig. 4. A Numerical Slide Example

Figure 5 shows an example of slides containing verbal text.

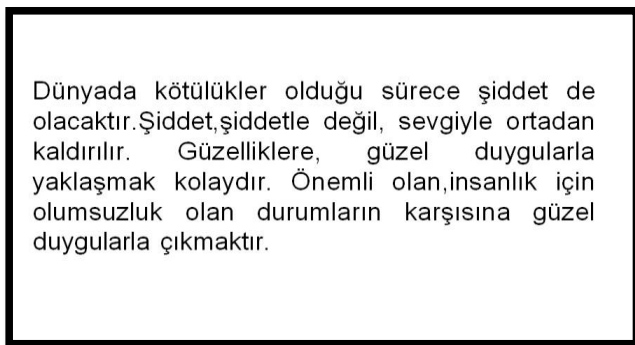


Fig 5. A Text Slide Example

### B. Pre-Process

During recording, 50 Hz noise originating from the network is automatically cleaned by the device. After the recordings, signals are bandpass-filtered between 0.1 and 120 Hz and are divided into 60 pieces, each of which is a 13.25 second long EEG segment. Afterwards, the last two slides and the first slide in each section deactivated to prevent any problems due to synchronization. For this reason, 27 verbal and numerical slides were used for experimental study on EEG recordings.

### C. Feature Extraction

Features are extracted from pre-processed EEG signals using HHT. HHT is an adaptive method that allows the analysis of non-linear, non-stationary signals as well as signals whose frequency and amplitude change by time [8]. By means of this method, complex, non-linear and non-stationary signals such as especially EEG signals, can be better analyzed than other signal processing methods. HHT is an empirical and adaptive method which combines Empirical Mode Decomposition (EMD) and Hilbert Transform. Hilbert Transform can be applied directly to single-component signals since these signals have only one frequency content at a given time interval. However, this is not possible for multi-component signals such as EEG. In order to solve this problem, Huang first divides the signal into the Intrinsic Mode Functions (IMFs) to include only one frequency value at a certain time [8]. Thereafter, the signal is expressed as the sum of these functions, each of which has a single frequency content instantaneously. In Empirical Mode Decomposition (EMD), the signal is divided into IMFs. In the last stage of the

HHT, The Hilbert transform is applied to the obtained IMFs to generate the energy-frequency-time distribution known as the Hilbert spectrum.

EEG signals were analyzed with a sliding window method where each window is one second long with 50% overlap. For each segment, 312 features were extracted by calculating the average amplitude and maximum amplitude values for delta (0.5-4Hz), theta (4-8 Hz), alpha (8-13 Hz), beta (13-30 Hz) and gamma (31-120 Hz) bands apart from the average amplitude and the maximum amplitude in the whole spectrum.

### D. Feature Selection

Features were selected by applying Consistency Measures and Correlation-based Feature Selection (CFS) methods which are filter methods to the feature matrix extracted using HHT. In the analysis of large data, these methods have been used because of the reduction of training and test time.

#### Consistency Measures

This method was used in many studies in the literature for feature selection [18, 19, 20]. Consistency Measures (CM) is a filter based method that considers consistency levels of class values to examine the usefulness of the subset of attributes [21].

The CM method begins with a subset of all the attributes. Then, a random subset of attributes is generated from a subset space of attributes. If the randomly generated subset of attributes contains attributes less than or equal to the current subset of attributes, then the consistency ratings of the existing and newly created subset of attributes are compared. If the newly created subset of attributes has a better consistency rate, this subset is selected. The process is repeated during a parameter set by the user.

TABLE I.  
NUMBER OF SELECTED FEATURES

No of Subject	Total Number of Features	Correlation based Feature Selection	Consistency Measures
Subject1	312	42	8
Subject 2	312	46	9
Subject 3	312	38	9
Subject 4	312	35	9
Subject 5	312	31	5
Subject 6	312	31	8
Subject 7	312	37	6
Subject 8	312	36	11
Subject 9	312	48	8
Subject 10	312	43	7
Subject 11	312	55	8
Subject 12	312	51	7
Subject 13	312	41	6
Subject 14	312	39	8
Subject 15	312	44	6
Subject 16	312	28	6
Subject 17	312	32	8
Subject18	312	32	6
<b>Average</b>	<b>312</b>	<b>39</b>	<b>8</b>

#### Correlation based Feature Selection

This method creates a new subset of attributes by selecting the attributes that have the highest correlation with the class within the set of attributes extracted from the signal. A subset of the

attributes which has the highest correlation with class but the lowest correlation each other is generated [22, 23, 24].

The number of selected attributes is shown in Table I for each Subject. When Table I is analyzed, it is seen that a very small number of attributes are selected in different numbers for each subject.

### III. RESULTS AND DISCUSSION

In this study, we reviewed the effect of feature selection algorithms such as CFS and CM on the experimental result, considering that HHT based features show better classification performance. The features were extracted by using HHT from EEG signals and were binary classified as arithmetical operation / silent reading.

The performance of Bayesian Network was worked out using the most commonly used parameters such as Accuracy, Precision and area under the ROC Curve (AUC) [25].

Examining Table II, it is seen obviously that arithmetical and reading operations were classified with an average accuracy of 92.60% and an average precision of 92.95%, without applying feature selection. Although the highest accuracy was 97.40%, the lowest accuracy is found as 85.50%. It is observed that there are a total of 13 subjects whose accuracy is 90% or better.

Table II  
THE CLASSIFICATION RESULTS WITH ALL FEATURES

No of Subject	Accuracy	Precision	AUC
Subject 1	0.9100	0.9100	0.9630
<b>Subject2</b>	<b>0.8550</b>	<b>0.8630</b>	<b>0.9370</b>
Subject3	0.8590	0.8610	0.9380
Subject 4	0.8990	0.9040	0.9710
Subject 5	0.9670	0.9670	0.9910
Subject 6	0.9590	0.9590	0.9940
<b>Subject7</b>	<b>0.9740</b>	<b>0.9740</b>	<b>0.9920</b>
Subject 8	0.9010	0.9070	0.9380
Subject 9	0.9440	0.9470	0.9780
Subject 10	0.9490	0.9500	0.9880
Subject 11	0.9390	0.9400	0.9810
Subject 12	0.8870	0.8940	0.9650
Subject 13	0.9700	0.9700	0.9970
Subject 14	0.8770	0.8990	0.9880
Subject 15	0.9700	0.9710	0.9980
Subject 16	0.9630	0.9630	0.9920
Subject 17	0.9340	0.9380	0.9830
Subject 18	0.9110	0.9140	0.9700
<b>Average</b>	<b>0.9260</b>	<b>0.9295</b>	<b>0.9758</b>

The results while applying consistency measure are given in Table III. The average accuracy and precision were found as 93.98% and 94.08%, respectively with an average number of 8 features. Comparing the results, the features selected by consistency is more successful than all features in classification. The lowest classification accuracy was 86.60% and the lowest precision was 86.90% as well as the highest classification accuracy and precision were 97.90%. It is also seen that a total of 16 subjects are classified, with accuracy of 90% and above.

Table IV shows the classification results achieved when applied correlation based feature selection.

Table III.  
THE CLASSIFICATION RESULTS USING CONSISTENCY MEASURES

No of Subject	Accuracy	Precision	AUC
Subject1	0.9300	0.9300	0.9800
<b>Subject2</b>	<b>0.8660</b>	<b>0.8690</b>	<b>0.9500</b>
Subject3	0.8770	0.8770	0.9540
Subject4	0.9410	0.9420	0.9890
Subject5	0.9720	0.9720	0.9970
Subject6	0.9300	0.9300	0.9860
<b>Subject7</b>	<b>0.9790</b>	<b>0.9790</b>	<b>0.9980</b>
Subject8	0.9100	0.9140	0.9730
Subject9	0.9400	0.9410	0.9830
Subject10	0.9630	0.9630	0.9960
Subject11	0.9460	0.9460	0.9860
Subject12	0.9010	0.9040	0.9740
Subject13	0.9700	0.9700	0.9960
Subject14	0.9450	0.9470	0.9910
Subject15	0.9720	0.9730	0.9970
Subject16	0.9580	0.9580	0.9930
Subject17	0.9390	0.9420	0.9910
Subject18	0.9770	0.9770	0.9950
<b>Average</b>	<b>0.9398</b>	<b>0.9408</b>	<b>0.9849</b>

Table IV.  
THE CLASSIFICATION RESULTS USING CORRELATION BASED FEATURE SELECTION

No of Subject	Accuracy	Precision	AUC
Subject1	0.9390	0.9400	0.9850
Subject2	0.8900	0.8930	0.9740
<b>Subject3</b>	<b>0.8670</b>	<b>0.8680</b>	<b>0.9510</b>
Subject4	0.9680	0.9680	0.9970
<b>Subject5</b>	<b>0.9860</b>	<b>0.9860</b>	<b>0.9990</b>
Subject6	0.9570	0.9570	0.9950
Subject7	0.9810	0.9820	0.9990
Subject8	0.9120	0.9190	0.9890
Subject9	0.9590	0.9610	0.9940
Subject10	0.9830	0.9830	0.9980
Subject11	0.9760	0.9760	0.9910
Subject12	0.9360	0.9380	0.9890
Subject13	0.9840	0.9850	0.9990
Subject14	0.9630	0.9630	0.9950
Subject15	0.9810	0.9820	0.9990
Subject16	0.9820	0.9820	0.9990
Subject17	0.9620	0.9640	0.9970
Subject18	0.9790	0.9790	0.9960
<b>Average</b>	<b>0.9558</b>	<b>0.9570</b>	<b>0.9914</b>

The average accuracy and precision were found as 95.58% and 95.70% as shown in Table IV. The results were found to be higher than the result gained using all the features. In addition, it is demonstrated that the poorest accuracy and precision values were 86.70% and 86.80%, respectively for Subject3 even though Subject5 had the highest precision and the accuracy value were 98.60%. In general, it is clearly seen that the findings in Table IV are more successful than those of Table II and Table III.

Examining the all results, for BCI systems, it is indicated that higher classification performance can be achieved by reducing the number of features with correlation-based feature selection. As a result of the all results, it is understood that the correlation-based feature selection method is more compatible with the

Bayesian Network algorithm for the features from database were used in this study.

The results show the importance of selecting features for the development of faster and higher performance BCI systems and the method to be selected for this process.

#### IV. CONCLUSION

In this paper, we presented an approach to examine the effect of feature selection algorithms on cognitive tasks based on EEG signals.

The features were extracted by using Hilbert Huang Transform from EEG signals and were binary classified as arithmetical operation / silent reading. For feature selection, we used Correlation based Feature Selection and Consistency Measures algorithm which are filter methods. Bayesian Network was employed for classification. Effect of feature selection algorithms are evaluated for this cognitive task analysis. The classification results indicated that Feature Selection algorithms have a positive effect on EEG signal classification performance. CFS and CM feature selection algorithms are a powerful and useful tools to select EEG features.

#### REFERENCES

- [1]. T. Aflalo. "Decoding motor imagery from the posterior parietal cortex of a tetraplegic human." *Science*. vol. 348. no. 6237. pp. 906–910. May 2015.
- [2]. E. Demirci, "Playing games with brain waves". *TUBITAK Science Technical Journal*. 44 (520). 18-24p, 2011
- [3]. L. R. Hochberg. "Reach and grasp by people with tetraplegia using a neurally controlled robotic arm." *Nature*. vol. 485. no. 7398. pp. 372–375. 2012.
- [4]. F. Cabestaing, T. M. Vaughan, D.J. Mcfarland, J. R. Wolpaw, "Classification of evoked potentials by Pearson's correlation in a Brain-Computer Interface". *Matrix*. 67. 156-166pp, 2017.
- [5]. A. Dogan, M.H. Calp, E.S. ARI, H. Ozkose. "An examination on Brain Computers Interfaces in Human Computer Interaction: Characteristics and Working Principle"
- [6]. E. Sevinç. "Brain Computer Interfaces", [http://www.rehabilitasyon.com/action/makale/1/Beyin\\_Bilgisayar\\_Arayuzleri-2299](http://www.rehabilitasyon.com/action/makale/1/Beyin_Bilgisayar_Arayuzleri-2299) (2006)
- [7]. O. Rioul, and M. Vetterli. "Wavelets and signal processing." *IEEE signal processing magazine* 8.4 (1991): 14-38.
- [8]. N. E. Huang et al. "The empirical mode decomposition and the Hilbert spectrum for nonlinear and non-stationary time series analysis." *Proc. R. Soc. A Math. Phys. Eng. Sci.*. vol. 454. no. 1971. pp. 903–995. Mar. 1998.
- [9]. J. Kortelainen, E. Vayrynen, U. Huuskonen, J. Laurila, J. Koskenkari, J.T. Backman, S. Alahuhta, "Using Hilbert-Huang Transform to assess EEG slow wave activity during anesthesia in post-cardiac arrest patients.", 2016 38th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC). Orlando. FL. 2016. pp. 1850-1853.
- [10]. R. Wang, Y. Wang and C. Luo. "EEG-Based Real-Time Drowsiness Detection Using Hilbert-Huang Transform." 7th International Conference on Intelligent Human-Machine Systems and Cybernetics. Hangzhou. 2015. pp. 195-198.
- [11]. K. Rai, V. Bajaj and A. Kumar. "Hilbert-Huang transform based classification of sleep and wake EEG signals using fuzzy c-means algorithm." *International Conference on Communications and Signal Processing (ICCSP)*. Melmaruvathur. 2015. pp. 0460-0464.
- [12]. R. Swarnalatha, and D. V. Prasad, "Detection of Sleep Bruxism Based on EEG Hilbert Huang Transform". 5th International Conference on Biomedical Engineering and Technology (ICBET 2015). IPCBEE vol.81 (2015) (2015) IACSIT Press. Singapore
- [13]. R. Jenke, A. Peer and M. Buss. "Feature Extraction and Selection for Emotion Recognition from EEG." in *IEEE Transactions on Affective Computing*. vol. 5. no. 3. pp. 327-339. July-Sept. 1 2014.
- [14]. Y. Ji, X. Bu, J. Sun and Z. Liu, "An improved simulated annealing genetic algorithm of EEG feature selection in sleep stage." 2016 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA), Jeju, 2016, pp. 1-4.
- [15]. B. Hu, X. Li, S. Sun and M. Ratcliffe, "Attention Recognition in EEG-Based Affective Learning Research Using CFS+KNN Algorithm." in *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, vol. PP, no. 99
- [16]. C. Cimpanu, L. Ferariu, T. Dumitriu and F. Ungureanu. "Multi-Objective Optimization of Feature Selection procedure for EEG signals classification." 2017 E-Health and Bioengineering Conference (EHB). Sinaia. 2017. pp. 434-437.
- [17]. A. Onan, S. Korukoglu. "Evaluation of feature selection methods in text classification", *Academic Knowledge* 2016, Turkey (2016).
- [18]. S. Zhang, Z. Zhao. "Feature Selection Filtering Methods for Emotion Recognition in Chinese Speech Signal". 2008
- [19]. L. Jalali, M. Nasiri and B. Minaei. "A hybrid feature selection method based on fuzzy feature selection and consistency measures." 2009 IEEE International Conference on Intelligent Computing and Intelligent Systems. Shanghai. 2009. pp. 718-722.
- [20]. B. Chakraborty and G. Chakraborty. "Fuzzy Consistency Measure with Particle Swarm Optimization for Feature Selection." 2013 IEEE International Conference on Systems, Man, and Cybernetics. Manchester. 2013. pp. 4311-4315
- [21]. Liu, H., Setiono, R., "A probabilistic approach to feature selection: a filter solution", *Proceedings of the Thirteenth International Conference on Machine Learning*, 319-327 (1996).
- [22]. S. Zhang, and Z. Zhijin, "Feature selection filtering Methods for emotion recognition in Chinese speech signal." *Signal Processing*. 2008.
- [23]. T.J. Lee, S.M. Park, K.B. Sim, "Electroencephalography Signal Grouping and Feature Classification Using Harmony Search for BCI". *Journal of Applied Mathematics* 2013. pp. 1-9.
- [24]. S. G. Eraldemir, E. Yildirim, S. Yildirim, Yakup Kutlu, "Feature selection for cognitive EEG signals based channel selection and classification". *Innovations and Applications in Intelligent Systems*. 2014 (ASYU 2014). 1-6.
- [25]. A. Subaşı, M. I. Gursoy, M. "EEG signal classification using PCA, ICA, LDA and support vector machines. *Expert Systems with Applications*", 37(12), 8659–8666. <https://doi.org/10.1016/j.eswa.2010.06.065>, 2010

#### BIOGRAPHIES



**Server Göksel ERALDEMİR** was born in Iskenderun, Turkey. He received the Bachelor degree in Computer Education and Instructional Technology from the University of OMU, Turkey, in 2002. Then he worked as a computer teacher in Reyhanlı high school from 2002 to 2009 and after in late 2009 he started to work as a lecturer at the university. Then he received a Master's Degree in electrical and electronics engineering from the University of Mustafa Kemal, in 2014. His current research interests are in biomedical signal processing and data mining.

## BIOGRAPHIES



**MUSTAFA TURAN ARSLAN** was born in Kilis, Turkey. He received the B.S. and M.S. degrees in computer engineering from the University of Erciyes, Turkey, in 2011 and 2016, respectively. Currently, he has been a Ph.D. student in electrical-electronics engineering department in Adana Science and Technology University since February 2018. His current research interests are biomedical signal processing, data mining and artificial intelligence.

## BIOGRAPHIES



**ESEN YILDIRIM**, received the B.S. degree in Electrical and Electronics Engineering from Cukurova University, Adana, Turkey, in 1997, and the M.S. and Ph.D. degrees in Electrical Engineering from University of Southern California (USC), Los Angeles, in 2000 and 2006, respectively. She is currently an Associate Professor of Electrical and Electronic Engineering, at Adana Science and Technology University, Adana, Turkey. Her general research interests include biomedical signal processing, epileptic seizure detection, functional connectivity, learning methods, and emotion recognition from physiological signals.

# A Distributed K Nearest Neighbor Classifier for Big Data

T. Tulgar, A. Haydar and İ. Erşan

**Abstract**—The K-Nearest Neighbor classifier is a well-known and widely applied method in data mining applications. Nevertheless, its high computation and memory usage cost makes the classical K-NN not feasible for today's Big Data analysis applications. To overcome the cost drawbacks of the known data mining methods, several distributed environment alternatives have emerged. Among these alternatives, Hadoop MapReduce distributed ecosystem attracted significant attention. Recently, several K-NN based classification algorithms have been proposed which are distributed methods tested in Hadoop environment and suitable for emerging data analysis needs. In this work, a new distributed Z-KNN algorithm is proposed, which improves the classification accuracy performance of the well-known K-Nearest Neighbor (K-NN) algorithm by benefiting from the representativeness relationship of the instances belonging to different data classes. The proposed algorithm relies on the data class representations derived from the Z data instances from each class, which are the closest to the test instance. The Z-KNN algorithm was tested in a physical Hadoop Cluster using several real-datasets belonging to different application areas. The performance results acquired after extensive experiments are presented in this paper and they prove that the proposed Z-KNN algorithm is a competitive alternative to other studies recently proposed in the literature

**Index Terms**—Big Data Classification, Hadoop, K-Nearest Neighbor, MapReduce.

## I. INTRODUCTION

In the age of the fourth industrial revolution, business's decision-making is highly based on the data retrieved by the internet-connected devices that are capable of collecting and processing ever-growing amounts of information [1].

**A. T. TULGAR**, is with Department of Computer Engineering, Girne American University, Girne, TRNC via Mersin 10 Turkey, (e-mail: tamertulgar@gau.edu.tr).

**B. A. HAYDAR**, is with Department of Computer Engineering, Girne American University, Girne, TRNC via Mersin 10 Turkey, (e-mail: ahaydar@gau.edu.tr).

**C. İ. ERŞAN**, is with Department of Computer Engineering, Girne American University, Girne, TRNC via Mersin 10 Turkey, (e-mail: ibrahimersan@gau.edu.tr).

Manuscript received August 29, 2017; accepted Nov 16, 2017.  
DOI: [10.17694/bajece.419551](https://doi.org/10.17694/bajece.419551)

To achieve precise forecasts, data coming from different environments (e.g. different social media tools, data warehouses, cloud storages etc.) need to be intelligently analyzed by businesses.

Analyzing data retrieved from numerous data sources results in tackling with vast amounts of unstructured raw data, which are big in terms of volume, variety and velocity of acquisition. The process of analyzing such data is recently a popular research area, known as the Big Data Analysis [2].

One of the most important data mining tasks is classification. Classification, which is the task of assigning objects to one of several predefined categories, is a pervasive problem that encompasses many diverse applications [3].

To classify data in the big data age, centralized techniques lack the low classification delay performance, which is vital to cope with the high velocity data streams of big data.

To deal with this timing requirement of the modern data classification, several distributed ecosystems have been tested and used by different researches. One of these distributed ecosystems is popularly known as the Apache Hadoop and the Google's MapReduce Framework [4].

K Nearest Neighbor Classification (K-NN) has been one of the most popular classification algorithms [5]. The classical K-NN algorithm is based on calculating the distances between the test data instance to be classified and all of the instances in the training data set and finding the closest K number of training instances. After detecting the K number of closest training instances, the K-NN algorithm applies majority voting which is the process of detecting the data class with the maximum number of instances among the K selected instances.

Since the classical K-NN algorithm is completely based on individual instance proximities, it heavily suffers from high computation costs. In addition, since the algorithm's decision-making strategy is relying on the individual instance proximities rather than stronger class representations, the algorithm's classification accuracy is also not adequate for modern big data analysis that requires rapid and accurate classification results.

On the other hand, K-NN's individual instance distances strategy makes K-NN a strong candidate for distributed data classification, which is the basis of achieving acceptably low classification delays while classifying big data.

Taking into account the K-NN's suitability to distributed environments, many K-NN based studies which try to improve the K-NN algorithms performance, and working on Hadoop

and MapReduce environment have been recently proposed in the literature [6-16].

In this paper, a new K-NN and MapReduce based algorithm is proposed named as the Z-KNN algorithm. The Z-KNN algorithm tries to remedy the classification accuracy performance of the classical K-NN classifier.

The main idea of the Z-KNN algorithm is to base the classification decision of the classical K-NN algorithm on the class representations by calculating the centroids of the closest Z training instances belonging to the classes of the K closest instances detected by the K-NN algorithm. In other words, the classical majority voting approach is replaced by a stronger classification decision, which is also computationally not expensive.

The rest of this paper is organized as follows: Section II presents the proposed Z-KNN algorithm in detail. The experimental setup and the achieved performance results are presented in section III. Finally, the section IV concludes the paper and states the future works.

## II. THE PROPOSED Z-KNN ALGORITHM

In this section, the proposed Z-KNN algorithm and its MapReduce application will be explained.

### A. The Classical K-NN Algorithm

In a classification task, if a data instance is considered as a vector of feature values, then a data instance  $i$  can be denoted as  $v_i$  which corresponds to a vector containing  $p$  features  $\langle feat_1, feat_2, \dots, feat_p \rangle$ . Hence, a classification task can be defined as detecting the correct data classes of  $n$  test data instances  $tsv_1, \dots, tsv_n$  by using  $m$  training data instances  $trv_1, \dots, trv_m$ .

The classical K-NN algorithm is based on the simple idea of calculating the distances between a test data to be classified and all of the  $m$  number of data instances in the training set. After calculating all of the distances, the classical K-NN sorts the measured distances and uses the first K number of training Neighbors of the tested data.

The classification decision is then given by detecting which data class has the most number of instances among the selected

K nearest neighbors, which is known as the majority voting.

As it can be deduced from the summary of the classical K-NN algorithm given above, the whole decision is based on the individual instance distances between all  $tsv_i$  and  $trv_j$ 's.

Since the calculation of the instance distances is an independent task, being able to distribute the distance calculations to several processes makes the K-NN strategy suitable for distributed environments.

On the other hand, especially when a data set with high number of instances and high number of features per instance needs to be classified, the classical K-NN algorithm's classification accuracy performance becomes lower than its other well-known competitors, like K-Means classification [17].

Hence, it can be deduced that to become a classification algorithm suitable for modern data analysis needs, the K-NN's classification accuracy performance should be improved on a distributed environment.

Taking into account these needs, the Z-KNN algorithm is proposed and explained in sub-section B

### B. The Z-KNN Algorithm

The proposed Z-KNN algorithm is a distributed K-NN based classification algorithm, which is designed to work on MapReduce environment.

Hadoop MapReduce is a software framework for easily writing applications that process vast amounts of data in-parallel on large clusters of commodity hardware in a reliable, fault-tolerant manner [18].

A MapReduce job usually splits the input data-set into independent chunks which are processed by the map tasks in a completely parallel manner. The framework sorts the outputs of the maps, which are then input to the reduce tasks. Typically both the input and the output of the job are stored in a file-system. The framework takes care of scheduling tasks, monitoring them and re-executes the failed tasks [18].

In the MapReduce framework, any distributed task is designed as a combination of at least three functions: The Driver, Mapper and Reducer functions, which are inherited from the corresponding MapReduce classes [18].

The MapReduce framework of the Z-KNN algorithm for  $m$

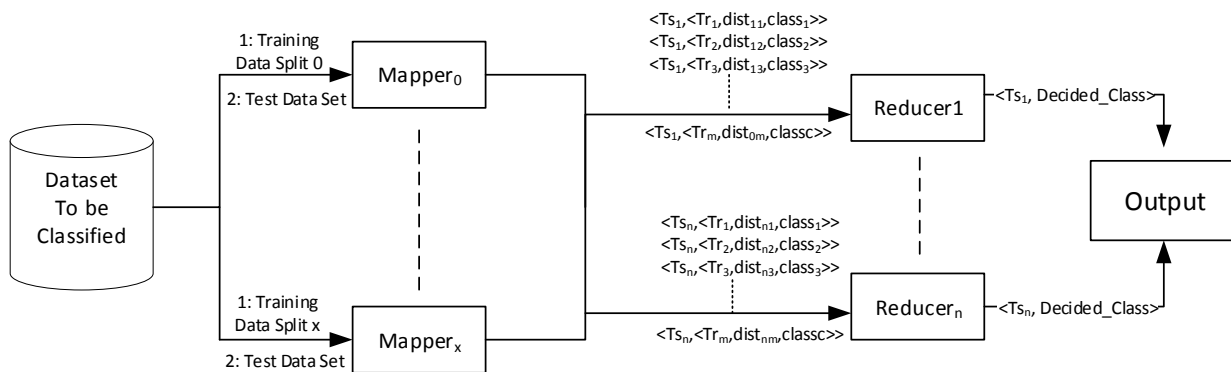


Figure 1. MapReduce Flow of the Z-KNN algorithm



number of instances in the training data set and n number of instances in the testing data set is presented in Figure 1.

As it can be seen in Figure 1, the Z-KNN algorithm's sub tasks are composed of the Mapper and the Reducer functions of the well-known MapReduce Framework.

The Mapper function is responsible of receiving the training data set splits from the MapReduce Driver, which is the main function for configuring the environment and managing the distributed processing running on top of the Hadoop framework, and calculating the distances between the testing instance to be processed and the training instances found in the received training data split.

According to the MapReduce Framework, the output of the Mapper function should be a <Key, Value> vector [18]. In the Z-KNN algorithm, the output Key of the Mapper is the testing instance id and the value is an object, which contains the used training instance, the distance between the testing and the training instances and the class id that the training instance belongs to.

Afterwards, the MapReduce framework shuffles the output <Key, Value> pairs emitted by the mapper functions so that the pairs with common keys are submitted to the same reducer function. In other words, a single reducer will process all of the calculated distances belonging to the same testing instance.

#### Algorithm 1: Z-KNN Mapper Function()

Input: <key, value>  
key : the record id of the training instance  
value: the set of feature values of the training instance

Output to MapReduce Env. : <key1, value1>  
key1: the record id of the test instance  
value1: a vector containing the training instance id, distance and the class id of the training instance

```

1: class_tr = readClassId(value)
2: for i=1 to n
3:   //the loop to iterate each test instance
4:   distij = DistanceFunction(trj, tsi)
5:   Context.write(i, object <trj, distij, class_tr>)
6: end for
7: return

```

#### 1) The Mapper Function:

As explained above the mapper function is responsible of calculating the distances between the training and testing instances. The complete algorithm of the Z-KNN mapper function can be found in Algorithm 1.

As an example to distance calculation, if a single testing instance ts<sub>i</sub> and a single training instance tr<sub>j</sub> are considered, than the distance between these two instances is calculated by Equation (1).

$$dist_{ij} = \sqrt{\sum_{k=1}^p (ts_{i_{feat_k}} - tr_{j_{feat_k}})^2} \quad (1)$$

If we assume that the class id of the training instance tr<sub>j</sub> is class

#### Algorithm 2: Z-KNN Reducer Function()

Input: <key1, <List value1's> distances>  
key1 : the record id of the test instance  
value1: an object which contains <tr<sub>j</sub>, dist<sub>ij</sub>, class\_tr>  
distances: List of all value1s

Output to MapReduce Env. : <key2, value2>  
key2: the record id of the test instance  
value2: the decided class id for test instance i

```

1: Sort_ascending(distances)
2: new LinkedList K_distances
3: new LinkedList Classes
4: new LinkedList Z_instances
4: for i=1 to K
5:   K_distances.add(distances.get(i))
6: end for
7: for all dist ∈ K_distances
8:   if dist.getClass() ∉ classes
9:     classes.add(dist.getClass())
10:  end if
11: end for
12: for all class_id ∈ classes
13:   Z_instances.clear()
14:   for i=1 to Z
15:     if distances.get(i).getClass() = class_id
16:       Z_distances.add(distances.get(i))
17:     end if
18:   end for
19:   μ = 1/Z ∑w=1Z (Z_instances.get(w))
20:   if DistanceFunction(μ, key1) < min
21:     min = DistanceFunction(μ, key1)
22:     decided_class_id = class_id
23:   end if
24: end for
25: key2 = key1
26: value2 = decided_class_id
27: Context.write(key2, value2)
28: return

```

A, then the output of a Z-KNN mapper becomes; <ts<sub>i</sub>, <tr<sub>j</sub>, dist<sub>ij</sub>, class A>>.

#### 2) The Reducer Function:

The reducer function contains the classification decision phase for the test instances, where the main contribution of the proposed Z-KNN algorithm can be seen.

The input of the reducer function is a list of all of the <key1, value1> pairs emitted by the mapper function of the MapReduce framework. It is worth to mention again that a single reducer receives the list of pairs belonging to a common key value. In other words, a single reducer receives the distances of a single test instance to all of the training instances calculated by the mappers.

Upon receiving the input, the reducer function finds the K closest neighbors from all of the training instances by examining the minimum K distances among all the values in the list. Next, the Z-KNN reducer detects to which classes these K neighbors belong (e.g. class A, class B and class C).

The main contribution in Z-KNN classifier depends on

correctly representing the detected classes A, B and C rather than relying only on the distances to the individual training data instances. The main motivation of this strategy is fueled by the fact that, when the size of the data is big and especially when the data is represented by multiple number of classes and large number of features, some data may have similar proximity to different classes instances. In such cases, the minimum distance from the individual instances may not correctly mean that the test instance will belong to that same class. Please consider the following example as a sample case:

Let assume that  $K$  is 5 and the following values are emitted to the reducer for test instance 9. Also, for the sake of example, let us assume that the test instance 9 should be classified to class B:

```
<ts9, <tr3, 0.11, A>
<ts9, <tr11, 0.14, B>
<ts9, <tr27, 0.15, B>
<ts9, <tr23, 0.12, C>
<ts9, <tr42, 0.12, A>
```

In this case, because of majority voting, the classical  $K$ -NN algorithm will conclude that the test instance 9 belongs to class A. The majority voting strategy of the classical  $K$ -NN will end up at this decision since class A and class B has equal number of instances among the  $K$  nearest neighbors, and the class A contains an instance, which has the closest proximity to the testing instance.

Nevertheless, the same class A has an instance, which is further away from the class B instances to the test instance 9.

To give equal chances to classes in the classification decision,  $Z$ -KNN proposes to represent the classes A, B and C among the  $K$  nearest neighbors by centroids and base the classification decision on the distance of the test instance to the centroids of the classes rather than relying on the individual data members' proximities.

To decrease the computation overhead of the proposed proximity to the centroid representation strategy, in the  $Z$ -KNN algorithm, a parameter  $Z$  is introduced.

The parameter  $Z$  in the  $Z$ -KNN classifier is the number of instances from each of the classes A, B and C that have the closest distances to the test instance.

As it was explained earlier in the Mapper function of the algorithm, for each test instance, the distances to every training instance is calculated and emitted to the Reducer function.

Also it is worth to mention that, benefiting from the MapReduce Framework's shuffling/sorting functionality, the coded Reducer Function and the Value class, which defines the value objects in the  $\langle$ Key, Value $\rangle$  pairs, are coded to have an sorted list of the values according to the ascending order of the distances.

That is to say, each reducer receives a list that is already sorted so that the reducers can directly take the first  $Z$  number of elements from each class. Hence, with no extra cost, the Reducer is able to use the already available proximity information.

Repeating the centroid calculation for each class, the  $Z$ -KNN Reducer calculates the class centroids using the first  $Z$  elements,

in other words closest  $Z$  training instances to the test instance to be classified, the reducer ends up with a number of centroids as many as the number of classes found among the  $K$  nearest neighbors.

The  $Z$  parameter contribution simply proposes that, instead of using the complete class population to calculate a class center, using only  $Z$  number of instances of a class, the reducer calculates the centroid for that class with a much lower computation cost and still maintaining a strong class representation compared to relying on individual instance proximities.

Then, the classification decision will be given by the  $Z$ -KNN reducer function, according to which centroid the test instance have the minimum distance.

In this way, the outliers in the class data will have less significance and the decision will be based on a stronger representation of the classes.

The complete  $Z$ -KNN reducer function's algorithm can be seen in Algorithm 2.

### III. THE EXPERIMENTAL SETUP AND THE RESULTS

#### A. The Experimental Setup

The MapReduce functions of the  $Z$ -KNN are coded in Sun JAVA JDK 1.8 [19].  $Z$ -KNN classification experiments are conducted on a small cluster of HP Workstations installed with Ubuntu Linux 16.04 and Apache Hadoop 2.7.4.

To validate the classification scheme, for each dataset used in the experiments, 10-fold cross validation is used, where each test is repeated 10 times and the averages of the 10 tests are considered so that the reliable results can be achieved.

In the experiments, real datasets downloaded from UCI Machine Learning Repository [20] are used. The 5 real datasets that are used in the experiments are summarized in Table I.

TABLE I  
THE REAL DATASETS USED IN THE EXPERIMENTS

Dataset	Instances	Features	Classes
ionosphere	351	34	2
wdbc	569	32	2
wine	178	13	3
seeds	210	7	3
satimage	6435	36	7
pendigits	10992	16	10

#### B. Datasets Used In The Experiments

1) *Ionosphere*: Ionosphere data set is the data coming from the classification of radar returns from the ionosphere. The dataset contains 351 instances belonging to 2 classes. Each instance contains values belonging to 34 features. This dataset is also used in [14].

2) *WDBC*: The Wisconsin Diagnostic Breast Cancer (WDBC) was first used in [21]. The dataset contains 569 instances belonging to 2 classes. Each instance contains values belonging to 32 features. WDBC dataset is also used in [14].

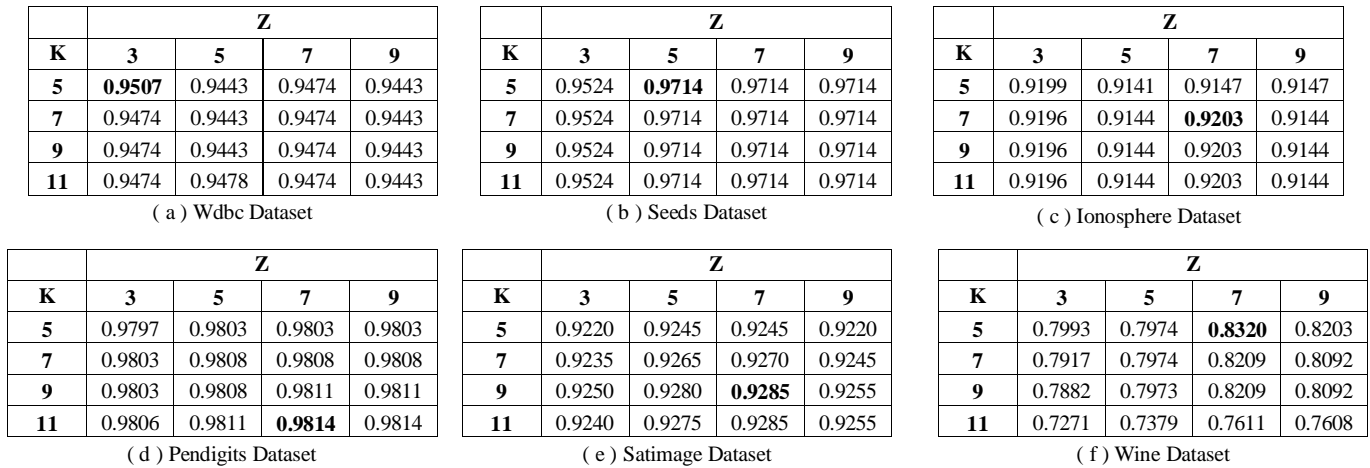


Figure 2. The classification Accuracy Results for (a) Wdbc, (b) Seeds, (c) Ionosphere, (d) Pendigits, (e) Satimage and (f) Wine datasets

3) *Wine*: Wine dataset contains data from chemical analysis to determine the origin of wines. The dataset is composed of 178 instances in 3 classes containing 13 features. Wine dataset is also used in the experiments of [14].

4) *Seeds*: The seeds dataset contains the measurements of geometrical properties of kernels belonging to three different varieties of wheat. The dataset contains 210 instances in 3 classes. Each instance is defined by the values of 7 features. Seeds data set is first used in [22] and also investigated in [14].

5) *Satimage*: The Satimage dataset was generated from Landsat Multi-Spectral Scanner image data. The dataset contains 6435 instances belonging to 7 classes. Each instance contains the data of 36 features. Satimage dataset is also used by [12]–[14].

6) *Pendigits*: Pen-Based Recognition of Handwritten Digits Data Set (pendigits) is a digit database of 250 samples from 44 writers [23]. This dataset contains 10992 instances belonging to 10 classes. Each instance contains the data of 16 features. Pendigits is also used by [12]–[14].

C. The Results

In this section, the results acquired after extensive experiments are presented. The performance of the Z-KNN algorithm is measured in terms of classification accuracy, which represents the ratio of the number of correct classifications to the number of all classifications. The classification accuracy results are given in Fig. 2.

As the overall classification accuracy performance, it can be seen in Fig. 2 that the Z-KNN managed to correctly detect the class of more than 92% of the tested data in all of the data sets.

Also, looking at the accuracy performance of the Z-KNN it can be seen that, for the majority of the datasets, the Z-KNN algorithm manages to detect the correct class of the test instances with K values 5 or 7, without needing to analyze more number of nearest neighbors and hence attaining a reasonable computation cost.

As for the Z parameter, it can be observed that the Z-KNN algorithm manages to achieve a high classification accuracy with 5 to 7 nearest neighbors in the class representation, which also shows that the addition of the Z parameter does not increase the computation cost significantly.

Especially on Pendigits and Satimage datasets, which contain higher number of instances, features and classes compared to other datasets, it is worth to mention that by attaining Z values smaller or equal to 7, Z-KNN shows realistic applicability also to real big data applications.

The accuracy performance of the proposed Z-KNN algorithm and its comparison against the classical K-NN’s accuracy is given in Table II.

TABLE II  
CLASSICAL K-NN VS Z-KNN CLASSIFICATION ACCURACIES

Dataset	Classical K-NN	Z-KNN
Wine	0.8295	0.8320
Wdbc	0.6548	0.9507
Seeds	0.8424	0.9714
Ionosphere	0.6286	0.9203
Pendigits	0.978	0.9814
Satimage	0.9065	0.9285

As it can be seen in Table II, The Z-KNN significantly improves the accuracy performance of the Classical K-NN algorithm in all data sets. In addition, the performance of the Z-KNN algorithm is compared against two algorithms recently proposed in [13-14]. The comparative results are presented in Table III.

TABLE III  
PERFORMANCE COMPARISONS

Dataset	LC-KNN [13]	SR-KNN [14]	Z-KNN
Wine	-	0.9707	0.8320
Wdbc	-	0.965	0.9507
Seeds	-	0.9019	0.9714
Ionosphere	-	0.8971	0.9203
Pendigits	0.9721	0.9452	0.9814
Satimage	0.8883	0.8806	0.9285

As it can be seen in the performance comparisons, Z-KNN performs better almost in all of the datasets compared to other KNN based proposals, which is demonstrating that the proposed Z instance representation significantly improves the accuracy performance of the classical K-NN and some of its variations.

The only dataset where the proposed Z-KNN algorithm is not performing better than the competitors is the Wine dataset. From the results, which were observed in [17], it can be deduced that the low performance of the Z-KNN can be explained by the data distribution features of the Wine dataset that can be remedied by introducing the variance effect contribution to the similarity analysis.

#### IV. THE CONCLUSION AND THE FUTURE WORKS

In this paper a new K-NN based algorithm, named Z-KNN is presented and the performance results are presented after extensive experiments conducted on Hadoop MapReduce environment.

The performance results show that, the main contribution, which proposes to use centroid representation of the data classes instead of relying on individual instance distances, proves to improve the classification accuracy over classical K-NN algorithm.

In the experiments, it was observed that the proposed Z-KNN algorithm proves to be a strong competitor with its high classification accuracy achieved for several different real datasets.

As the future works, it is planned to introduce the effect of the variance to the distance calculation, from the study proposed in [17]. It is expected that, especially the weakness that can be seen in the wine data set can be significantly improved when variance effect is introduced to the distance calculations.

In addition, instead of the classical distance measure, a new similarity measure will be introduced to the Z-KNN algorithm so that the algorithm becomes applicable to any kind of quantitative/categorical features containing datasets.

As an immediate improvement, it is planned to improve the Z instances usage during centroid calculations by introducing a weighted contribution of the Z instances to the centroids. With this improvement, it is expected that, especially if the weights of the Z instances can be set or calculated effectively, the overall classification accuracy of the Z-KNN algorithm can be improved significantly.

Lastly, after the planned future works, the Z-KNN algorithm will be applied to other datasets containing number of instances in the measure of  $10^6$  and above to further prove the algorithms applicability to Big Data applications.

#### REFERENCES

- [1] Klaus Schwab, "The Fourth Industrial Revolution", Crown Business, 2017.
- [2] D. Singh and .K. Reddy, "A survey on platforms for big data analytics", *Journal of Big Data* vol. 1, no. 8, 2014.
- [3] P. Tan, M. Steinbach and V. Kumar, "Introduction to Data Mining", 1st ed., Reading, MA: Addison-Wesley, 2005.
- [4] J. Dean, S. Ghemawat, "MapReduce: A Flexible Data Processing Tool", *Communications of the ACM*, vol. 53 no. 1, pp.72-77, 2010.
- [5] X. Wu et. Al., "Top 10 algorithms in data mining", *Knowledge and Information Systems*, vol. 14, no. 1, pp 137, 2008.
- [6] Fahad et. AL., "A Survey of Clustering Algorithms for Big Data: Taxonomy and Empirical Analysis", *IEEE Trans.on Emerging Topics in Computing*, vol. 2, no.3, pp. 267-279, 2014.
- [7] S. Zhang, M. Zong and D. Cheng, "Learning k for KNN Classification", *ACM Transactions on Intelligent Systems and Technology*, vol. 8, no. 3, pp. 43:1-19, 2017.
- [8] K. Niu, F. Zhao and S. Zhang, "A Fast Classification Algorithm for Big Data Based on KNN", *Journal of Applied Sciences*, vol. 13,no. 12, pp. 2208-2212, 2013.
- [9] Bifet, J. Read, B. Pfahringer and G. Holmes, "Efficient Data Stream Classification via Probabilistic Adaptive Windows", in *Proc. 28th Annual ACM Symposium on Applied Computing*, 2013, pp. 801-806.
- [10] S. S. Labib, "A Comparative Study to Classify Big Data Using fuzzy Techniques", in *Proc. 5th International Conference on Electronic Devices, Systems and Applications (ICEDSA)*, 2016.
- [11] M. El Bakry, S. Safwat and O. Hegazy, "A Mapreduce Fuzzy technique of Big Data Classification, in *Proc. SAI Computing Conference 2016*, pp. 118-128.
- [12] B. Quost and T. Denoeux, "Clustering and Classification of fuzzy data using the fuzzy EM algorithm", *Fuzzy Sets and Systems*, vol. 286, pp. 134-156, 2016.
- [13] Z. Deng, X. Zhu, D. Cheng, M. Zong and S. Zhang, "Efficient kNN classification algorithm for big data", *Neurocomputing*, vol.195, pp. 143-148, 2016.
- [14] S. Zhang, D. Cheng, M. Zong and L. Gao, "Self representation nearest neighbour search for classification", *Neurocomputing*, vol.195, pp. 137-142, 2016
- [15] G. Song, J. Rochas, L. El Beze, F. Huet and F. Magoules, "K Nearest Neighbour Joins for Big Data on MapReduce:A Theoretical and Experimental Analysis", *IEEE Trans. on Knowledge and Data Engineering*, vol. 28, no. 9, pp. 2376-2392, 2016.
- [16] J. Maillo, S. Ramirez, I. Triguero and F. Herrera, "kNN-IS: An Iterative Spark-based design of the k-Nearest Neighbours classifier for big data", *Knowledge-Based Systems*, vol. 117, pp. 3-15, 2017.
- [17] T.Tulgar, A.haydar and İ.Erşan, "Data Distribution Aware Classification Algorithm based on K-Means", *International Journal of Advanced Computer Science and Applications*, Article in Press, 2017.
- [18] T. White, "Hadoop: A Definitive Guide", 4th ed., O'Reilly, 2015.
- [19] J. Gosling, B. Joy, G. Steele, G. Bracha, A. Buckley, (2017,AUG 01). The Java Language Specification-Java SE 8 Edition Online. Available: <https://docs.oracle.com/javase/specs/jls/se8/html/index.html>
- [20] UCI Center for Machine Learning and Intelligent Systems, (2017, AUG 01). UC Irvine Machine Learning RepositoryOnline.Available: <https://archive.ics.uci.edu/ml/>
- [21] O.L. Mangasarian, W.N. Street and W.H. Wolberg, "Breast cancer diagnosis and prognosis via linear programming", *Operations Research*, vol. 43, no. 4, pp. 570-577, July-August 1995.
- [22] M. Charytanowicz, J. Niewczas, P. Kulczycki, P.A. Kowalski, S. Lukasik, S. Zak, "A Complete Gradient Clustering Algorithm for Features Analysis of X-ray Images", *Information Technologies in Biomedicine*, Springer-Verlag, Berlin-Heidelberg, pp. 15-24, 2010.
- [23] F. Alimoglu, E. Alpaydin, "Methods of Combining Multiple Classifiers Based on Different Representations for Pen-based Handwriting Recognition", in *Proc. Fifth Turkish Artificial Intelligence and Artificial Neural Networks Symposium (TAINN 96)*, June 1996.

## BIOGRAPHIES



**Tamer Tulgar** received his B.Sc. degree in Computer Engineering and the M.Sc. degree in Computer Engineering from the Eastern Mediterranean University in 1999 and 2001, respectively. He has a Ph.D. degree in Computer Engineering from the Eastern Mediterranean University (2008) in the area of Cellular Wireless Communication. Tamer Tulgar

worked as a Research Assistant in Eastern Mediterranean University during his Graduate Studies. Upon receiving his Ph.D. degree, he joined the the Girne Ameircan University, Department of Computer Engineering staff, where he currently works as Associate Professor. His research interests include Wireless Communication, Computer Networks, Data Mining, Machine Learning and currently Big Data Analysis.



**Ali Haydar** received the B.Sc. degree in Electrical and Electronics Engineering and the M.Sc. degree in Electrical and Electronics Engineering from the Middle East Technical University in 1991 and 1994, respectively. He has a Ph.D. degree in Electrical and Electronics Engineering from the Eastern Mediterranean University

(1999). He has worked in the research labs of TÜBİTAK during his graduate studies in a project in the field of speech recognition. His research interests include Artificial Neural Networks, Speech Recognition, Optimization, Fuzzy Logic and Data Analysis.



**İbrahim Erşan**, was born in Nicosia, North Cyprus, in 1974. He received the B.Sc. and M.Sc. degrees in electrical and electronic engineering from Eastern Mediterranean University (EMU), Famagusta, in 2000 and the Ph.D. degree in computer engineering from Girne American University (GAU), Kyrenia, in

2012. In 1997, he was an assistant lecturer in EMU, Electrical and Electronics Engineering Department. From 1998 to 2001, he was a research assistant in EMU, Information Technologies Research and Development laboratory. From 2001 to 2005, he was working in construction industry as electrical engineer and project manager. Since 2006, he is working in GAU and since 2017, he is an Associated Professor in Computer Engineering Department. He is currently the head of Computer Engineering Department. He is the author of more than 15 articles. His research interests include decision support systems, machine learning, neural networks and big data.

# 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.

**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](mailto:ibrikci@cu.edu.tr))

Manuscript received August 10, 2017; accepted Nov 16, 2017.  
DOI: [10.17694/bajece.419553](https://doi.org/10.17694/bajece.419553)

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

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
<i>c-CS-s</i>	control mice, motivated to learn, infused with saline	150
<i>c-CS-m</i>	control mice, motivated to learn, infused with memantine	150
<i>c-SC-s</i>	control mice, not motivated to learn, infused with saline	135
<i>c-SC-m</i>	control mice, not motivated to learn, infused with memantine	135
<i>t-CS-s</i>	trisomy mice, motivated to learn, infused with saline	135
<i>t-CS-m</i>	trisomy mice, motivated to learn, infused with memantine	135
<i>t-SC-s</i>	trisomy mice, not motivated to learn, infused with saline	105
<i>t-SC-m</i>	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:  $P(x_i | \text{parents}(x_i))$  based on DAG [15-17].

The probability of class given the particular sample of  $x_1, \dots, x_n$  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].

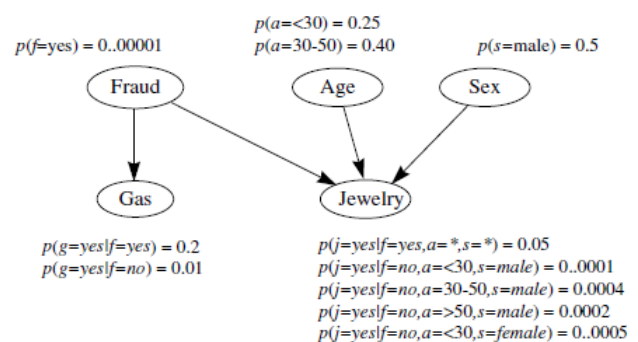


Figure 1. Motion Scenario

### 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 effortlessly.
- 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

```

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

```

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

```

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] Györödi, C., Györödi, 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*. 3<sup>rd</sup> 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 Boriratr, Sirapat Chiewchanwattana, Khamron Sunat, Pakarat Musikawan and Punyaphol Horata. *Harmonic extreme learning machine for data clustering*. Computer Science and Software Engineering (JCSSE), 13<sup>th</sup> 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 Ibrikli** received his BS degree in physics (Cukurova University, Adana, Turkey), MSc in computer science (Nova Southeastern University, Fort Lauderdale, 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.

# Analysis of Photonic Crystal Tuned by Nematic Liquid Crystals

F. Karaomerlioglu

**Abstract**—Modelling of left-handed metamaterial-based photonic nanostructures in functional optoelectronic devices have been studied in present paper. Two dimensional photonic crystal array on a narrow-band semiconductor base filled with a liquid crystal was analyzed. Photonic band structure and isofrequency contours of the nanostructures containing both inorganic and organic components were calculated for both TE and TM modes by using the plane wave expansion method. Silver telluride was used as a narrow-band semiconductor and E7 type as a liquid crystal. The photonic crystal structure that is designed as hexagonal rods shape in an air background is planned for the square lattice.

**Index Terms**—Metamaterial, Organic Compound, Photonic Crystal, Semiconductor.

## I. INTRODUCTION

Optics of nanotechnology and 3D artificial structures have actuated the research in the properties of photonic crystals (PCs) and have concluded the increase of applications of photonic band (PB) gap materials [1, 2]. The PCs essential characteristic is the light presence of allowed and forbidden frequency bands. It is feasible manipulating the light by PCs. Thanks to this property, it is designed new optical devices with PCs. In order to design devices, research on tuning the optical properties of PB gap has been an increase. Some research on one dimension (1D), two dimensions (2D), and three dimensions (3D) PCs has been studied [3-7].

It has been researched negative refraction recently. In opposition to left-handed materials (double-negative metamaterials), single-negative materials and indefinite materials, PCs can display negative refraction behaviors that are only detected by the properties of their PB structures and isofrequency contours [8-14].

In recent years, novel investigations have reached a viewpoint of LCs owing to the tunable light. Refractive indices of LCs can be changed by rotating LCs' directors [15-17].

**F. KARAOMERLIOGLU**, is with Department of Electrical Engineering University of Mersin University, Mersin, Turkey, (e-mail: [filizkrm@mersin.edu.tr](mailto:filizkrm@mersin.edu.tr)).

Manuscript received August 18, 2017; accepted Nov 16, 2017.  
DOI: [10.17694/bajece.419555](https://doi.org/10.17694/bajece.419555)

It is familiar that the concept of topological order provides a new perspective. It has generated intense recent interest in searching for nontrivial topological materials, topological insulators (TIs) [18]. The differential feature of a TI is existence of robust conducting edge or surface states on the boundary of insulators. These typical boundary states have a topological origin, and so have potential for applications in spintronics and quantum computation devices [19]. Until now, TIs have been theoretically and experimentally affirmed [19, 20].

In this paper, it is proved and enhanced the optical properties in the 2D PC structure in TI rods based on the left-handed metamaterial tuned by LCs in theory. The investigation was achieved by controlling the intensity of the optical properties that had different materials added to a certain structure.

## II. METHODOLOGY

The plane wave expansion (PWE) methods' principles are depended on a direct numerical solution of the Maxwell's equations. Bloch's theorem [21] is used to expand the  $H(\vec{r})$  field in terms of plane waves because the light waves are transmitted in periodic structures, as

$$H(\vec{r}) = \sum_{\vec{G}} h(\vec{G}) \hat{e}_{\vec{G}} e^{i(\vec{k} + \vec{G})\vec{r}} \quad (1)$$

where  $\vec{k}$  is a wave vector in the Brillouin zone of the lattice and  $\hat{e}_{\vec{G}}$  is the direction that is perpendicular to the wave vector  $(\vec{k} + \vec{G})$  because of the transverse character of the magnetic field  $H(\vec{r})$ ,  $\nabla \cdot H(\vec{r}) = 0$ .

## III. RESULTS AND DISCUSSION

Using the PWE methods, the PC structure, composed of a PC in TI rods based on the left-handed metamaterial infiltrated with LCs, is designed for the square lattice. PCs structures that are designed as hexagonal rods shape are computed. Silver telluride ( $\text{Ag}_2\text{Te}$ ) was used as TI material and E7 type as nematic LCs.  $\text{Ag}_2\text{Te}$  is a narrow-gap semiconductor and has been predicted to be new families of TIs with a highly anisotropic.

This paper is intended for characterising and comparing 2D PC structures which differ by the characteristics of their PB gap and isofrequency dependences.

**A. Photonic band structure of 2D PC with hexagonal rods**

It is considered the results obtained from the computing of PB structure of the spectrum for the 2D PC of the TI rods type. This type composes of the elements in the form of dielectric hexagonal shaping a square lattice filled with and without LC.  $Ag_2Te$  is a highly anisotropic TI.  $Ag_2Te$  has two different basis refractive indices as the ordinary-refractive index  $n_o = 3.7977$  and the extraordinary refractive index  $n_e = 4.6960$  at  $\lambda > 1 \mu m$ . The calculations are performed for TI PCs with the permittivity of the hexagonal rods 14.423 and the period of the structure  $a = 1 \mu m$ . LCs have commonly two kinds of dielectric constants;  $\epsilon_o$  and  $\epsilon_e$ , ordinary and extraordinary dielectric constants. According to the electric fields perpendicular and parallel light waves, the director of the LC have ordinary and extraordinary refractive indices, respectively. The ordinary-refractive index of E7 type LCs is  $n_o = 1.51$  and the extraordinary refractive index,  $n_e = 1.69$  at  $\lambda = 1.55 \mu m$  [22].

Schematic views of the proposed 2D PC of TI hexagonal rods without and with LC-infilled in an air background ( $\epsilon_a = 1$ ) in a square lattice are shown in Fig.1. PB structure for TE and TM mode is calculated along direction that includes the high symmetry points  $\Gamma$ , X and M for the Brillouin zone in a square lattice. It is assumed that  $d_1 = 0.45a$  and  $d_2 = 0.2a$  denote the outer and inner edge of TI hexagonal rods.

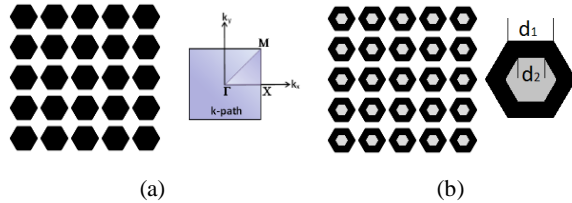


Fig. 1. 2D PC structure of TI hexagonal rods for square lattice (a) without (b) with LC-infilled.

Fig.2 shows the seed PB structure of the PC (inset, from Fig.1 (a)) with the permittivity of the dielectric hexagonal and the permittivity of free space 1. For the ordinary refractive index, this PC has two PB gap of TE mode and three PB gap of TM mode. In TE mode the PC band gap along the direction between high symmetry points  $\Gamma-X$  direction of the Brillouin zone lies in the frequency range from  $0.237(2\pi c/a)$  to  $0.240(2\pi c/a)$  and from  $0.448(2\pi c/a)$  to  $0.456(2\pi c/a)$ . For TI hexagonal rods without LC-infilled, relative widths are 1.24% from band 1 to band 2 and 1.70% from band 4 to band 5 of TE mode, respectively (Fig.2 (a)). For TM mode, the frequency ranges have from  $0.190(2\pi c/a)$  to  $0.194(2\pi c/a)$ , from  $0.309(2\pi c/a)$  to  $0.326(2\pi c/a)$ , and from  $0.501(2\pi c/a)$  to  $0.515(2\pi c/a)$ . Similarly, relative widths are 1.86% from band 1 to band 2, 5.39% from band 3 to band 4, and 2.76% from band 7 to band 8 of TM mode in Fig. 2 (b).

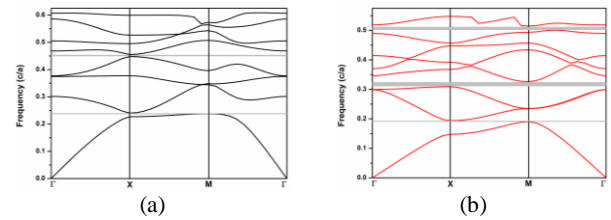


Fig.2. PB structure of (a) TE (b) TM mode for TI hexagonal rods without LC-infilled.

In order to determine the optical properties of PC, structure and material are very important. For that reason, it is changed the PC structure that TI hexagonal rods are infiltrated with LC obtaining optimum results. Dispersion of PC in combination with the LC causes the appearance of different PB gaps in the continuous spectrum of PC, imperceptible on the scale of Fig.2. These effects are illustrated in Fig.3. It can be seen from Fig.3 (a) that the presented fragment of the PB structure of the spectrum exhibits different band gap situated in the frequency range between  $0.245(2\pi c/a)$  and  $0.265(2\pi c/a)$  and between  $0.411(2\pi c/a)$  and  $0.418(2\pi c/a)$  of TE modes. Relative widths are 8.07% from band 1 to band 2 and 1.61% from band 2 to band 3. When TI hexagonal rods are infiltrated with LC of the extraordinary refractive index, two band gaps in TM mode is shown in Fig.3 (b). PB gap has relative widths of 1.19% from band 3 to band 4 and 3.65% from band 5 to band 6, and the frequency range between  $0.325(2\pi c/a)$  and  $0.329(2\pi c/a)$  and between  $0.443(2\pi c/a)$  and  $0.459(2\pi c/a)$  of TM modes.

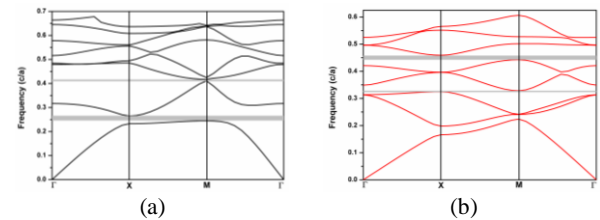


Fig.3. PB structure of (a) TE mode (b) TM mode for TI hexagonal rods with LC-infilled.

A comparison of Fig.2 and Fig.3 shows that the spectra of both types PC have different PB gaps between different bands.

**B. Isofrequency surface of 2D PC with hexagonal rods**

Dependency of the isofrequency has a simple physical meaning to analyze 2D geometries. Because this dependence describes all of the possible waves with the given frequency and various wave vectors, the directions of the reflected and the refracted rays can be determined by elementarily finding the points in isofrequency dependences of media at a known orientation of the boundary and a given angle of incidence of the wave [23].

It is made use of symmetry calculating the isofrequency surfaces over the irreducible zone of the entire Brillouin zone. First, it is considered the isofrequency surface of TI hexagonal rods without LC-infilled for the first two TM bands of a

square lattice. In Fig.4, it is reproduced PC with TI hexagonal rods with the same parameters. For the first band, the map was discretized using five field points per edge of the unit cell in Fig.4 (a). The map was discretized using four field points per edge of the unit cell for the second band in Fig.4 (b).

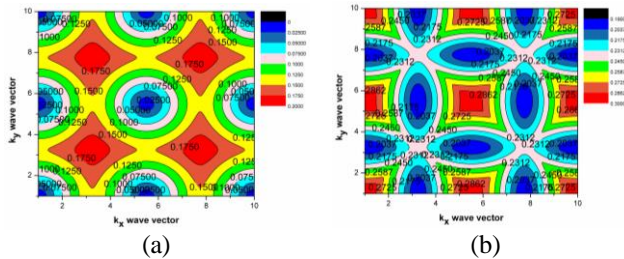


Fig.4. The isofrequency contours of PC with TI hexagonal rods without LC-infill for the square lattice (a) first band (b) second band.

In Fig.5, it is derived for PC with TI hexagonal rods of LC-infill for the first two TM bands. For the first band, the map was discretized using five field points per edge of the unit cell in Fig.5 (a). The map was discretized using four field points per edge of the unit cell for the second band in Fig.5 (b).

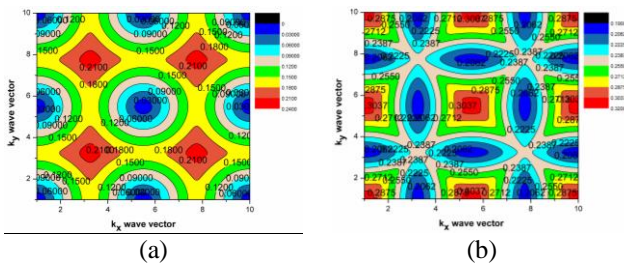


Fig.5. The isofrequency contours of PC with TI hexagonal rods with LC-infill for the square lattice (a) first band (b) second band.

The conventional PB structure plot only shows modes along the  $\Gamma-X-M-\Gamma$  line. The isofrequency contour allows to see all of the possible wave vectors uniformly sampled in a space. In fact, the PB structure shows all frequencies correspond to each given wave vector, while the isofrequency contour shows all wave vectors correspond to each given frequency. The information of PB structure and the isofrequency contour are correlated and complement each other.

Although the arms have different lengths, the figure uses a fixed number of sample points along each arm of the  $\Gamma-X-M-\Gamma$  contour. In order to estimate the shape of the contour from PB structure plot this is an important detail. It is easy to see the shape of the contour when it is calculated the isofrequency contour. (Figures 4 and 5). The circular region means isotropic propagation for the first band, while other shapes indicate anisotropic behaviors for the second band.

It can be found that PB structures deduced from the observed isofrequency contours are in overall agreement with that of the simulation results. The observability of the isofrequency contour due to any distortion in structures is important and could be used to research the dispersion of periodic structures and understand light propagation properties in the periodic photonic structures.

#### IV. CONCLUSIONS

It was analyzed the optical properties in a 2D PC structure of TI hexagonal rods filled without and with LCs in a square lattice. For TE and TM mode, the PB structure is calculated along with the high symmetry point for the Brillouin zone.

These results have been shown that the dispersion of 2D photonic structure in combination with the nematic LC leads to qualitative changes in PB structure of the electromagnetic spectrum.

In practical usage, kind of LC infilled PCs based on a left-handed metamaterial are promising materials in order to use in the design of solar cell, super-lens applications, and a novel optoelectronic devices.

#### ACKNOWLEDGEMENTS

I would like to express my gratitude to Prof. Dr. Amirullah M. Mamedov and Ekmel Ozbay for invaluable discussions, guidance and suggestions.

#### REFERENCES

- [1] E. Yablonovitch, Photonic band-gap structures, *J. Opt. Soc. Am. B* 10: 283-295, 1993.
- [2] J. D. Joannopoulos, S. G. Johnson, J. N. Winn and R. D. Meade, *Photonic Crystals: Molding the Flow of Light*, Princeton University Press, Princeton, NJ, 2008.
- [3] C. Sibilía, T. M. Benson, M. Marciniak, T. Szoplik, *Photonic Crystals: Physics and Technology*, Springer, Italia, 2008.
- [4] K. Sakoda, *Optical Properties of Photonic Crystals*, Springer, Germany, 2005.
- [5] M. C. Gupta, J. Ballato, *The Handbook of Photonics*, CRC Press, USA, 2007.
- [6] J. M. Brosi, *Slow-Light Photonic Crystal Devices for High-Speed Optical Signal Processing*, University of Karlsruhe, Germany, 2009.
- [7] A. E. Serebryannikov, A. Y. Petrov, E. Ozbay, Toward photonic crystal based spatial filters with wide angle ranges of total transmission, *Applied Physics Letters* 94: 181101, 2009.
- [8] G. Sun, A. G. Kirk, Analyses of negative refraction in the partial bandgap of photonic crystals, *Optics Express* 16 (6): 4330-4336, 2008.
- [9] E. Cubukcu, K. Aydin, E. Ozbay, S. Foteinopolou, C. Soukoulis, Negative Refraction by Photonic Crystals, *Nature* 423: 604, 2003.
- [10] R. Moussa, S. Foteinopoulou, Lei Zhang, G. Tuttle, K. Guven, E. Ozbay, C. M. Soukoulis, Negative refraction and superlens behavior in a two-dimensional photonic crystal, *Physical Review B*. 71: 085106, 2005.
- [11] K. Guven, K. Aydin, K. B. Alici, C. M. Soukoulis, E. Ozbay, Spectral negative refraction and focusing analysis of a two-dimensional left-handed photonic crystal lens, *Physical Review B* 70: 205125, 2004.
- [12] L. Shi, H. Yin, X. Zhu, X. Liu, J. Zi, Direct observation of iso-frequency contour of surface modes in defective photonic crystals in real space, *App. Phys. Lett.* 97: 251111, 1-3, 2010.
- [13] Y. Y. Wang, L. W. Chen, Tunable negative refraction photonic crystals achieved by liquid crystals, *Opt. Express* 14: 10580-10587, 2006.
- [14] G. V. Eleftheriades, K. G. Balmain, *Negative-Refraction Metamaterials: Fundamental Principles and Applications*, John Wiley & Sons, Canada, 2005.
- [15] I. C. Khoo, S. T. Wu, *Optics and nonlinear optics of liquid crystals: Electro-optical properties of liquid crystals*, World Scientific; Singapore, pp.100-258, 1993.
- [16] F. Karaomerlioglu, A. M. Mamedov, E. Ozbay, Organic semiconductor-based photonic crystals for solar cell arrays: band gap and optical properties, *J Modern Optics*, 2014, doi.10.1080/09500340.2014.967320.
- [17] F. Karaomerlioglu, A. M. Mamedov, E. Ozbay, Optical properties of metamaterial-based devices modulated by a liquid crystal, *Appl. Phys. A* 117(2): 611-619, 2014.
- [18] M. Z. Hasan and C. L. Kane, "Topological insulators", *Rev. Mod. Phys.* 2010, 82(4), 3045-3067.

- [19] C. He, L. Lin, X. C. Sun, X. P. Liu, M. H. Lu, and Y. F. Chen, "Topological photonic states", *Int. J. Mod. Phys. B* 2014, 28(2), 1441001(1-15).
- [20] S. Lee, J. In, Y. Yoo, Y. Jo, Y. C. Park, H. J. Kim, H. C. Koo, J. Kim, B. Kim, K. L. Wang, "Single Crystalline  $\beta$ -Ag<sub>2</sub>Te Nanowire as a New Topological Insulator", *Nano Lett.* 2012, 12, 4194-4199.
- [21] C. Kittel, *Introduction to Solid State Physics*, John Wiley & Sons, New York, 2005.
- [22] J. Li, S. T. Wu, S. Brugioni, R. Meucci, S. Faetti "Infrared refractive indices of liquid crystals", *J. Appl. Phys.* 2005, 97, 073501(1-5).
- [23] E. H. Lock, The properties of isofrequency dependences and the laws of geometrical optics, *Physics-Uspeski* 51(4): 375-393, 2008.

### BIOGRAPHIES



**FİLİZ KARAOMERLIOĞLU** received B.S. degree in Hacettepe University, Department of Physics Engineering, Ankara, Turkey in 1997. She received M.Sc. degree in Electrical-Electronics Engineering and Ph.D. degree in Physics from Cukurova University, Adana. She was a Research Assistant with Electrical-Electronics Engineering. From 2013 to

2014 she was a Visiting Researcher with Nanotechnology Research Center (NANOTAM), Bilkent University, Ankara. She joined Mersin University (Mersin, Turkey) in 2009, where she is currently an Assistant Professor in Electrical-Electronics Engineering Department. Her research interests include optics, optical devices, optical materials, optics and photonics, photonic crystals, nanophotonics and metamaterials.

# Speech Emotion Classification and Recognition with different methods for Turkish Language

C.Bakir, M.Yuzkat

**Abstract**— In several application, emotion recognition from the speech signal has been research topic since many years. To determine the emotions from the speech signal, many systems have been developed. To solve the speaker emotion recognition problem, hybrid model is proposed to classify five speech emotions, including anger, sadness, fear, happiness and neutral. The aim this study of was to actualize automatic voice and speech emotion recognition system using hybrid model taking Turkish sound forms and properties into consideration. Approximately 3000 Turkish voice samples of words and clauses with differing lengths have been collected from 25 males and 25 females. In this study, an authentic and unique Turkish database has been used. Features of these voice samples have been obtained using Mel Frequency Cepstral Coefficients (MFCC) and Mel Frequency Discrete Wavelet Coefficients (MFDWC). Moreover, spectral features of these voice samples have been obtained using Support Vector Machine (SVM). Feature vectors of the voice samples obtained have been trained with such methods as Gauss Mixture Model( GMM), Artificial Neural Network (ANN), Dynamic Time Warping (DTW), Hidden Markov Model (HMM) and hybrid model(GMM with combined SVM). This hybrid model has been carried out by combining with SVM and GMM. In first stage of this model, with SVM has been performed subsets obtained vector of spectral features. In the second phase, a set of training and tests have been formed from these spectral features. In the test phase, owner of a given voice sample has been identified taking the trained voice samples into consideration. Results and performances of the algorithms employed in the study for classification have been also demonstrated in a comparative manner.

**Index Terms**—MFCC, MFDWC, emotion, HMM, hybrid model.

## I. INTRODUCTION

Today, with the development of technology, security problems have also come to light. Biometric systems, such as authentication in particular, constitute an important part of the security issue. For this reason, it is necessary to determine the forensic soundings of the voice recordings

**C. BAKIR**, is with Department of Computer Engineering University of Iğdir University, Istanbul, Turkey, (e-mail: [cigdem.bakir@igdir.edu.tr](mailto:cigdem.bakir@igdir.edu.tr)).

**M. YUZKAT**, is with Department of Computer Engineering University of Mus Alparslan Technical University, Istanbul, Turkey, (e-mail: [m.yuzkat@alparslan.edu.tr](mailto:m.yuzkat@alparslan.edu.tr)).

Manuscript received August 22, 2017; accepted Nov 16, 2017.  
DOI: [10.17694/bajece.419557](https://doi.org/10.17694/bajece.419557)

subject to various crimes and the emotions of the people in these voice recordings at that moment. Especially in commercial transactions, some studies have been carried out to prevent the transfer of information belonging to persons to other persons. Handwriting recognition, signature recognition, face recognition, iris recognition, voice recognition constitute several of these studies [8].

Despite the fact that speech recognition technology has a very long history, attempts to extract emotion from human voice are still new and attract great attention. Obtaining the necessary data for extracting emotion constitutes an important problem. Because, there are so many kinds of emotions and it is very difficult to determine these emotions.

Various studies have been performed to determine the emotion of voice and speaker. Shami et al. have performed the study of emotion recognition from speech data with k-nearest neighbors, (kNN), Support Vector Machines (SVM) ve Ada-Boosted decision tree machine learning techniques on four different databases. In this study, the success of feature extraction techniques, AIBO and segmentation based approach, SBA on different databases and different classification techniques are compared. Both feature extractions gave different results on different databases [1].

Chen et al. have developed a three-level model to distinguish six emotions as independent of the speaker. Various features were selected from 288 individuals by using the Fisher ratio for each level of emotion. In order to measure the success of the proposed system, Principal Component Analysis (PCA) dimension reduction method was used and for classification, ANN and SVM were used. The results obtained have been presented comparatively. However, since the frequency of speech changes abruptly in some emotions, more study is required to be performed in this respect [2].

He et al. proposed two different methods of feature extraction for emotion classification from speech data. In the first method, they applied the EMD (Empirical Model Decomposition) method which calculates the average entropy of speech data. In the second method, however, they studied with a method that calculates the average spectral energy in the lower bands of the speech spectrogram. He et al. calculated the success of these two methods by using GMM and kNN classification algorithms on two different databases. They also compared the success of these two methods by using the MFCC feature extraction method [3].

Polzehl et al. conducted emotion recognition studies by using acoustic characteristics of speech data of children. They



tried to distinguish angry feelings and feelings which are not angry. In this study, frame-based cepstral properties reduced in size were classified by acoustic properties ANN and SVM. Furthermore, in the study, feature selection was made with the Data Acquisition Ratio [4].

Nwe et al. have worked to distinguish the six emotions on the speech data. They classified feelings with HMM. In this study, a database containing 90 different emotions taken from two speakers was used. However, this work was also carried out depending on the speaker [5].

Bhaskar et al. have proposed a hybrid approach to classify feelings in speech and text. In the study they made, both the textual and speech features were combined. For classification, they used the multi-class SVM method. However, only 11 features have been used in the study. More features need to be selected to achieve the desired performance [6].

Lee et al. Carried out an emotion study by using the Recurrent Neural Network (RNN) algorithm. In this study, they applied the Bidirectional long short-term memory (BLST) algorithm to determine the time-varying emotions. In this way, the changes that occurred in the emotions, that is, unspecified emotions whose tag changed were tried to be determined [7].

Feature extraction, classification techniques used of the study performed, experimental study of results and conclusion were given respectively in the section 2, section 3, section 4 and section 5.

## II. FEATURE EXTRACTION METHODS

The study has been realised on a unique database, which was formed from Turkish sound samples taken from both men and women. These sound samples are trained by getting dispersed to various feature vectors with MFDWC, MFCC and LPCC. In the second stage, the feature vectors of the recorded sound signals are trained with classification algorithms such as artificial neural network (ANN), Dynamic Time Warping (DTW), Hidden Markov Model (HMM) and Gauss Mixture Model (GMM). The speech for recognition is decided by looking at sound signals in the test and training data after the system is trained. Furthermore, the classification success in recognising the gender of speaker was calculated separately for 5 feature vectors and the success of the methods was presented comparatively by training the feature vectors, obtained from speech signals.

### A. Mel Frequency Cepstral Coefficients (MFCC)

MFCC is a feature extraction method, that is used in sound processing. It is used to extract important information and features by dividing the sound data to its subsets. The steps of feature extraction technique of MFCC is indicated in Figure 1[18].

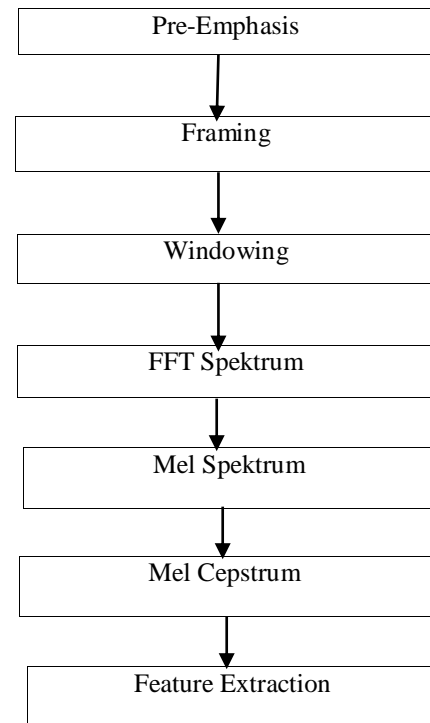


Fig.1. Feature extraction steps of MFCC

Two filters are used in MFCC feature extraction method. The first filter has a linear distribution of frequency values under 1000 Hz and the other has a logarithmic distribution of frequency over 1000 Hz. Pre-emphasis stage is the first stage in obtaining MFCC feature vector.

The sound signals, which have high frequency, are passed through a filter at this stage. This way, the energy of the sound is increased at high frequency. The sound signals are analog. The sound signals are converted from analog to digital by getting divided into small frames between 20 and 40 ms during the framing stage and it is divided into N frames. The sound signal is moved by sliding the sound signal at the windowing stage. This way, the closest frequency lines and the frame, which will come by windowing, that is used are combined. The window type, width and sliding amount are determined at this stage. Each of N frames is transmitted from the time space to the frequency space with Fast Fourier Transformer (FFT). The spectral features of sound signals are shown in frequency space. MEL spectrum is obtained by calculating the total weight of these spectral features. This MEL spectrum is formed from triangle waves and are formed by getting passed through a series of filters. MEL spectrum reduces the noise by lowering two neighbour frequencies. The logarithm of signal is taken at the stage of MEL spectrum and the signal is transmitted back again from frequency space to the time space. MEL frequency cepstrum factors are obtained by using DCT (Discrete Cosine Transform) in time space.

### B. Mel-frequency Discrete Wavelet Coefficients (MFDWC)

The study in question has been performed, based on a unique database comprising Turkish voice samples collected from men and women. These voice samples were separated into various feature vectors with MFDWC, and trained. MFDWC is a feature extraction method employed in the speech processing. It is used to extract significant information and features by dividing voice data into subsets. Feature extraction steps of MFDWC technique is shown in the Figure 2 [19].

Sample speech signal is shown between the 40-40000 Hz range in the MFDWC feature extraction method. Speech signal is divided into frames after the pre-processing step. Hamming window has been used in this study in order to smoothen the transition of speech samples between the frames. One Mel shows the frequency of voice tone. Mel-scale is scaled between actual frequency of voice signal and estimated voice frequency. For this reason total energy of every frame is calculated. Classification success in speaker identification has been calculated on an individual basis for MFDWC-5 vectors by training the feature vectors obtained from voice signals by means of different methods.

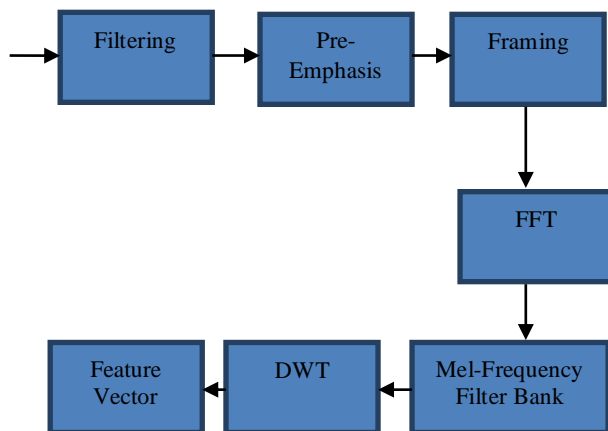


Fig.2. Feature extraction steps of MFDWC

### III. METHOD

In Figure 3, the steps of the study are given. In this study, the sound samples, taken from 25 males and 25 females in different age ranges, have been separated to their feature vectors with SVM. The education and test samples have been formed from these voice feature vectors. These train and test samples have been coached according to ANN, DTW, HMM and recommended hybrid model and speech emotion recognition transaction has been realized automatically. Furthermore, the results, obtained with ANN, DTW, HMM and hybrid method, have been given comparatively.

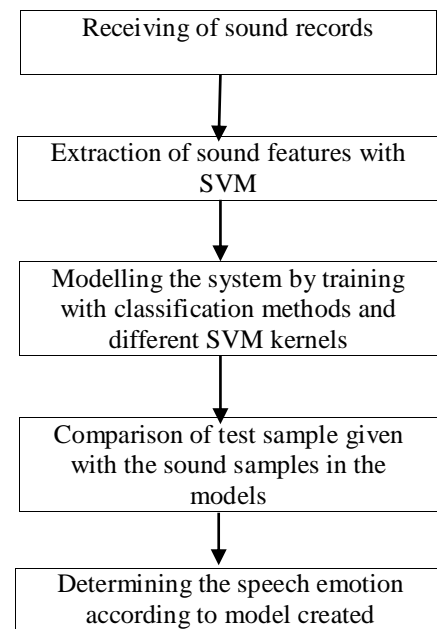


Fig.3. Study steps

### A. Artificial Neural Network

ANNs have a very wide fields of application up to automotive, banking, defense industry, electronics, entertainment, finance, insurance, manufacture, oil and gas, robotics, telecommunication and transportation industry.

Artificial neural networks are information systems which mirror human brain function, and classify the data through learning. They have been developed, being based on a principle of human brain functioning. In other words; ANNs have been developed with a logic similar to the biological neural networks, and are data processing structures connected to each other with weights.

ANNs comprise of input layer, output layer and hidden layers. Data is received into neural networks through input layer. And it is transferred to outside through output layer. Layers between input and output layers constitute hidden layers.

Neurons in the feed-forward neural networks are connected just in the forward direction [11]. Each layer of neural network contains the connection of next layer and these connections are not in the backward direction. In a sense, there is a hierarchical structure between neurons, and the neurons located in one layer can only communicate data to the next layer. Structure of a feed-forward ANN is shown in the Figure 4.

Backward propagation network shows how to train a neuron [12]. Trainer is a sort of learning. Network is maintained both with the sample inputs and expected outputs when the trainer method is employed. Expected outputs are compared with actual outputs for the networks the inputs of which are given. Error is calculated in case the expected outputs are used, and weights of

various layers are adjusted in the backward direction from output layer to input layer. In other words, it is given for both input data and output data. Network updates its coefficients in order to obtain the expected output.

ANN is the most widely used method. In this algorithm, error in the output layer is calculated at the end of each iteration, so this error is transmitted to all neurons in the direction from output layer to input layer, and weights are readjusted according to the error margin. Such error margin is distributed to the previous neurons located before the said neuron in proportion to their weights.

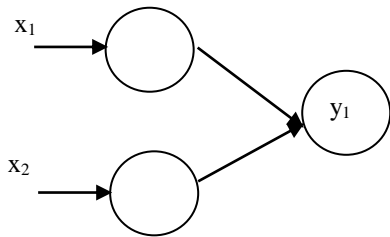


Fig.4. Feed-forward neural network

Layers are located one after another in a multilayer artificial neural network. Outputs of neurons in a layer will be given as their weights, to the input of next layers, and these weight are used in the calculation of outputs for the next layer. Weights of the hidden layer between input and output layers are calculated [12].

*B. Dynamic Time Warping*

Dynamic Time Warping (DTW) finds out to which speaker the voice signal given belongs, by calculating the similarity between the time-variant two speech signals. The most optimal time curve can be identified between two signals with this method.

$$Q = q_1, q_2, \dots, q_i, \dots, q_n \tag{1}$$

$$C = c_1, c_2, \dots, c_j, \dots, c_m \tag{2}$$

Q and C in the equation 1 and equation 2 demonstrate two distinct speech signals; n and m show the lengths of these speech signals [6]. In this case, the ratio of similarity between Q and C signals is calculated using Euclid length as in the equation 3 [14].

$$d(q_i, c_j) = (q_i - c_j)^2 \tag{3}$$

A matrix (i,j) is generated for Q and C. Accumulated distance matrix is calculated using this matrix.

$$D(i,j) = \min[D(i-1,j-1), D(i-1,j), D(i,j-1)] + d(i,j) \tag{4}$$

*C. Hidden Markov Model*

A lot of studies have been carried out with regard to the Hidden Markov Models (HMM) in many fields from past to today. HMM has been used in a wide manner in face recognition, speech recognition, voice recognition, hand script recognition, human body motion recognition, bioinformatics, estimation of gene, cryptanalysis, protein structure and sequence, DNA sequence and pattern recognition.

In Hidden Markov Model (HMM) the aim is to try to estimate future situations that will likely occur in cases when the existing situations are given as an input to the system. HMM is a stochastic process since it generates different output whenever it is operated. In addition, system in Markov models may move from its own state to another state according to the probability distribution, or remain in the same state. Probabilities occurred in the states are called as transition probabilities. States are not seen by the observer as distinct from HMM normal Markov model. However transition subject to the states may be observed. HMM speaker recognitions systems comprise of the following steps [13].

- $S = \{S_1, S_2, \dots, S_Q\}$  shows current status of the speech signals generated where there are Q numbers of states.
- Initial state probabilities is determined in a discrete time, t. ( $\pi = \{P_T, (S_i, |t=0, S_i, \epsilon S)\}$ )
- Transition probabilities are calculated according to the current states.  $a_{ij} = (P_T (S_j \text{ t in time t} | S_j \text{ in time t-1}), S_i \in S, S_j \in S)$
- F, which is the number of features observed, is determined.
- Probability distribution of speech signal will be calculated in this way. ( $b_x = \{b_j(x) = P_T(x(S_i), S_i \in S, x \in F)\}$ )
- HMM generated is demonstrated by  $\lambda = (a, b, \pi)$ .

*D. Gauss Mixture Model*

Gauss Mixture Model is a statistical method based on the weight combination of the Gaussian distribution of one or more audio signals. The sum of the weighted combinations of Gaussian intensity is shown in equation 5 [10].

$$p(x|\lambda) = \sum_{i=1}^M p_i b_i(x) \tag{5}$$

$x$  shows the feature vector,  $D$  shows the dimensional random vector  $b_i(x)$ ,  $i = 1, \dots, M$  shows the density components, and  $p_i$  shows the mixture weight. The parameters of this model are found by the ExpectationMaximization (EM) algorithm. All classes in the training data are expressed by

independent Gaussian density function. The most optimal density components to determine the mixture are found. Equation 6 is used to find the Gaussian model parameters that will maximize  $p(x|\lambda)$  [10].

$$p(x|\lambda) = \prod_{t=1}^T p(x_t|\lambda) \tag{6}$$

The GMM density function is shown in equation 7 [15].

$$p(x) = \sum_{i=1}^N w_i N(x; \mu_i, \Sigma_i) \tag{7}$$

$N$  shows the Gauss density function,  $w_i$ ,  $\mu_i$  and  $\Sigma_i$  show weight, mean and covariance matrix of the Gaussian component  $i$ , respectively. The GMM super-vector consists of the sum of the averages of each Gaussian component [15].

$N$  shows the Gauss density function,  $w_i$ ,  $\mu_i$  and  $\Sigma_i$  show weight, mean and covariance matrix of the Gaussian component  $i$ , respectively. The GMM super-vector consists of the sum of the averages of each Gaussian component [15].

$$\mu = \begin{pmatrix} \mu_1 \\ \mu_2 \\ \vdots \\ \mu_N \end{pmatrix} \tag{8}$$

Each emotion is trained by the spectral properties generated by the GMM super-vectors shown in equation 8.

*E. Hybrid Model (Gauss Mixture Model with combined SVM)*

SVM is a classification algorithm that determines the class of each training vector in high dimensional space. The SVM determines the classes that will determine the support vectors of the data and the output of the hyper plane and the system. At the moment of training, it determined the support vectors by linear, polynomial or sesamoid functions. In this study, linear and polynomial SVM kernels were used for GMM super-vectors.

$$K(x_i, x_j) = x_i^T x_j \tag{9}$$

$$K(x_i, x_j) = (x_i^T x_j + 1)^n \tag{10}$$

The stages of the hybrid model are shown in Figure 5.

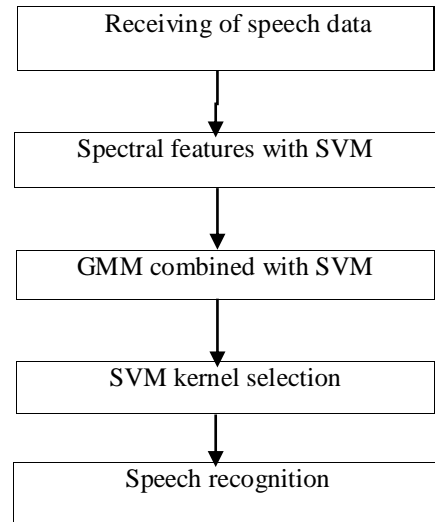


Fig.5. Hybrid model steps

IV. EXPERIMENTAL STUDY

A unique and genuine Turkish language database has been employed in this study. Names, family names, ages, speeches and genders of the persons were added to this database. In this database, the numbers of five senses in males and females as angry, fearful, sad, happy and neutral are shown in Table 1.

Voice samples have been tested by training them, using available feature vectors by means of ANN, HMM ,DTW, GMM and hybrid methods. Success rates of speech samples obtained utilizing are given for ANN, HMM ,DTW, GMM in the Table 2.

TABLE I  
VOICE DATABASE

	Female	Male	Total
Anger	124	241	365
Fear	178	157	335
Sadness	274	256	530
Happiness	179	364	543
Neutral	572	634	1206

TABLE II  
THE SUCCESS OF THE METHODS WITH SVM

	ANN	HMM	DTW	GMM
Male	74.62	75.71	69.97	71.60
Female	75.34	77.63	70.79	72.39

TABLE III  
SUCCESS IN CLASSIFICATION FOR MFCC (5 FEATURE VECTORS)

	ANN	HMM	DTW	GMM
Male	72.64	77.25	67.21	70.25
Female	71.47	75.68	70.24	68.17

Success rates of speech samples obtained utilizing MFCC 5 feature vectors are given for ANN, HMM, DTW and GMM in the Table 3. HMM gave more successful results when compared to other techniques.

TABLE IV  
SUCCESS IN CLASSIFICATION FOR MFDWC (5 FEATURE VECTORS)

	ANN	HMM	DTW	GMM
Male	71.37	75.09	65.38	69.36
Female	70.28	74.34	68.09	68.55

Success rates of speech samples obtained utilizing MFDWC 5 feature vectors are given for ANN, HMM, DTW and GMM in the Table 4. HMM gave more successful results when compared to other techniques.

TABLE V  
HYBRID METHODS FOR DIFFERENT SVM KERNELS

	Linear kernel	Polynomial kernel
Male	76.78	80.67
Female	79.85	81.37

Success rates of speech samples obtained hybrid methods for different SVM kernels (linear, polynomial) in the Table 5. Hibrit Model gave more successful results when compared to all other techniques.

## V. CONCLUSION

Speech recognition and speech emotion recognition plays an important role in our day due to security and many other reasons. Speech emotion recognition of systems have been developed, being based on an unique database obtained by utilizing Turkish language in this study. Classification success of the methods employed in the study have been calculated and results are demonstrated in a comparative manner. Hybrid Method provided more successful results compared to the other speech emotion methods when the results are taken into consideration. This hybrid model has been carried out by combining with SVM and GMM. In first stage of this model, with SVM has been performed subsets obtained vector of spectral features. Hybrid model yielded better results compared to other methods that are used in other literature. Moreover, success rates of speech samples obtained employing MFCC and MFDWC feature vector. Success rates of speech samples obtained employing 5 feature vector. MFCC gave more successful results compared to MFDWC.

## REFERENCES

- [1] Mohammed Shami, Wemen Verhelst, "An evaluation of the robustness of existing supervised machine learning approaches to the classification of emotions in speech", *Speech Communication*, 2007, 49(3), p.201-212.
- [2] Lijiang Chen , Xia Mao, Yuli Xue , Lee Lung Cheng , "Speech emotion recognition: Features and classification models", *Digital Signal Processing*, 22(6), 2012, p.1154-1160.
- [3] Ling He, Margaret Lech, Namunu C. Maddage, Nicholas B. Allen, "Study of empirical mode decomposition and spectral analysis for stress and emotion classification in natural speech", *Biomedical Signal Processing and Control*, 2011, 6(2), p.139-146.
- [4] Tim Polzehl , Shiva Sundaram , Hamed Ketabdar , Michael Wagner and Florian Metzke, "Emotion Classification in Children's Speech Using Fusion of Acoustic and Linguistic Features", *Interspeech 2009: 10th Annual Conference of the International Speech Communication Association*, 2009.
- [5] Halicioglu, Tin Lay Nwe, Foo Say Wei and Liyanage C De Silva, "Speech Based Emotion Classification", *TENCON 2001. Proceedings of IEEE Region 10 International Conference on Electrical and Electronic Technology*, 2001.
- [6] Jasmine Bhaskar, Sruthi Ka and Prema Nedungadi, "Hybrid Approach for Emotion Classification of Audio Conversation Based on Text and Speech Mining", *Procedia Computer Science*, 2015, 46, p.635-643.
- [7] Jinkyu Lee and Ivan Tashev. "High-level Feature Representation using Recurrent Neural Network for Speech Emotion Recognition", *Interspeech 2015*, 2015.
- [8] S.Oh and C.Suen, "A class-modular feed forward neural network for handwriting recognition", *Pattern Recognition*, 2002, 35(1), p.229-244.
- [9] Dimitros and Kontropulos, "Emotional speech recognition: Resources, features, and methods", *Speech Communication*, 2006, 48(9), p.1162-1181.
- [10] D.A. Reynolds and R.C. Rose, "Robust Text-Independent Speaker Identification Using Gaussian Mixture Speaker Models", *IEEE Trans. Speech Audio Proc.*, 1995, 3, p. 72-83.
- [11] Seok, Oh and Ching, Suen, "A class-modular feed forward neural network for handwriting recognition", *Pattern Recognition*, 2002, 35(1), p.229-244.
- [12] Lihang, Li, Dongqing, Chen and Sarang, Lakare etc, "Image segmentation approach to extract colon lumen through colonic material tagging and hidden markov random field model for virtual colonoscopy", *Medical Imaging*, 2002.
- [13] Edmondo, Trentin and Marko, Gori, "A survey of hybrid ANN/HMM models for automatic speech recognition", *Elsevier Neurocomputing* 37, p.91-126, 2001.
- [14] Lindasalwa, Muda and Mumtaj, Began, "Voice Recognition Algorithms using Mel Frequency Cepstral Coefficient (MFCC) and Dynamic Time Warping (DTW) Techniques", *Journal Computing*, 2010, 2(3), p.138-143, ISBN 2151-9617, 2010.
- [15] Hao Hu, Ming-XingXu, and Wei Wu, "GMM Supervector Based SVM with Spectral Features for Speech Emotion Recognition", *Acoustics, Speech and Signal Processing, 2007. ICASSP 2007. IEEE International Conference on*, 2007.
- [16] Cigdem Bakir, "Automatic Speaker Gender Identification for the German Language", *Balkan Journal of Electrical&Computer Engineering*, 2015, 4(2), p.79-83, 2015.
- [17] Cigdem Bakir, "Automatic voice and Speech Recognition System for the German Language", *1st International Conference on Engineering Technology and Applied Sciences*, 2016, p.131-134.
- [18] Lindasalwa Muda, Mumtaj Began and I. Elamvazuthi, " Voice Recognition Algorithms using Mel Frequency Cepstral Coefficient (MFCC) and Dynamic Time Warping (DTW) Techniques", *Journal of Computing*, vol.2, issue 3, p.138-143, ISSN 2151-9617, 2010.
- [19] M., Fahid M. and M.A., "Robust Voice conversion systems using MFDWC", *2008 International Symposium on Telecommunications*, p.778-781, 2008.



#### BIOGRAPHIES

**CIGDEM BAKIR** was born in İstanbul. She received the B.S. degrees in computer engineering from the University of Sakarya, in 2010 and the M.S. degree in computer engineering from Yildiz Technical University, İstanbul, in 2014. Since 2012, she was a Research Assistant with the Yildiz Technical University. She works a Research Assistant with the Iğdir University. Her research interests include recommendation systems, information security, data mining, image processing and biomedical signal processing.



**MECIT YUZKAT** received the B.S. degrees in computer engineering from the University of Trakya and M.S. degrees in computer engineering from the University of Yildiz Technical University. He works a Research Assistant with the Mus Alparslan University. Her research interests include process mining algorithm, data mining, image processing and biomedical signal processing.

# Spectral Ratio Method for Fault Detection in Rotating Machines


J. Dikun, D. Stanelyte, L. Urmoniene

**Abstract**—This study presents the ratio, which is defined between two vibration signals in the spectral domain, to be used in extracting the fault signatures from the signals. These two signals are considered as two different cases of the electric motor of 5 HP in terms of the faulty and healthy motor cases and hence, the comparison between two spectral variations is used as a method to show the fault characteristic. In this manner, the bearing damage of the electric motor of 5 HP are given within the range of 0-4 kHz and its J-curve is presented as an indication of the motor aging.

**Index Terms**—Spectral Analysis, Electric motor, Machinery aging, Fault detection.

## I. INTRODUCTION

ELECTRIC motors play very important role in most industrial application. In this manner, safety operation of the electric motors under the different conditions are highly connected with the process reliability. In order to determine the faulty characteristics of the electric motors, there are two important approaches in the literature. One of them is model-based detection, another one is also signal based approach. In the signal-based approach, the most popular method is spectral method, which is based upon the Fourier Transform [1,2]. Nowadays, another popular method is wavelet transform and its different types in usage [ 2-4]. The wavelet transform based methods or applications are very powerful methods. As an example, Multi-Resolution Wavelet Analysis (MRWA) is in the form of the signal decomposition and hence, the signal to be analysed can be separated to subbands as filter outputs to extract the faulty signal band [3-7]. In addition, Continuous Wavelet Transform (CWT) is another alternative method to indicate the faults. The CWT has an additional property that is defined as redundancy, and hence it is used for the early detection of the faulty cases [3-6].

**J. DIKUN**, is with Department of Electrical and Mechanical Engineering, Klaipada State University of Applied Sciences, Klaipeda, Lithuania, (e-mail: [jeldik@bk.ru](mailto:jeldik@bk.ru)). 

**D. STANELYTE**, is with Department of Electrical and Mechanical Engineering, Klaipada State University of Applied Sciences, Klaipeda, Lithuania, (e-mail: [d.stanelyte@kvk.lt](mailto:d.stanelyte@kvk.lt)) 

**L. URMONIENE**, is with Department of Electrical and Mechanical Engineering, Klaipada State University of Applied Sciences, Klaipeda, Lithuania, (e-mail: [lione.urmoniene@gmail.com](mailto:lione.urmoniene@gmail.com)). 

Manuscript received August 25, 2017; accepted Nov 16, 2017.  
DOI: [10.17694/bajece.419642](https://doi.org/10.17694/bajece.419642)

Meanwhile artificial intelligence based methods, like neural networks and fuzzy logic approaches are used for the feature extraction of the electric motor current or vibration signals [7-9]. Sometimes, instead of these methods, just statistical calculations can be sufficient to detect the faulty cases by means of the statistical parameter variations [6-11].

In this study, as an alternative approach to the fault detection methods defined in the literature above, a simple calculation defined as a ratio between two vibration signals of an induction motor of 5 HP is introduced and used to extract the faulty case. This study has so many advantages over the other classical methods: One of them is feature extraction in frequency domain and the determination of motor aging curve. In addition, the frequency range that is related with the faulty case can be identified depending on the apriory knowledge [11-15].

## II. POWER SPECTRAL DENSITY CALCULATION AND SPECTRAL RATIO

The Fourier transform is used to analyse a time-domain signal [1-3]. Nowadays, depending on developing of the fast computers, the Discrete Fourier Transform (DFT) are easily used for the signal analysis.

For a given data of N-samples, the transform at frequency  $m\Delta f$  is given by following equation:

$$X(m\Delta f) = \sum_{k=0}^{N-1} x[k\Delta t] \exp(-j2\pi Nmk) \quad (1)$$

Here:  $\Delta f$  - is the frequency resolution,

$X(m\Delta f)$ - the DFT of the signal  $x(t)$ ,

The  $Y(m\Delta f)$  becomes of the DFT the signal  $y(t)$ .

The autopower spectral densities (APSDs) of  $x(t)$  and  $y(t)$  are estimated as below:

$$S_{XX}(f) = \frac{1}{N} |X(f)|^2 \quad (2)$$

Where:  $f = m\Delta f$

With the similar way, it is rewritten for the signal  $y(t)$ , as  $S_{YY}$ :

$$S_{YY}(f) = \frac{1}{N} |Y(f)|^2 \quad (3)$$

The statistical accuracy of the estimate in Eqs. (2) and (3) increase as the number of the data points  $N$ . Hence, the spectral ratio can be given by the following equality:

$$R(t) = \frac{S_{YY}}{S_{XX}} \quad (4)$$

III. MEASUREMENT SYSTEM AND APPLICATION

In this study induction motor of 5 HP is used for the test motor. After the accelerated aging studies of the motor [1,2], in order to detect and extract the bearing damage signatures, the following data collection system is used as shown in Figure 1. During this procedure, the sampling rate is selected at 12 kHz and cutoff frequency of the low pass filter used in the signal conditioner unit is at 4 kHz. Also, the bandwidth of the accelerometer is 20 kHz.

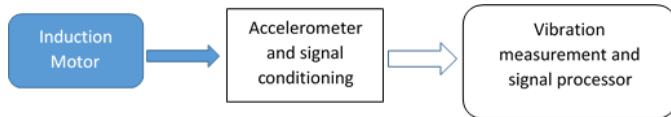


Fig.1. Schematic Diagram for measurement system.

After the accelerated aging tests, motor performance is tested between the healthy motor case (initial case) and faulty motor case (final case) in terms of the vibration signals. In this manner, these vibration signals are given by the following figures, Figure 2 (a) and (b).

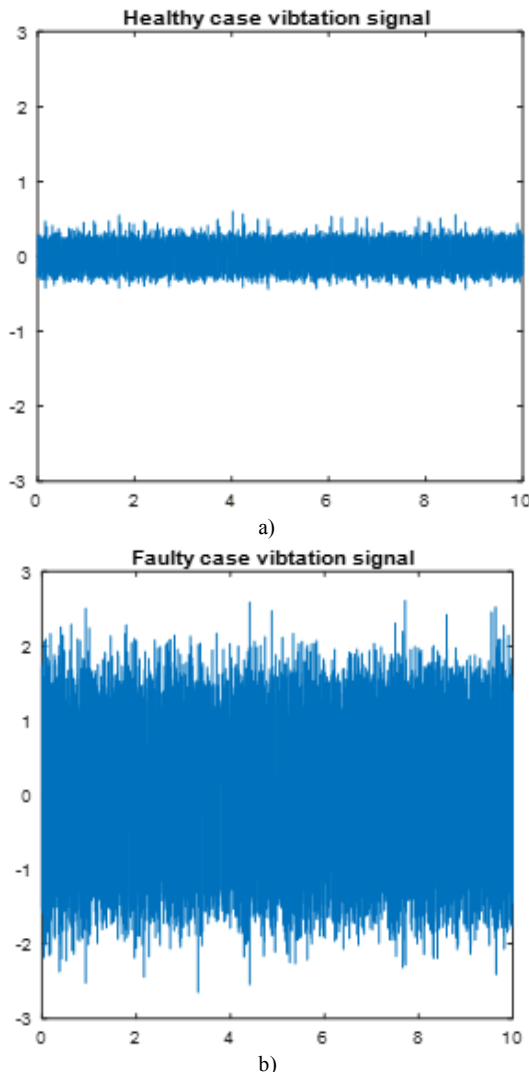


Fig.2. Vibration measurements a) Healthy case, b) Faulty Case.

In addition to the vibration measurements in the time domain, their power spectral density variations can be shown in the frequency domain as follows:

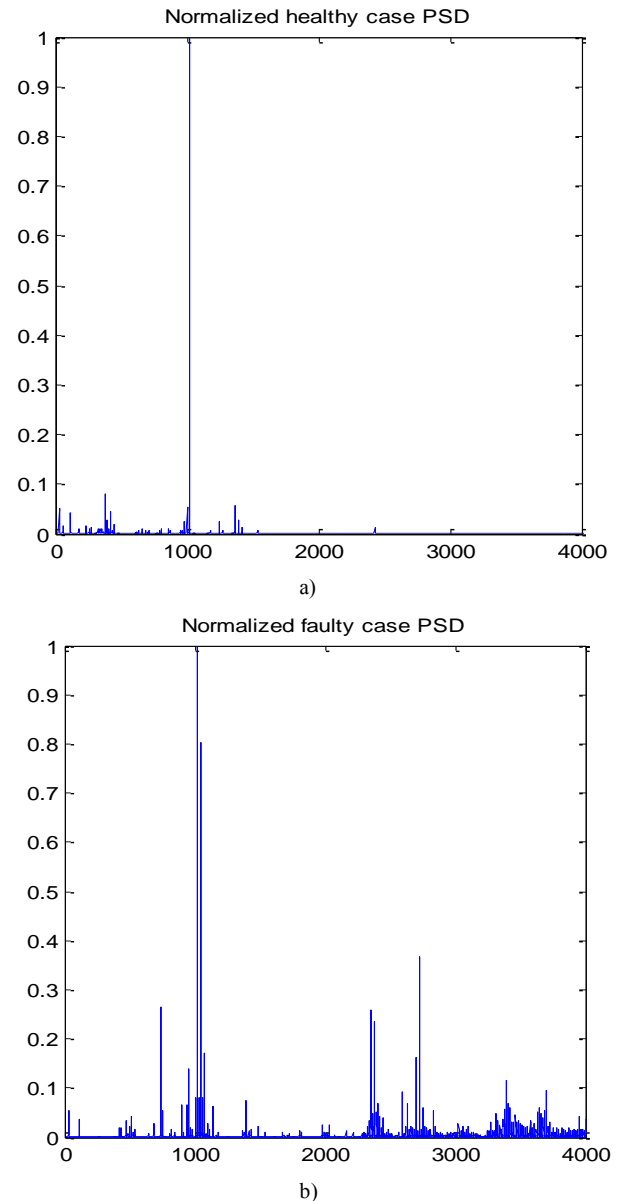


Fig.3. Power Spectral Densities of the Vibration signals a) Healthy case, b) Faulty case.

Comparing the vibration signals, it is seen that there is a increasing in the signal amplitudes and, also, some additional frequencies are appeared between the 2 and 4 kHz as an indicator of the bearing damage occurred at the end of the aging tests. Hence, these are called as the signatures of the bearing damage. After the feature extraction, which is related with the motor bearing damage. On the same data of the power spectral densities, the spectral ration variation that is defined by the Eq. (4) is calculated and it is shown by the following figures as well as aging curve.



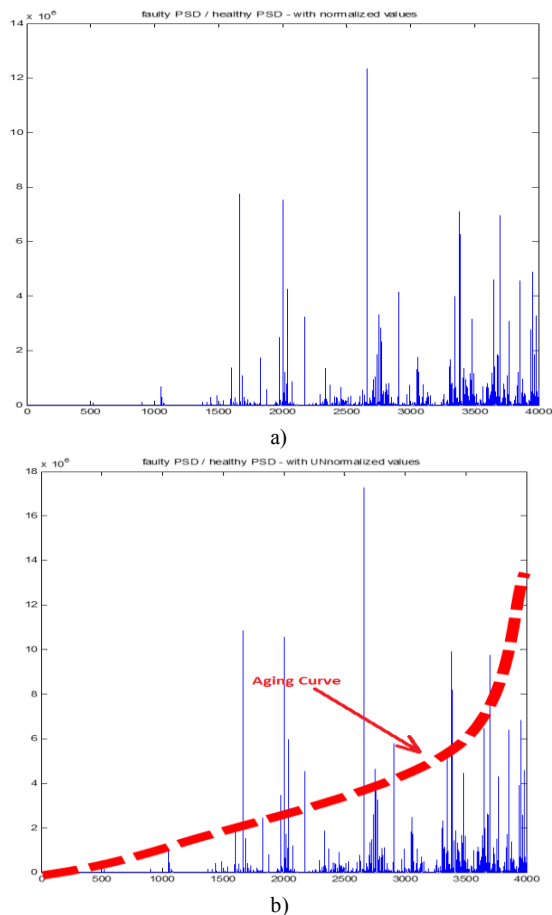


Fig.4. Spectral Ratio of the Vibration spectra measurements (a) and aging curve (b).

#### IV. CONCLUSION

In this study, some characteristics based upon bearing damage of the induction motor of 5 HP was examined and its aging curve was extracted from the experimental data. As seen in the figure (4), the upper critical frequencies of the motor in terms of the safety operation conditions can be indicated at around of the turning point of the aging curve after 3.5 kHz. This is also expected situation because the bearing damage was characterized between 2 and 4 kHz.

#### ACKNOWLEDGEMENT

Authors present their deep thanks to Prof. Dr. Serhat Şeker for his valuable contributions in the meaning of teaching the signal based diagnostic methods of the rotating machines and for his helps in the data providing. This study was done under the ERASMUS staff exchange program collaboration between the ITU (Istanbul Technical University) and KSUAS (Klaipeda State University of Applied Sciences).

#### REFERENCES

- [1] Şeker S., Ayaz E., A Study on Condition Monitoring for Induction Motors Under the Accelerated Aging Processes, IEEE Power Engineering Review, V.22, N.7, pp.35-37, July2002.
- [2] Seker, S; Ayaz, E Feature extraction related to bearing damage in electric motors by wavelet analysis Journal of Franklin Institute-Engineering and Mathematics, 340 (2), 2003, pp.125-134.

- [3] Ozturk A.; Seker S. "On the Frequency Resolution of Improved Empirical Mode Decomposition Method" International Review of Electrical Engineering-IREE, Vol. 5, No. 4, pp. 1798-1805, Part b, 2010.
- [4] Seker, S; Ayaz, E; Turkcan, E Elman's recurrent neural network applications to condition monitoring in nuclear power plant and rotating machinery Engineering Application of Artificial Intelligence, 16 (7-8): pp.647- 656 Oct-Dec 2003
- [5] Senguler Tayfun; Karatoprak Erinc; Seker Serhat "A New MLP Approach for the Detection of the Incipient Bearing Damage", Advances in Electrical and Computer Engineering, Vol.10, No. 3, pp. 34-39, DOI: 10.4316/AECE.2010.03006, 2010.
- [6] D. Sonmez, S. Seker, M. Gokasan, "Entropy-based fault detection approach for motor vibration signals under accelerated aging process" Journal of Vibroengineering Paper # 851, Vol.14, No.3, September 2012.
- [7] D. Bayram, S. Şeker, " Redundancy Based Predictive Fault Detection on Electric Motors by Stationary Wavelet Transform", IEEE, Transaction on Industrial Applications, vol. 53, pp.2997-3004, 2017.
- [8] A.H. Bonnet and G.C. Soukup, "Cause and Analysis of Stator and Rotor Failures in Three Phase Squirrel-Cage Induction Motors", IEEE Transactions on Industry Applications, Vol.28, No.4, pp. 921-937, August 1992.
- [9] K.R. Cho, J.H. Lang, and S.D. Umas, "Detection of Broken Rotor Bars in Induction Motors Using State and Parameter Estimation", IEEE Transaction on Industry Applications, Vol.28, No.3, pp. 702-709, May/June 1992.
- [10] S.W. Bowers, K.R. Piety, and R.J. Colsher, "Evaluation of the Field Application of Motor Current Analysis", Proceedings of the Meeting of the Vibration Institute, 1993.
- [11] R. Schoen, T.G. Habetler, F. Kamran, and R.G. Bartheld, "Motor Bearing Damage Detection Using Stator Current Monitoring", 1994 IEEE Industrial Application Meeting, Vol.1, pp.110- 116, 1994.
- [12] M.J. Costello, "Shaft Voltages and Rotating Machinery," IEEE Transaction on Industry Applications, Vol. 29, No. 2, pp. 419-425, 1993.
- [13] S.V. Bowers and K.R. Piety, "Proactive Motor Monitoring Through Temperature Shaft Current and Magnetic Flux Measurements", CSI 1993 Users Conference, September 20-24, 1993, pp.2-3.
- [14] J.R. Nicholas, "Predictive Condition Monitoring of Electric Motors", P/PM Technology, pp. 28-32, August 1993.
- [15] G.A. Bisbee, "Why Do Motor Shaft and Bearing Fail", TAPPI Journal, Vol. 77, No. 9, pp. 251-252, September 1994.

#### BIOGRAPHIES



**Jelena DIKUN** Received the BS Degree in Electrical Engineering Science from Klaipeda University in 2010, the MS Degree in Electrical Engineering Science from Klaipeda University in 2012. Currently she is working as a lecturer in Lithuanian Maritime Academy as well as in Klaipeda State University of Applied Sciences. Her primary research interests include electromagnetic radiation, ships' electrical systems, electrical machines and apparatus, electrical measurements technologies.



**Lione URMONIENE** Received the PhD Degree in Electrical Engineering Science from Kaunas University in 2012. Currently she is working as an Assistant Professor in Klaipeda University as well as in Klaipeda State University of Applied Sciences. Her scientific interests includes oscillated electrical machines, electrical drives, drives control systems and automated drives control systems.



**Daiva STANELYTE** Received the MS Degree in Informatic Engineering Science from Klaipeda University in 2010. Currently she is a PhD student of Kaunas Technology University at Energy and Thermo Engineering Faculty. At present time, she is working as a head department of Mechanical and Electrical Engineering in Klaipeda State University of Applied Sciences. Her field of scientific research are industry systems automation and energy production, distribution and consumption.

# Investigating the Impact of Team Formation by Introversion/Extraversion in Software Projects

V. Garousi, and A. Tarhan

**Abstract**—Human factors have an important effect on performance of software teams and resulting software products. One of the seldom-studied aspects of human factors is the effect of personality-based team formation on team cohesion and quality of the software product. In this study, we investigate the above effect by conducting an exploratory case study during a term-long undergraduate software engineering course containing a project component with 50 undergraduate students. We grouped the students based on the social-interaction dimension (introversion/extraversion) of the well-known Myers–Briggs Type Indicator (MBTI) personality assessment model. We then collected the relevant metrics to explore/analyze the two parameters of interest in our study: team cohesion, and project grade as an indicator of project output (i.e. resulting product quality). Our results show that there is some (although weak) relationship between the team formation scheme (based on either introversion or extraversion) with group performance and project grade. The results also show that mixed grouping of personality types has no significant effect on team cohesion but is advantageous in achieving higher project grades especially for people with low GPAs.

**Index Terms**—human factors, team formation, MBTI, team cohesion, product quality, software projects, empirical study.

## I. INTRODUCTION

Software Engineering (SE) is a team activity by nature, and human and social factors have a strong impact on the success of any SE endeavor and the software product developed by software teams [1]. In the pursuit of more effective and efficient software development, software teams must be composed of people who work well together. How to properly form these teams, the interaction between team members, and how individual personalities influence performance and software quality, have been among the important concerns in the SE field from the 1960s to the present day [2]. Many leading figures in the field have claimed that it is fundamentally people that make the difference between success or failure of software projects [3].

**V. GAROUSI** is with the Information Technology Group of Wageningen University, the Netherlands (e-mail: [vahid.garousi@wur.nl](mailto:vahid.garousi@wur.nl)).

**A. TARHAN** is with Department of Computer Engineering, Hacettepe University, Ankara, Turkey (e-mail: [atarhan@hacettepe.edu.tr](mailto:atarhan@hacettepe.edu.tr)).

Manuscript received August 25, 2017; accepted Nov 16, 2017.

DOI: [10.17694/bajece.419645](https://doi.org/10.17694/bajece.419645)

Building effective software teams that would lead to project success, however, is not trivial [2]. Understanding human aspects in SE teams is crucial because having the right people in a team can *make* or *break* a project. Thus, there is a need to explore factors that bind team members and understand the elements that enable effective team performance. In this context, various factors such as personality types and skill levels should be taken into account. While there exist a large body of research in this area, e.g. [4-12], there is a need for more empirical evidence and in-depth studies which look into each personality-related factors in more detail, e.g., introversion and extraversion.

In this study, we aim at assessing the impacts of personality-based team formation on team cohesion and project output, from the viewpoints of researchers and practitioners. We conducted an exploratory case study during a term-long undergraduate SE course containing a project component with 50 students. We first assessed personality types using the widely-used Myers–Briggs Type Indicator (MBTI) [13, 14], which is the most commonly used model in SE literature [15]. For team formations, we considered the introversion/extraversion dimension of MBTI. We then investigated the effects of team formation on team cohesion and project output. The results of our case study provide insights for practitioners and can be useful when building software teams.

The remainder of this paper is organized as follows. Section II discusses the background and related work. Section III describes our research method. Section IV presents the results of the study. Section V summarizes the findings, implications and limitations of our study. Finally, Section VI concludes this study and states the future work directions.

## II. BACKGROUND

### A. Team Related Factors

The most related body of work to our study are the empirical studies about team-related factors in SE. From the large set of such studies, we have sampled a list as shown in Table I. For each study, we show the publication year, paper title, and the independent and dependent variable(s) studied in the study.

From the list of possible independent and dependent variables that are worthy of investigation, some have been studied in the previous work as listed in Table I. In this study, we focus on personality-based team formation as the independent variable and team cohesion and project output as dependent variables. To the best of our knowledge, our study is the first one focusing on this particular combination of independent and dependent variables.

TABLE I. EMPIRICAL STUDIES IN SE STUDYING TEAM-RELATED FACTORS

Ref. & Year	Paper title	Independent variable(s)	Dependent variable(s)
[4] 2005	Examining team cohesion as an effect of software engineering methodology	Software engineering methodology	Team cohesion
[5] 2006	A follow up study of the effect of personality on the performance of software engineering teams	Personality	Team performance
[6] 2009	How do personality, team processes and task characteristics relate to job satisfaction and software quality?	Personality, team processes and task characteristics	Job satisfaction and software quality
[7] 2010	Analyzing personality types to predict team performance	Personality types	Team performance
[9] 2010	Software engineering group work: personality, patterns and performance	Personality of group members	Using design patterns, learning achievements
[8] 2013	A worked example of the relations between personality and software team processes	Personality	Team process
[10] 2014	A mixed methods investigation of ethnic diversity and productivity in software development teams	Ethnic diversity	Productivity, innovation and problem solving
[11] 2014	A replicated quasi-experimental study on the influence of personality and team climate in software development	Team cohesion and conflict	Team performance
[12] 2015	Are team personality and climate related to satisfaction and software quality? Aggregating results from a twice replicated experiment	Team personality and climate	Satisfaction and software quality

TABLE II. DIMENSIONS IN MBTI PERSONALITY ASSESSMENT

<b>Introversion/ Extraversion (I/E)</b>	<i>The extraverted types</i> learn best by talking and interacting with others; and by interacting with the physical world, they can process and make sense of new information. <i>The introverted types</i> prefer quiet reflection and privacy; and information processing occurs as they explore ideas and concepts internally.
<b>Sensing/ Intuition (S/N)</b>	<i>Sensing types</i> enjoy a learning environment in which the material is presented in a detailed and sequential manner. They attend to what is occurring in the present, and can move to the abstract after they have had the experience. <i>Intuitive types</i> prefer a learning atmosphere in which an emphasis is placed on meaning and associations; they value insight higher than careful observation, and naturally recognizes patterns in work.
<b>Thinking/ Feeling (T/F)</b>	<i>Thinking types</i> desire objective truth and logical principles and are natural at deductive reasoning. <i>Feeling types</i> place an emphasis on issues and causes that can be personalized while they consider other people's motives.
<b>Judging/ Perceiving (J/P)</b>	<i>Judging types</i> thrive when information is organized and structured, and they are motivated to complete assignments in order to gain closure. <i>Perceiving types</i> flourish in a flexible learning environment in which they are stimulated by new and exciting ideas.

### B. MBTI Personality Assessment

The Myers–Briggs Type Indicator (MBTI) is a popular tool for personality assessment, which was developed based on the theories of Carl Jung [16]. It serves as an introspective self-report questionnaire designed to indicate psychological preferences in how people perceive the world and make decisions [13, 14]. There are four dimensions in this indicator as shown in Table II. MBTI has been widely used in different research communities, e.g., social sciences, psychology and SE [15]. Various studies have appeared on the usage of MBTI and other personality tests in SE, e.g., [3, 15, 17]. According to a systematic literature review (SLR) on personality assessment in SE [15], MBTI is the most commonly used model in the SE literature. For all the above reasons, in this study, we selected MBTI as the personality assessment model.

### C. Team Building and Personality Types

There are many discussions and studies in other domains which report collaborations among extraverts and introverts could be challenging [18, 19], e.g.: “*Extraverts can think that introverts are slow, have few ideas to share and are unemotional. They interpret those calm faces as meaning introverts lack in emotion and passion. Introverts think that extraverts are shallow because they talk a lot. Not being direct and concise can be seen as lacking depth.*” [19].

It was argued in [2] that personality-type analysis could help take the guesswork out of putting together a high-performance software project team. The authors invited 92 Information Systems (IS) professionals from 20 software development teams in Hong Kong to complete a questionnaire-based survey. The surveys showed how team leaders scored on the information gathering dimension (sensing/intuitive) had a significant impact on team performance. Only the decision-making dimension (thinking/feeling) of the systems analyst personality had a significant influence on team performance. Only the social-interaction dimension (introversion/extraversion) of the programmer personality was strongly related to team performance. Among the conclusions were that it was unnecessary to have diversity of personalities among team members (excluding team leader) due to the fact that members needed to perform multiple tasks of the software development life cycle and that heterogeneity was not good for all phases.

The study in [20] examined the relationships between the ‘Big Five’ personality factors (conscientiousness, extraversion, neuroticism, agreeableness, and openness to experience) and objective team performance, and derived implications for selecting successful product teams. Successful teams were characterized by higher levels of general cognitive ability, higher extraversion, higher agreeableness, and lower neuroticism than their unsuccessful counterparts.

The study in [21] proposed a formal model for assigning human resources to teams in software projects. Using the Delphi method, the authors proposed a set of software project roles and competencies. Psychological tests and data mining tools identified useful rules for forming software project teams. These were used to build a formal model, which was later built into a tool that automatically calculated role assignments. This decision-support tool was claimed to help managers in assigning people to roles and forming software teams. The model was validated by assignment scenarios in two software development organizations.

The study in [22] presented a mix-method replicated study for team building in software projects. The findings indicated that carefully selecting team members for software teams was likely to positively influence the projects in which these teams participate. Besides, it seemed that the type of development method could moderate (increase or decrease) this influence.

The study in [23] discussed a comparison of the performance of student groups formed randomly, with those formed by using the learning styles questionnaire. The study found no significant differences in the performances of these two sets of groups, for which it discussed several possible reasons.

#### D. Team Cohesion

One of the most studied team variables in literature is “team cohesion”. Team cohesiveness is the degree to which team members like each other, identify themselves positively with the team and want to remain with its members [24]. It reflects the degree of attraction among group members. A study of cohesiveness is considered essential for understanding group dynamics in teams. Two meta-analyses in the psychology discipline [25, 26] have reported a positive relationship between cohesiveness and performance. According to these studies, cohesive teams demonstrate increased collective efficacy and greater team success. Furthermore, cohesive team members are less anxious, more satisfied, have higher self-esteem, conform to group norms, make personal sacrifices for the team, share responsibility for team failure and are less likely to indulge in social loafing.

### III. RESEARCH METHOD

#### A. Goal and Research Questions (RQs)

The goal of our study was to conduct an exploratory evaluation on impacts of personality-based team formation on team cohesion and also quality of project output (software). To focus the study on one independent variable and prevent the impact of more than one independent variables (the so called “confounding bias”), we focused on only one dimension of the MBTI – the social-interaction dimension (introversion and extraversion). Based on the stated goal, we raised the following two research questions (RQ):

- RQ 1 - What are the impacts of personality-based team formation on team cohesion?
- RQ 2 - What are the impacts of personality-based team formation on project output (i.e. resulting product quality)?

#### B. Research Design

We designed our research approach by adapting the Goal, Question, Metric (GQM) methodology [27]. We replaced the questions with RQs, and identified independent and dependent variables as the metrics to be used in our research to answer the RQs. The design that we developed is shown Table III.

TABLE III. GQM DESIGN OF OUR RESEARCH APPROACH

Goal: To conduct an ‘exploratory’ evaluation on impacts of personality-based team formation on team cohesion and project output		
RQ1: What are the impacts of personality-based team formation on team cohesion?		
Independent var.	M1: MBTI social interaction personality type (I/E)	
	M2: Students’ grade point average (GPA)	
Dependent var.	M3: Team cohesion morale index (TCMI)	
RQ2: What are the impacts of personality-based team formation on project output?		
Independent var.	M1: MBTI social interaction personality type (I/E)	
	M2: Students’ grade point average (GPA)	
Dependent var.	M4: Project grade (as an indicator of project output)	

For personality-based team formation, we needed a suitable metric for assessing students’ personality types. We instructed the students in the beginning of the semester to take MBTI using a free online test [14] for identifying their own personality types of social interaction dimension (introversion/extraversion). We also gathered students’ latest cumulative

grade point average (GPA) in their program, and treated the GPA as indicators of their technical abilities. We used MBTI social interaction personality types and GPAs of students as independent variables in our research.

Metrics for the dependent variables included the ones for measuring the team cohesion and the project output. For quantitatively measuring team cohesion, we searched in both the formal and the grey literature, and selected a rubric set [28] developed by an Agile practitioner and coach. This rubric is used by Agile practitioners to quantitatively measure team morale in Agile teams in the industry. The rubric, which is shown in Table IV, has been developed using the rigorous foundations from the psychology literature [29], and consists of eight questions. The answer of each question is based on a 5-point Likert scale as follows: {1: Very low, 2: Low, 3: Average, 4: High, 5: Very high}. To quantitatively calculate the team morale of an individual member, the average value of the scores on the eight questions is calculated and set as the Team Cohesion and Morale Index (TCMI). An example calculation is shown in Table IV.

Note that while Likert scale is originally an ordinal scale, analyzing Likert scales as interval values (and calculating average based on such data) is possible when the sets of Likert items can be combined to form indexes, with the caveat (assumption) that “*this combination forms an underlying characteristic or variable*” [30]. Also, we ensured precise wordings for the five response levels above to clearly imply “*a symmetry of response levels about a middle category*” [31]. Therefore, equal spacing of response levels was clearly indicated, and the argument for treating it as interval-scale data was supported [31].

TABLE IV. 8-QUESTION RUBRIC USED TO MEASURE TEAM COHESION AND MORALE (ADOPTED FROM [28])

#	Question	Sample values
1	I am enthusiastic about the work that I do for my team	1: Very low
2	I find the work I do for my team of meaning and purpose	4: High
3	I am proud of the work that I do for my team	3: Average
4	To me, the work that I do for my team is challenging	2: Low
5	In my team, I feel bursting with energy	5: Very high
6	In my team, I feel fit and strong	4: High
7	In my team, I quickly recover from setbacks	3: Average
8	In my team, I can keep going for a long time	1: Very low
	Team Cohesion and Morale Index (TCMI)=	Avg.=2.8

For quantitatively measuring project output, we used project grade as an indicator of resulting project output. We calculated project grades based on the grades of deliverables planned and submitted by students throughout project life cycle. Students delivered project artifacts at five milestones: (0) Vision and project plan, (1) Requirements document, (2) Design document, (3) Demo of the prototype version of the software, and (4) Final software product. Students were asked to submit team cohesion (morale) via an online questionnaire in each milestone (0-4). Project teams had to explain their team work and the work of each student individually in project reports which were then used for marking the works of teams and the students. The teaching team assessed the students’ works in each delivery, and used the following rubric to reduce subjectivity in marking: Functional quality (% of test cases passed), code readability,

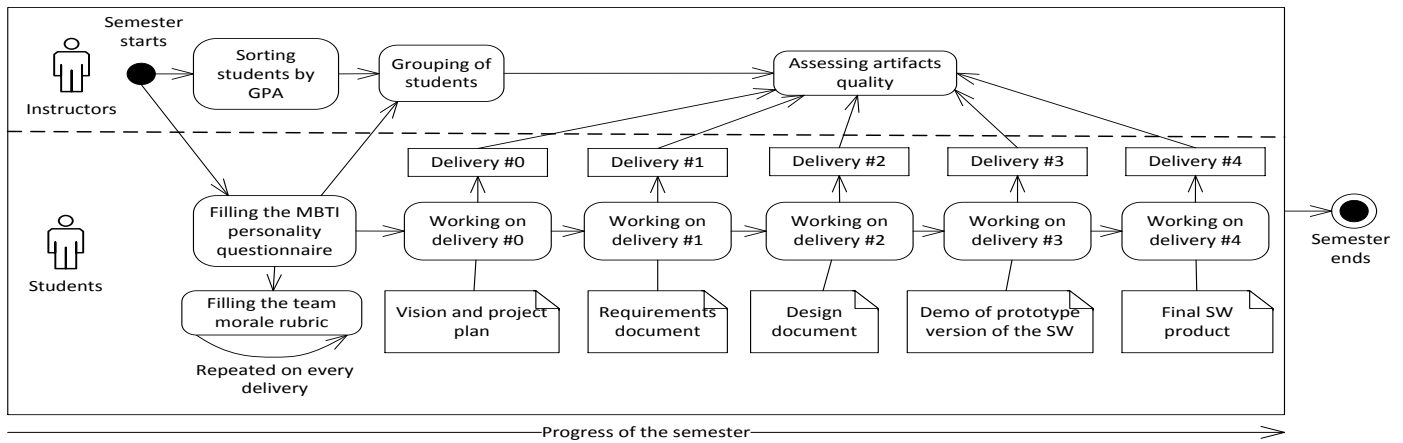


Fig. 1. Activity diagram showing the planning and execution of the empirical study.

and extent and quality of documentation. Although the authors had intended to use a more detailed rubric for evaluating deliveries, they had to simplify the rubric to include several key factors above due to shortage of human resources.

Fig. 1 depicts an activity diagram showing the planning and execution stages of our empirical study. For its design and execution, we took into consideration the recommendations on using students in empirical studies (e.g., [32, 33]) and also received ethics approval from Hacettepe University.

The teaching team consisted of two instructors (the co-authors of this paper) plus two teaching assistants (TAs). As shown in Fig. 1, in the beginning of the semester, the instructors sorted the students by their GPAs and used the MBTI personality data to group the students. The approach that we took for grouping is explained in the next subsection.

C. Study Subjects and Team Formation Approach

The study was conducted in the context of ‘Software Engineering Laboratory’ course, which is the practical counterpart of the 3<sup>rd</sup> year ‘Software Engineering’ course, in Hacettepe University’s Department of Computer Engineering. During the Spring 2016 offering of the course, in which the study was conducted, the course had exactly 50 students.

An important issue was to decide how to group students (i.e. form teams). As per the study’s goal (impacts of personality-based team formation by social-interaction dimension of the MBTI), we sorted the students by their GPAs and then grouped each three students into one team such that, in a controlled manner, a group with the closest GPAs would have all introvert members, another group would have all extravert students, and another one having a mix of introverts and extraverts. Grouping would continue from students with higher GPAs towards those with lower GPAs until every student belonged to a group. This grouping mechanism would yield a set of groups with similar GPAs in which the only differentiating factor would be the extraversion and introversion attitudes.

As shown Fig. 1, we instructed all the students in the beginning of the term to take an online MBTI test [14] and send their results to the instructors. Once we had the MBTI assessment results (types) and the GPAs, we used the grouping approach discussed above to form the groups. Fig. 2 shows the

results of grouping process. We set the group sizes to three students, except the very last group (which had five students). As a result, 16 groups were formed, shown as #1 ... #16 in the figure. As per our grouping approach, all three members of group #1 were extraverts (labeled as ‘All Ext’ in Fig. 2). All three members of group 2 were introverts. Group 3 was a mix of extraverts and introverts, etc.

Based on the MBTI data, at the end, we had three all-introvert groups. Nine groups had all extraverts, and four groups were mixes of introverts and extraverts. We would have liked to have a balanced mix of the three group types, but the MBTI data of students did not allow this (i.e., we did not have as many introverts as we wanted to put in the groups).

Fig. 3 shows a dot-plot of the distribution of the GPA values of the 50 students in the class. GPA values were out of 4 and corresponded to the performance of the students in their previous school terms. The values were taken from the student records. The mean (average) was 2.55. The minimum and maximum were 1.11 and 3.45, respectively.

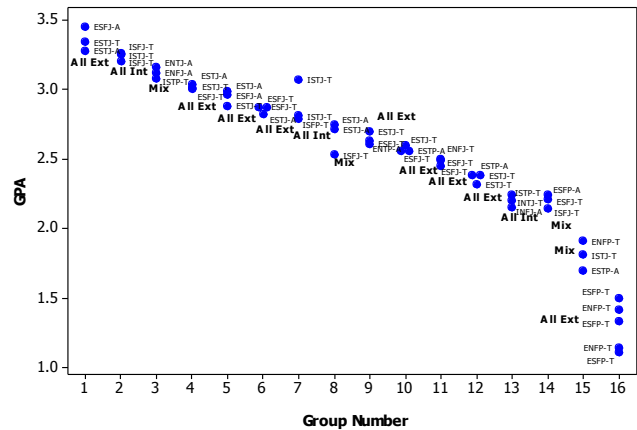


Fig. 2. GPAs and MBTI social interaction types of the students in 16 groups

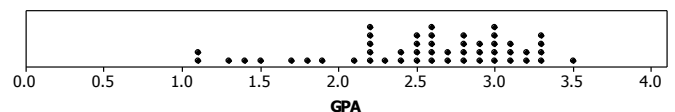


Fig. 3. Dot-plot of the GPA values of the 50 students in the class

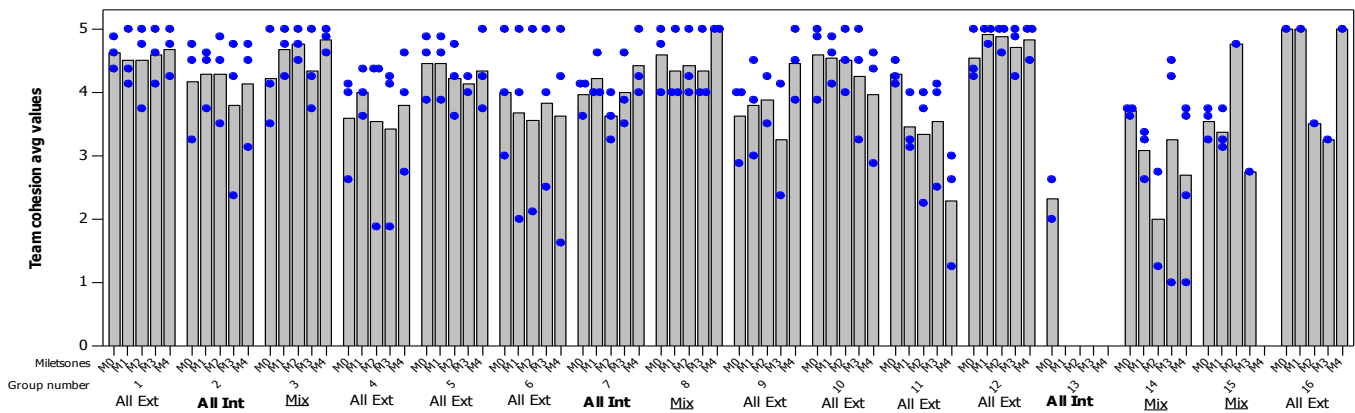


Fig. 4. Team cohesion average values reported by each group member (student) in each milestone

#### D. Study Objects and Project Development Context

The project was to develop an online library management software. Students were provided with the high-level requirements of the system written in English, and in a UML (Unified Modeling Language) use-case diagram.

Students were asked to use the Open Unified Process (OpenUP) development process and its artifacts' templates [34]. As shown in the design of the empirical study (Fig. 1), the development project had five milestones and students submitted various software artifacts (documentation or code) in each of the steps, as per the OpenUP's specifications. We asked for artifacts for the following phases: requirements, architecture, implementation, testing, and project management.

For requirements and design stages, OpenUP requires modeling by using UML. To establish consistency in the entire class among all student groups, students were asked to use the Visual Paradigm UML tool [35].

### IV. RESULTS

#### A. RQ1: Team Formation and Team Cohesion Morale Index

Our rationale behind RQ1 was to assess the implications and outcomes of team dynamics and to assess the impacts of team formation (if any) on team cohesion and morale as measured by Team Cohesion and Morale Index (TCMI) metric. Fig. 4 shows, as an individual-value plot, the team cohesion average values reported by each group member (student) in each milestone. The bars show the average values of the individual values for each milestone, e.g., M0 thru M4. Data for group 13 was not available since the group decomposed soon after the term had started. Please note that the team who abandoned the class was one of the teams with the lowest GPAs, which explains their decision to drop the course. Only two TCMI values for M0 for this group were reported. This did not lead to a negative effect on our case study since our group formation took into account academic success, and our design had a preventive nature against such occurrences as previously mentioned. Recall from Section III.C that groups were sorted by descending order of GPAs, e.g., members of group #1 had the highest GPAs and those in group #16 had the lowest GPAs.

For ease of review and analysis, we have also included the types of groups (either all extraverts, all introverts, or mixed) in Fig. 4 (below the group numbers). As we could observe, grouping based on introversion/extraversion 'alone' did not

have any noticeable impact on team cohesion, as groups with all extraverts, all introverts, or mixed all reported different levels of the TCMI measured, regardless of group formation types. One expectation in this context could have been that, in groups with homogeneous (compatible) team members (all extraverts or all introverts), TCMI measure would be higher than in groups with mixes of extraverts and introverts, since mixed groups could have higher chances for arguments and disagreements, thus leading to lower TCMI measure.

We found no significant correlation between the social-interaction dimension (introversion/extraversion) and team cohesion (i.e. TCMI values). This observation is similar to the findings of the study in [11] in which no significant correlation between the extraversion personality factor and team satisfaction was found.

We also investigated whether there was any correlation between TCMI values reported by each student and her/his GPA (i.e., whether students with higher GPAs felt better team cohesions). Fig. 5 shows the scatterplot of these values for all students. The Pearson correlation of the two datasets is 0.24 (p-Value = 0.12) – thus, showing a weak correlation, meaning that for a student with higher technical capabilities, it would be expected for her/him to have a higher perception of team morale and team cohesion feelings; and vice versa.

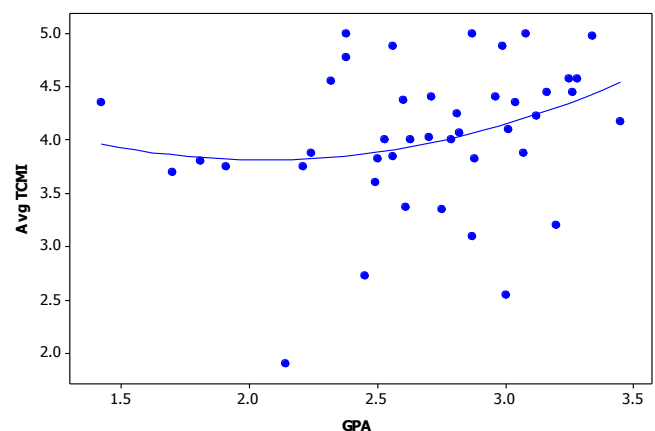


Fig. 5. Scatterplot of TCMI values reported by each student and her/his GPA

#### B. RQ2: Team Formation and Project Grade

As the response to RQ2, we discuss the impacts of team formation on project grade as an indicator of resulting project

output. Fig. 6 shows the individual-value plots of (a) GPA and (b) project grades with respect to team formation, i.e., all extraverts (All Ext) – 29 students, all introverts (All Int) – 9 students, and mixed (Mix Ext-Int) – 12 students.

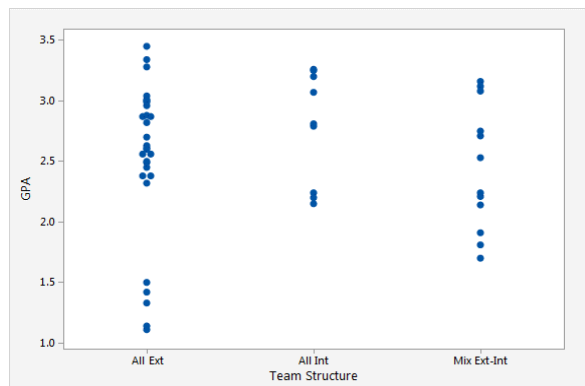
As Fig. 6 (a) shows, students in the teams of all-extraverts mostly had high GPAs (between 2.3 and 3.45), though a small number of such students had GPAs below 1.5. It is also seen from the figure that students in the teams of all-introverts had GPAs above 2.15. When it comes to students in the mixed teams of extraverts and introverts, we see that GPAs were distributed between the values of 1.7 and 3.16. In addition, the students in all-extraverts-teams had values on the edges of the GPA scale while the range for the students in all-introverts-teams was smaller in variance (between 2.1 and 3.2).

Fig. 6 (b), on the other hand, represents the individual-value plots of project grade for the three team types. For the teams having all extraverts, most of the students (75%) received grades above 65% and that four out of 29 students failed. The teams having all introverts either were very successful (having grades above 93%) or performed very poorly (in three out of nine groups). The mixed teams, interestingly, were generally

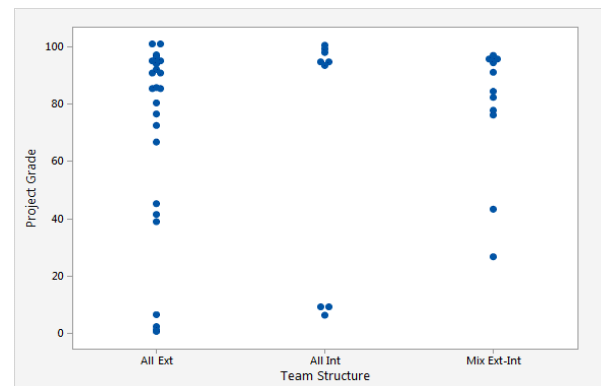
high-performers (with two exceptions) with their grades between 76% and 97%.

To evaluate the data in Fig. 6 (a) and (b) together, and to better understand the relationships between GPAs and project grades for different team types, we sketch in Fig. 7 the average values of project grades versus the average values of GPAs of the teams with respect to the three team types. The figure shows the relation between project grades and GPAs of group members per team formation type. It can be observed that the groups with lower GPAs failed except the mixed ones, i.e., having both extraverts and introverts. It seems mixed grouping of personality types worked better than discrete grouping of all extraverts or all introverts, in terms of achieving higher project grades, especially for students with low GPAs.

We also investigated the relationship between project grades and GPAs of the students by their personality type. Fig. 8 shows the distributions of data points, with their best-fit-curves, for the types of extraverts (Ext) and introverts (Int). Coincidentally, best-fit-curves are fully overlapping, denoting that there was no statistically-significant difference in project grades between the students from the two personality types (extraverts and introverts).



(a) Plot of GPA values



(b) Plot of project grades

Fig. 6. Individual-value plots of GPAs and project grades of students by team structure

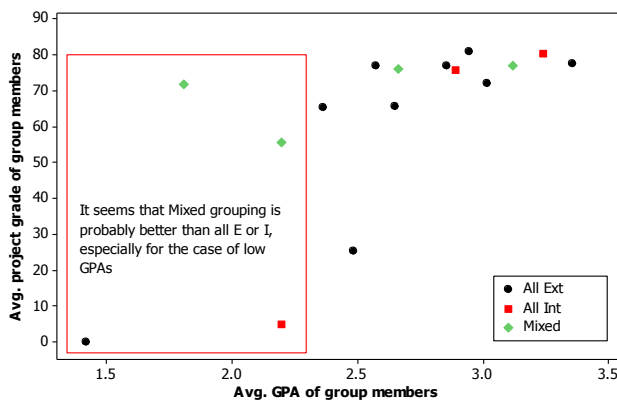


Fig. 7. Avg. project grade versus avg. GPA of teams by team formation

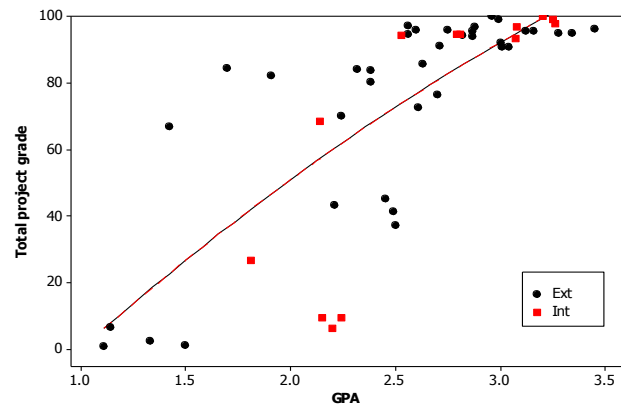


Fig. 8. Project grade versus GPA, grouped by personality types: E vs. I

## V. DISCUSSION

### A. Summary of Findings

RQ1 was intended to investigate the implications and outcomes of team formation on team cohesion and morale as measured by the Team Cohesion and Morale Index (TCMI) metric. The results for RQ1 showed that grouping based on the social-interaction dimension ‘alone’ (introversion vs. extraversion) did not have any noticeable impact on team cohesion, as groups with all extraverts, all introverts, or mixed types reported different levels of TCMI values, regardless of the group formation types. This observation was similar to the findings of the study [11] in that no significant correlation between the extraversion personality type and team satisfaction was found. We also examined if there was any correlation between TCMI values reported by each student and her/his GPA (i.e. if the students with higher GPAs felt better about team cohesions). We noticed a weak correlation between these two variables, possibly meaning that the higher the technical capabilities of a developer, the stronger his/her feelings of team morale and team cohesion would be; and vice versa.

RQ2 was aimed to understand the impacts of team formation on project grades as an indicator of resulting project output. The results for RQ2 showed that, for the teams having all extraverts, most of the students received grades above 65% and that only few students in such groups failed. Students in the all-introvert groups were either very successful (with grades above 93%) or performed very poorly. The mixed teams, interestingly, were generally high-performers. That is, mixed grouping of personality types worked better than discrete grouping of all extraverts or all introverts, in terms of project grades, especially for people with low GPAs (i.e. low technical abilities).

### B. Potential Threats to Validity

We discuss the limitations and potential threats (construct, internal, conclusion, external) [36] to the validity of our study and the steps that we took to minimize or mitigate them in the following paragraphs.

*Construct validity* is concerned with issues that to what extent the study truly represented the theory behind it [36]. The potential issues in this regard were whether we properly conducted personality-based team formation, and actually measured team cohesion and project grades. Adapting the GQM methodology [27] for our research design and standardizing the metrics and the instruments used in this study, we addressed those issues and minimized the associated threats. However, threats might have remained regarding the variability of MTBI test results depending on the mood of the students while answering the questions, and the percent rating scheme of the questionnaire, e.g., introversion/extraversion scores could be close (e.g. 49% vs. 51%) or far apart (e.g. 1% vs. 99%). Also, the team formation approach that we used based on the students’ GPAs might be considered as another threat, as we put the best students together, average students together, and not-so-good students together. Still, we adopted this grouping mechanism because it resulted in a set of groups with similar GPAs, which was important to keep the social interaction dimension (i.e. extraversion/introversion) the only differentiating factor in our research design.

*Internal validity* reflects the extent to which a causal conclusion based on a study is warranted [36]. To prevent confounding bias, we focused only on the social-interaction dimension (introversion/extraversion) of the MBTI model, and thus prevented the likely impact of more than one independent variables. In terms of selection bias, the subjects of the study were composed of 50 undergraduate students who had enrolled in the ‘Software Engineering Laboratory’ course. To prevent any negative influence, we considered the recommendations on using students in empirical studies (e.g., [32, 33]) in the design of the study, and had ethics approval from our university. While the subjects (i.e. the students) were not yet software engineers, they had very similar profiles. To reduce possible variability in team activities and project deliverables, the students followed the basic life-cycle that we tailored from OpenUP and used its artifacts’ templates [34]; and also used a popular UML modeling tool [35].

*Conclusion validity* of a study deals with whether correct conclusions are reached through rigorous and repeatable treatments [36]. To reduce the bias in reaching conclusions for each research question, we relied on statistical analysis. Thus, interpretation of the findings and implications of our research depends on statistical significance and are strictly traceable to data. In addition, by careful definition of evaluation process, its outputs and their grading rubrics, we enabled valid and repeatable investigation of the RQs.

*External validity* is concerned with to what extent the results of this study can be generalized [36]. The study was done in a single university course with only 50 undergraduate students formed into 16 groups. It provides a limited voice of evidence from a small data set, and therefore generalizing its findings is not possible. Though our study added to the body of evidence on this area, replications of it in other contexts would be needed to increase generalizability of our findings.

## VI. CONCLUSIONS AND FUTURE WORK

In this study, we investigated the effect of personality-based team formation (based on social-interaction dimension of MBTI, i.e. introversion/extraversion) on team cohesion and project output (i.e. resulting product quality). We conducted an exploratory case study during a term-long software engineering course containing a project component with 50 students. We grouped the students using the Myers–Briggs Type Indicator (MBTI) personality assessment model; and collected data to explore the team cohesion as measured by team cohesion and morale index, and project grade as an indicator of project output. Our results showed that there was no relation between team formation types and team cohesion, and that some (although weak) relationship existed between the formation schemes and group performance and resulting project output.

Our study provides a limited voice of evidence from a small data set. While our study added to the body of evidence in this area, it also highlighted the very complex nature of human characteristics and its manifestation in team formation and likely results. As researchers, we need to look in further depth into team dynamics and human aspects in software teams. We could consider other dimensions of personality types (i.e., sensing/intuition, thinking/feeling, judging/perceiving) in addition to social interaction dimension (i.e. extraversion/introversion), and their influence on team cohesion and



resulting project output. The findings from such research, including our investigation, might provide insights for practitioners who want to build teams to evaluate and increase the efficiency of their software teams.

Our future work directions include: (1) investigating to see whether Agile teams have a higher team morale than other teams (e.g., working in Waterfall); (2) investigating the effects of other MBTI dimensions on team cohesion and project output; (3) investigating the effects of uniform teams (all extraverts or all introverts) on development activities and if the performance differs in various SDLC phases: Analysis, design, implementation, and testing; and, (4) investigating to see whether more homogeneous teams (in terms of personality) are more suitable for software development compared to less homogeneous teams.

#### ACKNOWLEDGMENT

This study received ethics approval from Hacettepe University in 2015. The authors thank to Tuğba Erdoğan and Pelin Canbay, the TAs of the course, in their assistance in delivery of the laboratory course and marking the project deliverables.

#### REFERENCES

- [1] T. DeMarco and T. Lister, *Peopeware: productive projects and teams*. Dorset House Publishing Co., Inc., 1987, p. 188.
- [2] N. Gorla and Y. W. Lam, "Who should work with whom? building effective software project teams," *Commun. ACM*, vol. 47, no. 6, pp. 79-82, 2004.
- [3] L. T. Hardiman, "Personality Types and Software Engineers," *Computer*, vol. 30, no. 10, pp. 10-10, 1997.
- [4] C. A. Wellington, T. Briggs, and C. D. Girard, "Examining team cohesion as an effect of software engineering methodology," *SIGSOFT Softw. Eng. Notes*, vol. 30, no. 4, pp. 1-5, 2005.
- [5] J. Kam and T. Cowling, "A follow up study of the effect of personality on the performance of software engineering teams," presented at the Proceedings of the ACM/IEEE international symposium on Empirical Software Engineering, 2006.
- [6] S. T. Acuña, M. Gómez, and N. Juristo, "How do personality, team processes and task characteristics relate to job satisfaction and software quality?," *Information and Software Technology*, vol. 51, no. 3, pp. 627-639, 2009.
- [7] O. Mazni, S.-L. Syed-Abdullah, and N. M. Hussin, "Analyzing personality types to predict team performance," in *International Conference on Science and Social Research (CSSR)*, 2010, pp. 624-628.
- [8] F. Q. B. da Silva, S. S. J. O. Cruz, T. B. Gouveia, and L. F. Capretz, "Using Meta-ethnography to Synthesize Research: A Worked Example of the Relations between Personality and Software Team Processes," in *ACM IEEE International Symposium on Empirical Software Engineering and Measurement*, 2013, pp. 153-162.
- [9] D. Bell, T. Hall, J. E. Hannay, D. Pfahl, and S. T. Acuna, "Software engineering group work: personality, patterns and performance," in *Proceedings of the Annual Conference on Computer Personnel Research*, 2010, pp. 43-47, 1796921.
- [10] J. Congalton, "A mixed methods in investigation of ethnic diversity and productivity in software development teams," PhD thesis, Information Systems, Massey University, Wellington, New Zealand, 2014.
- [11] M. Gómez and S. Acuña, "A replicated quasi-experimental study on the influence of personality and team climate in software development," (in English), *Empirical Software Engineering*, vol. 19, no. 2, pp. 343-377, 2014.
- [12] S. T. Acuña, M. N. Gómez, J. E. Hannay, N. Juristo, and D. Pfahl, "Are team personality and climate related to satisfaction and software quality? Aggregating results from a twice replicated experiment," *Information and Software Technology*, vol. 57, pp. 141-156, 2015.
- [13] Constance Lynn Counselling, "Myers-Briggs Type Indicator (MBTI)" <http://www.constancelynn.com/assessments/myers-briggs-type-indicator-mbti/>, Last accessed: Nov. 2017.
- [14] NERIS Analytics Limited, "Free online personality test for Myers-Briggs Type Indicator (MBTI)," <http://www.16personalities.com/free-personality-test>, Last accessed: Nov. 2017.
- [15] S. S. J. O. Cruz, F. Q. B. da Silva, C. V. F. Monteiro, C. F. Santos, and M. T. dos Santos, "Personality in software engineering: Preliminary findings from a systematic literature review," in *Annual Conference on Evaluation & Assessment in Software Engineering*, 2011, pp. 1-10.
- [16] D. P. McAdams, *The Person: An Integrated Introduction to Personality Psychology*. Harcourt, 2001.
- [17] S. McDonald and H. M. Edwards, "Who should test whom? Examining the use and abuse of personality tests in software engineering," *Commun. ACM*, vol. 50, no. 1, pp. 66-71, 2007.
- [18] J. B. Kahnweiler, *The Genius of Opposites: How Introverts and Extroverts Achieve Extraordinary Results Together*. Berrett-Koehler Publishers, 2015.
- [19] Forbes, "Can Introverts And Extroverts Ever Work Well Together? How Opposites Can Collaborate Brilliantly," <http://www.forbes.com/sites/kathycapriano/2015/09/07/can-introverts-and-extroverts-ever-work-well-together-how-opposites-can-collaborate-brilliantly>, 2015, Last accessed: Nov. 2017.
- [20] S. L. Kichuk and W. H. Wiesner, "The big five personality factors and team performance: implications for selecting successful product design teams," *Journal of Engineering and Technology Management*, vol. 14, no. 3-4, pp. 195-221, 1997.
- [21] M. André, M. G. Baldoquín, and S. T. Acuña, "Formal model for assigning human resources to teams in software projects," *Information and Software Technology*, vol. 53, no. 3, pp. 259-275, 2011.
- [22] F. da Silva *et al.*, "Team building criteria in software projects: A mix-method replicated study," *Information and Software Technology*, vol. 55, no. 7, pp. 1316-1340, 2013.
- [23] M. Huxham and R. Land, "Assigning Students in Group Work Projects. Can We Do Better than Random?," *Innovations in Education & Training International*, vol. 37, no. 1, pp. 17-22, 2000.
- [24] M. E. Shaw, R. Robbin, and J. R. Belsler, *Group Dynamics: The Psychology of Small Group Behavior*. McGraw-Hill, 1981.
- [25] C. R. Evans and K. L. Dion, "Group cohesion and performance a meta-analysis," *Small group research*, vol. 22, no. 2, pp. 175-186, 1991.
- [26] B. Mullen and C. Copper, "The relation between group cohesiveness and performance: An integration," *Psychological Bulletin*, vol. 115, no. 2, pp. 210-227, 1994.
- [27] V. R. Basili, "Software modeling and measurement: the Goal/Question/Metric paradigm," Technical Report, University of Maryland at College Park 1992.
- [28] C. Verwijs, "Agile Teams: Don't use happiness metrics, measure Team Morale," <http://blog.agilistic.nl/agile-teams-dont-use-happiness-metrics-measure-team-morale/>, Last accessed: Nov. 2017.
- [29] L. v. Boxmeer, C. Verwijs, R. d. Bruin, J. Duel, and M. C. Euwema, "A direct measure of Morale in the Netherlands Armed Forces Morale Survey: theoretical puzzle, empirical testing and validation," in *Presented at International Military Testing Association Symposium*, 2007.

- [30] I. E. Allen and C. A. Seaman, "Likert Scales and Data Analyses," <http://asq.org/quality-progress/2007/07/statistics/likert-scales-and-data-analyses.html>, Last accessed: Nov. 2017.
- [31] Anonymous author(s), "How to Use the Likert Scale in Statistical Analysis," <http://statisticscafe.blogspot.nl/2011/05/how-to-use-likert-scale-in-statistical.html>, Last accessed: Nov. 2017.
- [32] J. Carver, L. Jaccheri, S. Morasca, and F. Shull, "Issues in using students in empirical studies in software engineering education," in *Proceedings International Software Metrics Symposium*, 2003, pp. 239-249.
- [33] M. Svahnberg, A. Aurum, and C. Wohlin, "Using students as subjects - an empirical evaluation," presented at the Proceedings of the ACM-IEEE international symposium on Empirical software engineering and measurement, 2008.
- [34] Eclipse foundation, "Open Unified Process (OpenUP) development process," <http://epf.eclipse.org/wiki/openup/>, Last accessed: Nov. 2017.
- [35] Visual Paradigm International, "Visual Paradigm tool," <https://www.visual-paradigm.com>, Last accessed: Nov. 2017.
- [36] C. Wohlin, P. Runeson, M. Höst, M. C. Ohlsson, B. Regnell, and A. Wesslén, *Experimentation in Software Engineering: An Introduction*. Kluwer Academic Publishers, 2000.

#### BIOGRAPHIES



**VAHID GAROUSI** received his B.S. and M.S. degrees in Computer Engineering from Sharif University of Technology (Iran) in 2000, and the University of Waterloo (Canada) 2003. He earned his PhD in Software Engineering in Carleton University (Canada) in 2006. He is currently an Associate Professor of Software Engineering in Wageningen University, the Netherlands. Previously, he was an Associate Professor of Software Engineering in Hacettepe University in Ankara, Turkey (2015-2017) and in the University of Calgary, Canada (2006-2014). Vahid was an IEEE Computer Society Distinguished Visitor from 2012 to 2015. His research interests in software engineering include: software testing and quality assurance, model-driven development, and software maintenance.



**AYÇA TARHAN** received the B.S. and M.S. degrees in computer engineering from Ege University in 1995 and from Dokuz Eylül University in 1999, and the Ph.D. degree in information systems from Informatics Institute of Middle East Technical University (METU) in 2006. She was a visiting researcher in 2013-2015 in the Department of Industrial Engineering and Innovation Sciences of Eindhoven University of Technology. Her research interests include internal and external software quality, software metrics, software development methodologies, process maturity, and business process management. Since 2007, she has been working as a Lecturer and an Assistant Professor in Computer Engineering Department of Hacettepe University in Ankara, Turkey.

# Real Measure of a Transmission Line Data with Load Fore-cast Model for The Future

M. Yilmaz

**Abstract**— In this study, an electric transmission line taken hourly data of feeders, belonging to the 1990-2017 year in Turkey by using actual consumption value, load forecasting analysis was done for the future. Short-medium-long term forecast range that results in hourly resolution, presented a mathematical approach to versatile applications. A statistical prediction tool that is called Exponentially Weighted Moving Average (EWMA) is used to predict the next year's demand for transmission in Turkey. In addition to this method, the estimated value of load factors near future, within a few years also has been shown to successfully predict the hour as possible. To load demand will increase in the future, it was presented solutions to be taking precautions.

**Index Terms**— Power grids, microgrids, power transmission, power system management, smart grids.

## I. INTRODUCTION

ELECTRICITY DEMAND is constantly increasing. The way to meet this increasing demand for energy in the most appropriate way possible by making the right forward planning. In order to meet energy demand and realize production at low cost, load estimation analysis is a very important issue. Some studies Show us, the power plants do not use the power of 5-10%, while 20% say that the value is used only 1-2 hours in a day.

For these reasons, it is important to predict in advance. Load forecast analysis is the first step in planning electrical energy. For a good system planning, the energy demand and peak load values must be estimated with the least number of errors. The installation of joints and / or new power plants to be made to the power plants is determined to meet the foreseen energy demand, taking into account the peak load values. According to the load estimation results, the capacity additions to the transmission - distribution systems together with the production and the investment costs related to them are determined.

The inability to store the electric energy increases the importance of the accuracy of the demand estimate. The accuracy of the load demand forecast; reliability and efficiency of electrical power systems, optimization between power plant units, hydrothermal coordination and fuel harvesting affect the operating characteristics of the energy system.

M. YILMAZ is with Department of Electrical and Electronics Engineering, Batman University, Turkey (e-mail: [musa.yilmaz@batman.edu.tr](mailto:musa.yilmaz@batman.edu.tr)).

Manuscript received August 25, 2017; accepted Nov 16, 2017.  
DOI: [10.17694/bajece.419646](https://doi.org/10.17694/bajece.419646)

Faults in the load forecast can cause significant problems in future power system planning. If the estimated value is below the future value, the system is overloaded, the energy quality is decreased and if the forecast is high, the cost is increased and the system is operating at low capacity. In order for the system not to work this way, as the available data increases, the estimates and corresponding plans must be renewed.

Studies on predicting the consumption of electricity is usually done by term load estimates (covering periods from one hour to one month), medium term load estimates (covering periods from one month to one year), long term load estimates (covering periods longer than one year).

The short-term load forecasting analysis determines the load sharing between the power plants and the commissioning status of the production units. Generally, the peak load values in the daily load curve are tried to be predicted in real time. Maintenance programs are prepared, river flow conditions for hydraulic power plants and water quantity to be kept in water reservoir are determined, the amount of fuel is determined in thermal power plants, data related to steam flow are determined and load of the plant units is provided according to estimated load values.

Mid-term load estimation is very important because it covers the planning of physical equipment. At this stage, the transmission system is expanded and transmission, distribution systems and units that can be taken over soon are determined. It is also used in planning distribution systems, collective planning studies and economic reviews to determine sales tariffs, maintenance periods and fuel sources.

In the long-term load forecasting, planning strategies are determined first. In addition to this, issues such as the need for fuel and the determination of fuel resources and the provision of capital are also realized in this period. The most needed long-term load estimate in practice. Because, at this stage, very important decisions are taken and high capital is used and production plans are made.

Load forecasting is important to industry and society in where and when needed, the electrical energy sufficient to meet the need for reliable and not necessary to determine the amount of more or less. In the realization of the electricity energy plans, each plant must be provided with the primary energy source, selection of site, feasibility studies, financing, execution without interruption, operation of the operation teams.

Since there are a lot of uncertainties in long-term load estimates, it is not possible to make a precise and precise estimate. When performing load estimation analysis; accurate determination of the variables affecting the change of the load, generalization with the mathematical load model and methods of obtaining the model parameters, the conditions in the input variables should be taken into consideration.

There are many approaches to load estimation. What is important is to determine the most accurate and most accurate values. When applying these approaches, the answer to the question of whether peak demand or load must be estimated separately is the direct calculation of the first option peak load. In this case, the result can go directly; but economic changes are ignored. The second option is to determine the load by estimating the load. This option means that the load factor is also calculated. In this alternative the energy is more uniform as it is determined by the load, and the population-dependent and economic factors are not neglected; but irregularly varying load factors can lead to erroneous predictions. In this case, a new question emerges. When making load estimates, should the past data be done as a whole or separately for each consumer group? In answer to this question; it is appropriate to make separate estimations for each group by dividing the consumers into separate groups. Finally, all these estimates are combined to determine the total load needed. As a result, misdirection of prediction is prevented. Another option is to estimate the total burden as a whole. This option has ease of use and a more comfortable observation of the growth tendency. Another question is; Should the upper boundary be used or the middle should be used? Based on the weather reports from the past when planning, it is based on the estimation of the load components. In this method, it is necessary to make some adjustments because the air changes do not have a regular course. Finally, detailed and precise mathematical calculations need to be made when estimating. The mathematical method is determined by the structure of the load. Before choosing a particular method, all possible methods should be tried and found to be the most appropriate.

In the literature, the main methods used for short-term load prediction analysis are: regression-based methods [2], Box Jenkins model [3] time series approach [4-5], Kalman filter [6], YSA models [7] hybrid approaches [8]. Recently, methods of using statistical methods and other artificial intelligence approaches as hybrids have also been suggested for solving this problem. Bayesian inference [9], self-organizing maps [10], wavelet transforms [11], and particle swarm optimization [12].

Medium and long term load forecasting analyzes are also very important in planning power systems. Time series approaches [13] and Fourier series (FS) [14] approaches were used for medium term load estimation. The long-term load estimate is important for long-term planning and for determining the peak loads during the year. The most important methods used in long-term load estimation are time series analysis [4-5], hierarchical artificial neural networks [15] and support vector machines [16].

## II. ESTIMATING METHODS

The selection of the forecasting technique to be used is important in determining future demand for freight. Depending on the nature of the load changes, one method may outperform the other. Before choosing a particular method, it is necessary to examine the behavior of the load. It can be understood that it is appropriate to select a suitable curve or a stochastic model for the behavior of the load. Since the electric networks show different characteristics, the structure of the existing system should be examined. It is

important to know the advantages and disadvantages of different systems in order to select the most appropriate technique according to the system examined. Basically there are two estimation methods, extrapolation and correlation. Extrapolation is made by assuming that the past data and the forces influencing this data will increase in the same way as in the past. There are many extrapolation methods. Some of these are made up of the interpretation of mathematical growth curves. The others are for years to be used for the growth averages of the past years. Correlation is the loading of loads through other factors (such as weather or economic conditions). The most important advantage of correlation is to evaluate the factors that affect growth according to their importance. For example, air is the digitization of the relationship between conditioning and load. The correlation method also helps in determining the cause if the estimates deviate from the true values. Some of the estimation methods used are:

Fourier series, Particle flock optimization, Hybrid, Fuzzy logic, Fuzzy logic, Kalman filter, Bayesian inference, Self-organizing maps, Time series analysis, Box Jenkins models and their derivatives, Artificial neural networks models and other methods.

## III. MATHEMATICAL MODELING AND APPLICATION

The mathematical model developed for load prediction analysis is an approach that allows short, medium, and long-term hourly load prediction analysis, unlike previous studies [17]. This model TEİAŞ in Turkey (Turkish Electricity Transmission Company) received the diagnosis hour remaining four years have been found using a total of twenty-six years of actual load data consists of annual value. The proposed method consists of three sub-sections that are intertwined. First part; modeling of annual load values, second part; the modeling of monthly load values during the year, the last part; Modeling of hourly load changes using 3D mathematical notation. A statistical prediction tool that is called Exponentially Weighted Moving Average (EWMA) is used to predict the next year's demand for transmission in Turkey.

In order to use hourly load data in a meaningful way, these data should be analyzed first and their dynamics should be understood. Load values have a dynamic structure and show similarities. But besides this, some unexpected situations, power plant failure, holiday periods, weather conditions and some other factors affect the change of load values. Another observation of load values is that there are two oscillations in the monthly and hourly periods, respectively. However, non-random parts of the load values, in other words portions of the oscillations showing similar variations, can be modeled using wave patterns or mathematical models. In addition, when the average load values of each year are examined, it is seen that the load values increase significantly with years.

One-year total energy values for each day were measured to see the daily variations of the requested load values during the year and the times when the demand was highest and lowest. In these measurements, a daily change of different seasons and periods is given in Figure 1. When analyzes are made, interim valuations are made to eliminate sudden falls (due to electrical failures, failures, etc.). This interim evaluation was done by taking the average of the values of

the previous month and the next month of the relevant time. The seasonal changes can be easily observed from Fig 2. As can be seen, electrical demand is at most winter months. Although not as much as the winter months, there are more demanding requests in July and August. In spring months, especially in May and June, the demanded value falls. In addition, it is seen that the demanded load values increased due to the annual increase. The hourly load values for the year 2017 are shown in 3D in Fig 3.

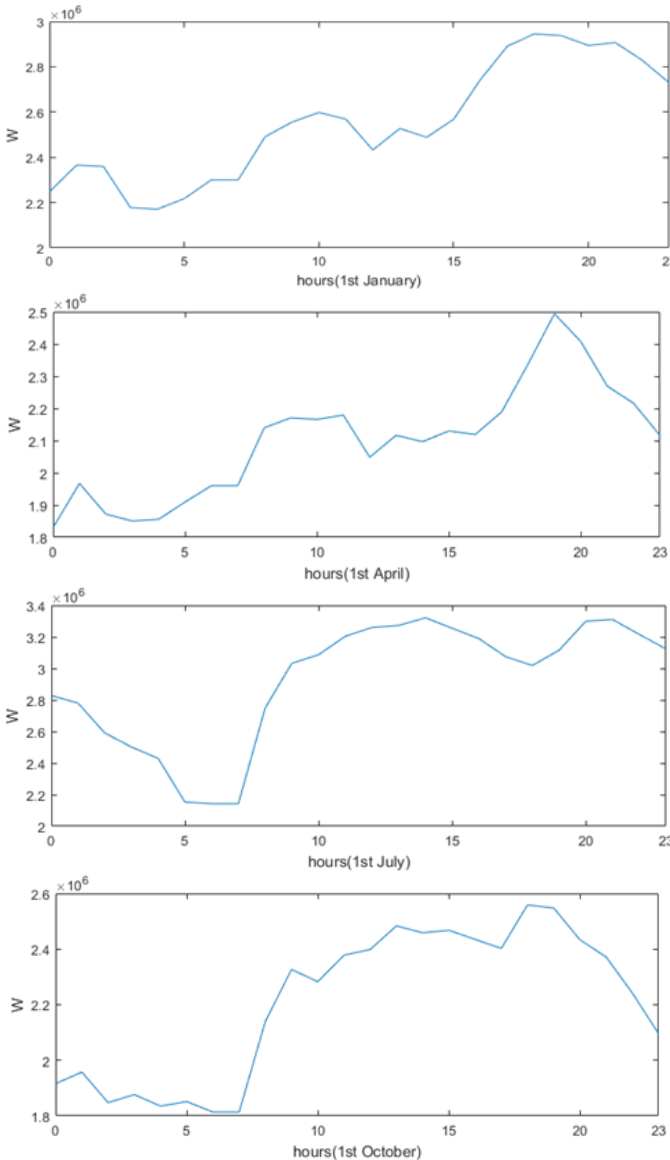


Figure 1. Daily energy demand values for some months of the 2017

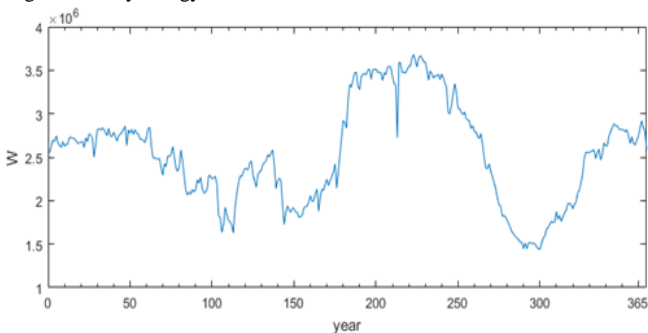


Figure 2. Annual average energy demand per day

The reason for the 3D representation of the load values, the combination of time and day-dependent changes of this

representation; in other words, it has a compact visual property. As can be seen from the figures, the 3D representation has more information and information on the change of load values. Because of the relative nature of the load changes, the mathematical modeling of Figure 3 is extremely complex. On the other hand, if the change in 3D graphics from one day to the next is separated by 3D matrix display, it will be possible to display this less complex model.

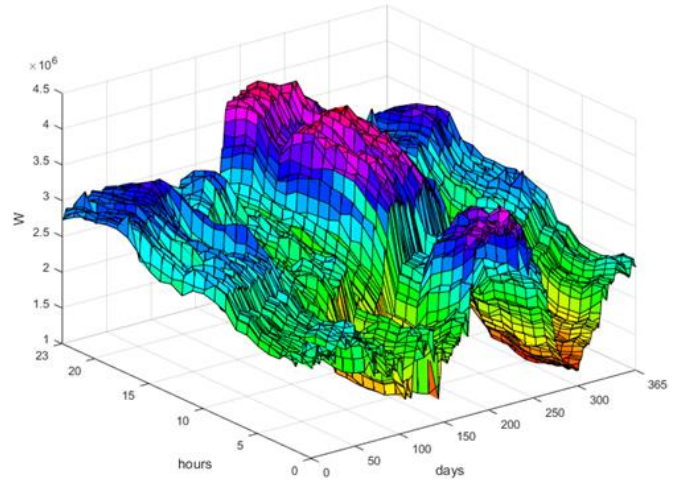


Figure 3. 3D representation of hourly demand data in 2017

In the method proposed in this study, the load values are modeled as three internal parts. The first part is the modeling of the annual mean load values. The second part is modeling the monthly residual load values within a year. The third part is the modeling of hourly variations within a month. This model is also modeled using matlab. This interstitial structure is shown in Fig.3.

#### IV. MODELING OF ANNUAL LOAD VALUES

Turkey is a country constantly increasing demand for electricity. According to TEİAŞ' data, electricity demand is increasing continuously between 1996-2016. The changes in annual load values are given in Fig. 4. The total chart for the years 1996-2016 is given in Figure 5.

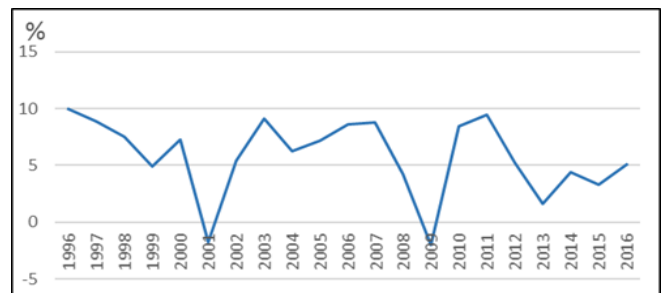


Figure 4. Turkey' electricity demand changes between the years 1996-2016.

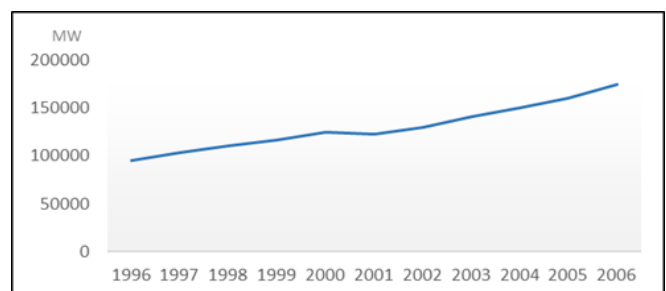


Figure 5. Turkey's electricity demand between the years 1996-2016

To be able to model, this graph can be examined by dividing into two parts. In this way, both the continuously increasing state (Figure 5) and the year-over-year changes (oscillations) can be modeled (functional difference between Figure 1 and Figure 2).

As know EWMA is a time series based statistical tool. To express EWMA, firstly we should define a time series and moving average [2].

A time series can be considered as  $x_t, t=1,2,3, \dots, i$  then the average of it can be calculated. If  $i$  is selected large, then an integer  $n$  which is selected smaller than  $i$ ,  $x_t$  can calculated a set of averages, or simple moving averages (of order  $n$ ) [2]:

$$\bar{x}_{t,1} = \frac{1}{n} \sum_{t=1}^n x_t \quad (1)$$

$$\bar{x}_{t,2} = \frac{1}{n} \sum_{t=2}^{n+1} x_t \quad (2)$$

$$\bar{x}_{t,i-n+1} = \frac{1}{n} \sum_{t=i-n+1}^t x_t \quad (3)$$

Where  $2 \leq n \leq i$  the each calculation of the average of the values over an interval of  $n$  data becomes as follows [2],

$$\bar{x}_t = \frac{1}{n} \sum_{t=t-n+1}^t x_t \quad (4)$$

This reveals that the average estimation at time  $t$  is the simple average of the  $n$  values at time  $t$  and the leading up to  $n-1$  time steps. If weights are applied that decrease the number of  $n$  that are next in time, the moving average will be called as exponentially smoothed. Therefore Moving averages are usually provided forecasting information about at a series time  $t+1$ ,  $S_{t+1}$ , is considered as the moving average for the period of including time  $t$ , e.g. today's forecast is based on an average of earlier values. Using (4) all  $n$ 's are equally weighted. These equal to weights assumed as  $\mu_t$ , every  $n$  weights would equal  $1/n$ , so the sum of the weights would be 1, where  $\mu_t = 1/n$ , then the (4) turns into [2]

$$\bar{x}_t = \sum_{t=t-n+1}^t \mu_t x_t \quad (5)$$

Using exponentially weighted moving averages the contribution to the mean value from  $n$ 's that are more removed in time is planned decreased, so emphasizing more local events. Basically a smoothing parameter is  $0 < \mu < 1$ . Where  $0 < \mu < 0.5$  designates more weight than to the prior ( $x_t$ ). If  $0.5 < \mu < 1$  less weight is assigned to  $x_{t-1}$  and more to  $x_t$ . In exponential smoothing it is needed to use a set of weights that sum to 1 to reduce in size geometrically [2]. The weights are used would be [2]

$$\mu(1 - \mu)^k, \quad (6)$$

where  $k=1,2,3,\dots$ . After some mathematical operations and reduction, the moving average that weighted with (5), (6) becomes [2]

$$\bar{x}_t = \sum_{k=1}^n \mu(1 - \mu)^{k-1} x_{t-k+1}. \quad (7)$$

Then (5) can be written as a repeated smoothed relation [2]

$$S_t = \mu x_t + (1 - \mu)x_{t-1} \quad (8)$$

In (8),  $x_{t-1}$  is indicated annual percentage growth rate of 1996 and  $x_t$  is indicated annual percentage growth rate of

2016 [2].

Modeling the hourly load values: In order to create a smooth hourly model of a year, a 'monthly model' was created using data from 2006. In order to find the monthly model, averages of the values of 12 months were taken. The hourly chart within a month is similar to the other days of the year. The arithmetic average for these 12 months represents the symbolic (nominal) monthly average. Before the average is taken, the total energy of the hourly curves of each month is normalized to normalize the active energy demand and to avoid the effects of increases and decreases within a year. As a result, the total energy of the created model is equal to the individual. The normalized hourly load changes are given in Figure 6.

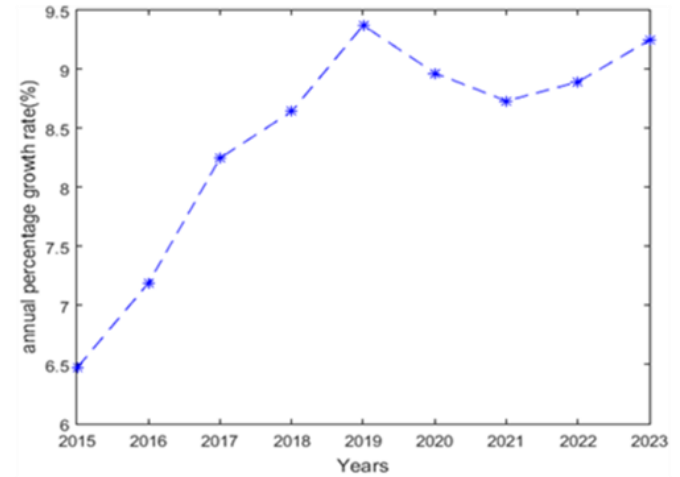


Figure 6. EWMA is used to predict next years' loads.

## V. CONCLUSION

The success of the load estimation analysis depends on the accuracy and accuracy of the statistical data of the current system. In this study, electric power generation, has been tested using the proposed method and the data presented in Turkey for energy transmission and distribution companies and financial units, which is a very important issue in terms of load forecasting analysis. Unlike other hourly load estimation methods, this method can estimate for several years. It has been observed that the estimation results obtained by the mathematical model consisting of subdivided subsegments give successful results on an hourly basis.

To create awareness about management of electric energy demand for next years the electrical load that will affect the electrical grid is shown firstly. Turkey is one of the developing countries. Therefore, the energy and technology demand of the country is increasing continuously. It will be needed a power plant investment due to electric demands in Turkey that depends on foreign energy. To overcome these problems existing power plants should be optimized which can be occurred by smart grid. The aim of the smart grids is to generate efficient and reliable energy and control the generation of power, consumption, storage, distribution and transmission in a flexible way by using of information technologies. If Turkey does not employ the smart grid in electrical power system, it cannot benefit from its renewable energy resources, and its external dependence on energy will increase.

## REFERENCES

- [1] Trudnowski, Dan J., Warren L. McReynolds, and Jeffery M. Johnson. "Real-time very short-term load prediction for power-system automatic generation control." *IEEE Transactions on Control Systems Technology* 9.2 (2001): 254-260.
- [2] Ahmadi, A., Gandoman, F. H., Khaki, B., Sharaf, A. M., & Pou, J. (2016). Comprehensive review of gate-controlled series capacitor and applications in electrical systems. *IET Generation, Transmission & Distribution*, 11(5), 1085-1093.
- [3] Asker, M. E., Ozer, A. B., & Kurum, H. (2016). Reduction of EMI with chaotic space vector modulation in direct torque control. *Elektronika ir Elektrotechnika*, 22(1), 8-13.
- [4] Bayindir, R., Irmak, E., Colak, I., & Bektas, A. (2011). Development of a real time energy monitoring platform. *International Journal of Electrical Power & Energy Systems*, 33(1), 137-146.
- [5] Colak, I., Bayindir, R., Fulli, G., Tekin, I., Demirtas, K., & Covrig, C. F. (2014). Smart grid opportunities and applications in Turkey. *Renewable and Sustainable Energy Reviews*, 33, 344-352.
- [6] Colak, I., Kabalci, E., & Bayindir, R. (2011). Review of multilevel voltage source inverter topologies and control schemes. *Energy Conversion and Management*, 52(2), 1114-1128.
- [7] Colak, I., Kabalci, E., & Bayindir, R. (2011). Review of multilevel voltage source inverter topologies and control schemes. *Energy Conversion and Management*, 52(2), 1114-1128.
- [8] Efe, S. B. (2016). Power Flow Analysis of A Distribution System Under Fault Conditions. *International Journal of Energy and Smart Grid*, 1(1), 22-27.
- [9] Fotouhi Ghazvini, M. A., Soares, J., Morais, H., Castro, R., & Vale, Z. (2017). Dynamic Pricing for Demand Response Considering Market Price Uncertainty. *Energies*, 10(9), 1245.
- [10] Lezama, F., Castañón, G., & Sarmiento, A. M. (2013). Routing and wavelength assignment in all optical networks using differential evolution optimization. *Photonic Network Communications*, 26(2-3), 103-119.
- [11] Nazarloo, Amin, et al. "Improving Voltage Profile and Optimal Scheduling of Vehicle to Grid Energy based on a New Method." *Advances in Electrical and Computer Engineering* 18.1 (2018): 81-88.
- [12] Kaur, Sandeep, and Ganesh Balu Kumbhar. "Incentive driven distributed generation planning with renewable energy resources." *Advances in Electrical and Computer Engineering* 14.4 (2014): 21-28.
- [13] Chanhom, Peerapon, Surasak Nuilers, and Natchpong Hatti. "A new V2G control strategy for load factor improvement using smoothing technique." *Advances in Electrical and Computer Engineering* 17.3 (2017): 43-51.
- [14] Nazaripouya, H., Chu, C. C., Pota, H. R., & Gadh, R. (2017). Battery Energy Storage System Control for Intermittency Smoothing Using Optimized Two-Stage Filter. *IEEE Transactions on Sustainable Energy*.
- [15] Nazaripouya, H., Wang, Y., Chu, P., Pota, H. R., & Gadh, R. (2015, July). Optimal sizing and placement of battery energy storage in distribution system based on solar size for voltage regulation. In *Power & Energy Society General Meeting, 2015 IEEE* (pp. 1-5). IEEE.
- [16] Emiroglu, Selcuk, Yilmaz Uyaroglu, and Gulcihan Ozdemir. "Distributed Reactive Power Control based Conservation Voltage Reduction in Active Distribution Systems." *ADVANCES IN ELECTRICAL AND COMPUTER ENGINEERING* 17.4 (2017): 99-106.
- [17] Reyes-Archundia, Enrique, et al. "Fault detection and localization in transmission lines with a static synchronous series compensator." *Advances in Electrical and Computer Engineering* 15.3 (2015): 17-22.
- [18] Pinto, T., Sousa, T. M., Praça, I., Vale, Z., & Morais, H. (2016). Support Vector Machines for decision support in electricity markets' strategic bidding. *Neurocomputing*, 172, 438-445.
- [19] Soares, J., Canizes, B., Ghazvini, M. A. F., Vale, Z., & Venayagamoorthy, G. K. (2017). Two-Stage Stochastic Model Using Benders' Decomposition for Large-Scale Energy Resource Management in Smart Grids. *IEEE Transactions on Industry Applications*, 53(6), 5905-5914.
- [20] Sousa, T., Morais, H., Vale, Z., Faria, P., & Soares, J. (2012). Intelligent energy resource management considering vehicle-to-grid: A simulated annealing approach. *IEEE Transactions on Smart Grid*, 3(1), 535-542.
- [21] Wang, Y., Shi, W., Wang, B., Chu, C. C., & Gadh, R. (2017). Optimal operation of stationary and mobile batteries in distribution grids. *Applied Energy*, 190, 1289-1301.
- [22] Wang, Y., Wang, B., Chu, C. C., Pota, H., & Gadh, R. (2016). Energy management for a commercial building microgrid with stationary and mobile battery storage. *Energy and Buildings*, 116, 141-150.
- [23] Yilmaz, M. (2017, March). The Prediction of Electrical Vehicles' Growth Rate and Management of Electrical Energy Demand in Turkey. In *Green Technologies Conference (GreenTech), 2017 Ninth Annual IEEE* (pp. 118-123). IEEE.

## BIOGRAPHIES



**MUSA YILMAZ** was born in 1979. He received his BS. in Electrical Education from Abant Izzet Baysal University (Turkey) in 2001. He earned his MS and PhD in Electrical Education from Marmara University (Turkey) in 2004 and 2013, respectively. He earned a postdoc in University of California Los Angeles (US) in 2016. He has been working as "assistant professor" since September 2014 at the Batman University, Department of Electrical and Electronics Engineering.

# Publication Ethics

The journal publishes original papers in the extensive field of Electrical-electronics and Computer engineering. To that end, it is essential that all who participate in producing the journal 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 behavior 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 coauthors

**The undersigned hereby assign(s) to *Balkan Journal of Electrical & Computer Engineering* (BAJECE) copyright ownership in the above Paper, effective if and when the Paper is accepted for publication by BAJECE and to the extent transferable under applicable national law. This assignment gives BAJECE the right to register copyright to the Paper in its name as claimant and to publish the Paper in any print or electronic medium.**

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

1. All proprietary rights other than copyright, including patent rights.
2. The right to make and distribute copies of the Paper for internal purposes.
3. The right to use the material for lecture or classroom purposes.
4. 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 BAJECE.
5. 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 noncommercial 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 BAJECE's official online version of the abstract is provided using the DOI (Document Object Identifier) link.





ISSN: 2147- 284X  
Year: April 2018  
Volume: 6  
Issue: 2

## CONTENTS

- A. Onan**; Sentiment Analysis on Twitter Based on Ensemble of Psychological and Linguistic Feature Sets..... **69-77**
- Y. Kaya**; Classification of PVC Beat in ECG Using Basic Temporal Features, .....**78-82**
- B. Karakaya, T. Kaya, A. Gulten**; FPGA-based ANN Design for Detecting Epileptic Seizure in EEG Signal, **83-87**
- M. Kara, M. Furat**, Client-Server Based Authentication Against MITM Attack via Fast Communication for IIoT Devices, .....**88-93**
- H. Karayığit, Ç. Aci, A. Akdağlı**; A Review of Turkish Sentiment Analysis and Opinion Mining, .....**94-98**
- S. G. Eraldemir, M. T. Arslan, E. Yildirim**; Investigation Of Feature Selection Algorithms On A Cognitive Task Classification: A Comparison Study, .....**99-104**
- T. Tulgar, A. Haydar, İ. Erşan**; A Distributed K Nearest Neighbor Classifier for Big Data, .....**105-111**
- F. G.Furat, T. Ibrıkcı**; Classification of Down Syndrome of Mice Protein Dataset on MongoDB Database, .**112-117**
- F. Karaomerlioglu**; Analysis of Photonic Crystal Tuned by Nematic Liquid Crystals, .....**118-121**
- C. Bakir, M. Yuzkat**; Speech Emotion Classification and Recognition with different methods for Turkish Language, .....**122-128**
- J. Dikun, L. Urmoniene, D. Stanelyte**; Spectral Ratio Method for Fault Detection in Rotating Machines, ....**129-131**
- V. Garousi, A. Tarhan**; Investigating the Impact of Team Formation by Introversion/Extraversion in Software Projects, .....**132-140**
- M. Yılmaz**; Real Measure of a Transmission Line Data with Load Fore-cast Model for The Future, .....**141-145**

## BALKAN JOURNAL OF ELECTRICAL & COMPUTER ENGINEERING

(An International Peer Reviewed, Indexed and Open Access Journal)

### Contact

Istanbul Technical University  
Department of Electrical Engineering,  
Ayazaga Campus, Maslak, Istanbul-Turkey

Web: <https://www.bajece.com>  
<http://dergipark.gov.tr/bajece>  
e-mail: [editor@bajece.com](mailto:editor@bajece.com)

