



# JISTA

*Journal of Intelligent Systems: Theory  
and Applications*

JANUARY 2019

ISSN: 2651-3927



**VOL 2 NO 1**

ARTIFICIAL INTELLIGENT > MACHINE LEARNING > MULTI-AGENTS

[WWW.JISTA.INFO](http://WWW.JISTA.INFO)



# Journal of Intelligent Systems: Theory and Applications

## Volume: 2 Issue: 1

### Editorial Boards

#### Honorary Editors

Zekai Şen, zsen@itu.edu.tr, Istanbul Technical University, Turkey

Burhan Turksen, bturksen@etu.edu.tr, TOBB ETU, Turkey

#### Editor-In-Chief

Harun Taşkın, taskin@sakarya.edu.tr, Sakarya University, Turkey

#### Associate of Editor-In-Chief

Özer Uygun, ouygun@sakarya.edu.tr, Sakarya University, Turkey

#### Editorial Board

Ali Allahverdi, ali.allahverdi@ku.edu.kw, Kuwait University, Kuwait

Andrew Kusiak, andrew-kusiak@uiowa.edu, The University Of Iowa, United States of America

Ayhan Demiriz, ademiriz@sakarya.edu.tr, Gebze Technical University, Turkey

Barış Yüce, b.yuce@exeter.ac.uk, Exeter University, United Kingdom

Cemalettin Kubat, kubat@sakarya.edu.tr, Sakarya University, Turkey

Cemil Öz, coz@sakarya.edu.tr, Sakarya University, Turkey

Dervis Karaboga, karaboga@erciyes.edu.tr, Erciyes University, Turkey

Ebubekir Koç, ekoc@fsm.edu.tr, Fatih Sultan Mehmet University, Turkey

Eldaw E. Eldukhri, eeldukhri@ksu.edu.sa, King Saud University, College Of Engineering Al-Muzahmia Branch, Saudi Arabia, United Kingdom

Ercan Öztemel, eoztemel@marmara.edu.tr, Marmara University, Turkey

Güneş Gençyılmaz, gunesgencyilmaz@aydin.edu.tr, Turkey

Hamid Arabnia, hra@cs.uga.edu, University Of Georgia, United States of America

Lyes Benyoucef, lyes.benyoucef@Isis.org, Aix-Marseille University, Marseille, France

Maged Dessouky, maged@rcf.usc.edu, University Of Southern California, Los Angeles, United States of America

Mehmet Emin Aydın, mehmet.aydin@beds.ac.uk, United Kingdom

Mehmet Recep Bozkurt, mbozkurt@sakarya.edu.tr, Sakarya University, Turkey

Mehmet Savsar, mehmet.savsar@ku.edu.kw, Kuwait University, Kuwait

Mohamed Dessouky, dessouky@usc.edu, University Of Southern California, Los Angeles, United States of America

M.H. Fazel Zarandi, zarandi@aut.ac.ir, Amerikabir University Of Technology, Iran

Türkey Dereli, dereli@gantep.edu.tr, Gaziantep University, Turkey

Witold Pedrycz, pedrycz@ee.ualberta.ca, University Of Alberta, Canada

Yılmaz Uyaroğlu, uyaroglu@sakarya.edu.tr, Sakarya University, Turkey



# Journal of Intelligent Systems: Theory and Applications

## Volume: 2 Issue: 1

### Contents

---

#### Review Articles

---

**Artificial intelligence in corneal topography**

1-6

*Nazar Saleh, Nebras Hussein*

#### Research Articles

---

**Eta Correlation Coefficient Based Feature Selection Algorithm for Machine Learning: E-Score Feature Selection Algorithm**

7-12

*Muhammed Kürşad Uçar*

**A Survey on Anomaly Detection and Diagnosis Problem in the Space System Operation**

13-17

*Seçil Taburoğlu*



# Artificial intelligence in corneal topography

Nazar Saleh\* , Nebras Hussein<sup>1</sup>

<sup>1</sup> Khwarizmi College of Engineering, Baghdad, Iraq.

## Abstract

The purpose of this paper is to explore the effectiveness and efficiency of various artificial intelligence (AI) techniques in extracting features from corneal topographies. A considerable number of dated and contemporary related research papers have been reviewed. The author has only checked the studies that considered developing at least one AI-based algorithm for data classification of topographic patterns. The results of this review emphasize the effectiveness and efficiency of machine learning algorithms in the clinical diagnosis of various eye refractive problems.

**Keywords:** Artificial intelligence, classification, feature extraction, topographic images.

## Korneal topografide yapay zeka

### Öz

Bu yazının amacı kornea topografyasından özelliklerin çıkarılmasında çeşitli yapay zeka tekniklerinin etkinliğini ve etkinliğini araştırmaktır. Önemli sayıda çağdaş ve güncel araştırma makaleleri gözden geçirilmiştir. Yazar, sadece topografik modellerin veri sınıflandırması için en az bir AI tabanlı algoritmanın geliştirilmesini düşündüğü çalışmalarını kontrol etmiştir. Bu derlemenin sonuçları, çeşitli göz kırıcı problemlerinin klinik tanısında, makine öğrenimi algoritmalarının etkinliğini ve verimliliğini vurgulamaktadır.

**Anahtar Kelimeler:** Yapay zeka, sınıflandırma, özellik çıkarma, topografik görüntüler.

## 1. Introduction

Topographical images of the human cornea anterior and posterior surfaces are becoming more important diagnostic tools than ever in the classification of the healthy and unhealthy corneas. These maps are created by a system called Pentacam. The final maps of this system carry a lot of details that help and guide the medical decision. Finding hidden details and reporting the expected clinical situation is in most of the time, related strongly to the experience and level of knowledge of the final decision maker (ophthalmologist).

At the same time, the artificial intelligence started to play the main role in the medical imaging analysis and smart diagnosis. These tools accompanied by the huge development level encountered in the computer (software and hardware) science introduce a new analysis area and unprecedented tools. The capability of recognizing particular features of these maps and defining the clinical interrelationship among them plays new roles in the artificial intelligence of the medical imaging field. These selected features are learned and specified earlier through a number of cases

that have been pre-diagnosed and reported by ophthalmologists.

### 1.2. Cornea

The cornea is the outlying part of the human eye and holds a significant focusing power. It is of a domed shape and transparent in nature. Despite this transparency, the cornea is characterized by a perfect organization of tissues and with neither substances nor blood vessels. It gets its nourishment and infections protection aspect from two other parts, the fluid layer that resides behind it and called the aqueous humans and the tears (Camarillo et al., 2002).

### 1.3. Corneal diseases

It is significant to interpret the biomechanical response of cornea for post-surgery cases and some diseases. The cornea is the topmost lens of the eye and accountable for 65 - 75% of the sight strength and therefore any disorder in the cornea curvature results in a sight disorder or refractive problems including, astigmatism, hyperopia, and myopia. High corneal

\* Corresponding Author. Aksaray University  
E-mail: nazar.s2009@yahoo.com

Received : Aug 31, 2018  
Revision : Sep 17, 2018  
Accepted : Nov 19, 2018



curvature leads to nearsightedness or myopia. Low corneal curvature causes farsightedness or hyperopia [2]. These two refractive diseases may also occur due to a malfunction of the lens. Additionally, uneven or abnormal corneal curvature causes blurred vision or astigmatism [3].

#### 1.4. Pentacam

Pentacam is a powerful scanner for the eye segments and based on the Scheimpflug camera measurement. It is a robust invention that helps ophthalmologists in investigating the corneal anterior surface. Although Scheimpflug technique was known for scanning anterior eye segments since the 70s, it was not before the 80s and 90s that it was released commercially. In 2005, it started to be used widely clinically [4].

#### 1.5. Corneal topographic maps

The four major topographic patterns are the pachymetry pattern, the anterior curvature sagittal pattern, in addition to both the anterior and posterior elevation pattern (Mazen M Sinjab, 2015) ( as indicated in Fig.1). These patterns reflect lots of details about the anterior eye surface which should be tested by the specialist.

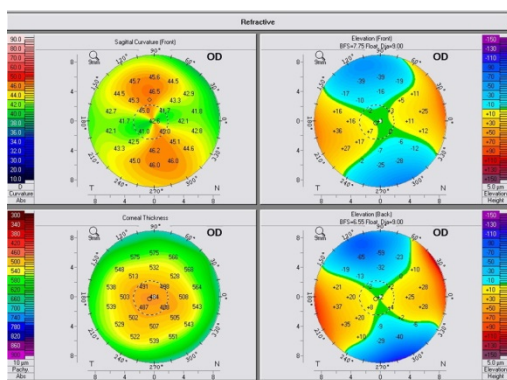


Figure 1. The four refractive topographical map

## 2. Artificial intelligence techniques

### 2.1. Artificial Neural Networks (ANNs) Based Classifiers

Some researchers, as early as 1995, attempted to evaluate the capabilities and usefulness of an automated system based on ANNs in interpreting the corneal topographic maps. They asserted that the difficulty of corneal topographic pattern interpretation, specifically when the topographic maps are similar to other pathologies, encouraged them to investigate the capabilities of an automated approach for diagnosing

shape abnormalities in the cornea. The study reported that the ANNs system may need further enhancements to be an objective computerized classification tool of video-keratography for helping the doctors in determining the abnormalities in corneal topographies (Maeda, Klyce, & Smolek, 1995).

Smolek and Klyce (1997) conducted a study to compare four video-keratographic based keratoconus detection methods to a new detection method based on ANNs. The researchers designed a neural network approach that was able to classify the keratoconus screening to automatically discover whether the keratoconus suspected (KCS) or keratoconus (KC) exists. The research findings assured that the ANNs approach succeeded in differentiating between the topographies of the KC from KCS, and even from other cases that bear a resemblance to KC. Moreover, the outcomes proved that the performance of the new ANNs detection method outweighed the other detection methods that the researcher examined in the study. The ANNs method recorded a higher performance in terms of specificity and accuracy. However, the researchers demonstrated that the sensitivity of all the detection methods including the ANNs was equal (Smolek & Klyce, 1997).

In 2008, for classifying particular types of corneal shapes, a study conducted to develop and compare two different classification techniques, the discriminant analysis, and ANNs, provided that the Zernike coefficients are given. It is noted that just a few studies have applied Zernike coefficients (ZC) as input parameters (Carvalho, 2005; Schwiegerling & Greivenkamp, 1996).

Despite the fact that the researchers used a comparatively small database, the study's outcomes confirm that for automated diagnosis of videokeratography patterns by the ANNs or discriminant analysis, a possible reliable parameter may be the Zernike coefficients to be used as input data as descriptors of the corneal shape (de Carvalho & Barbosa, 2008).

In 2018, some researchers, at Shiley Eye Institute at UC San Diego Health and University of California San Diego School of Medicine, have developed a new ANNs-based system to examine patients with common retinal diseases that cause blinding but are treatable. The diagnosis of both the macular degeneration related to age and diabetic macular edema was comparable to the experts' classification. They additionally indicated that the automated system introduced more interpretable and transparent diagnosis because it was able to determine the regions that the ANNs identified (Kermany et al., 2018; LaFee, 2018).

### 2.2. Support Vector Machines (SMVs) Based Classifiers

In 2005, a researcher developed a framework built on SMVs for ophthalmology pattern interpretation.



And to assist optometry doctors in the diagnosis of ocular refraction problems including short-sightedness, hypermetropia, and astigmatism. The outcomes of the project recommended a method for improving the understanding of eye patterns that were obtained from a system called HS (Hartmann-Shack), provided that other ocular errors measurements are measured and detected. The system was designed to interpret and analyze eye patterns as a whole without the need of having a reference pattern to evaluate the difference between the obtained data between the target and reference pattern (Netto, 2005).

In 2008, some researchers attempted to perform a classification of corneal topographies using the SVM technology. The study conducted to prove the success of a learning algorithm uses the SVM technology in distinguishing between corneal diseases such as keratoconus, suspect keratoconus, hyperopic and myopic laser vision correction, orthokeratology, and pellucid marginal degeneration from normal corneal conditions. The researchers pursued their study through testing topographies obtained by the Corneal Topographer ATLAS® Model 90000. The algorithm was trained and evaluated using previous data on 239 abnormal corneal cases and 85 normal cases. For the training purpose, four-fifths of the cases were selected randomly, while the remaining fifth was used for the evaluation purpose. The accuracy, sensitivity, and specificity of the algorithm were equal to or greater than 90%. The SVM-based model was able to identify the cases of abnormal corneas from normal conditions using the randomly picked training data (Bagherinia et al., 2008).

In 2012, a study developed an SVM-based algorithm for subclinical keratoconus and keratoconus diagnosis through tomography and topography data. It introduced a new data classifier algorithm for the detection of keratoconus using measurements of corneas obtained by Placido corneal topography together with a Scheimpflug camera. The SVM demonstrated a high specificity, sensitivity, and accuracy, in distinguishing four eye conditions, eyes with previous surgery history, eyes with subclinical keratoconus or keratoconus, and normal eyes. The algorithm accuracy was excellent for the cases regardless of if it was provided with the data obtained from the corneal thickness and posterior corneal surface or not. The rate of valid predictions was > 95% when provided with additional data, and 93% when no additional data was provided. It is obvious that when additional data of the posterior corneal surface was provided, the precision was higher. The pachymetry and the posterior and anterior corneal surfaces additional data increased the sensitivity of the classifier (Arbelaez, Versaci, Vestri, Barboni, & Savini, 2012).

A study developed a new diagnosis system for vision-related problems. It presented a system that is capable of performing two tasks, first to diagnose the Short-Sightedness (Myopia ) or Long-Sightedness

(Hypermetropia ) and second to identify the measurements of SPH (eyesight distance) and CYL (deviation). The researchers believe that it is easy for a physician to diagnose the problem if the SPH or CYL lies within the normal measurements, however, when any of them lies out of normal figures then it is not such an easy task to detect and it takes more time for the doctor to determine the condition. Therefore, they proposed a model based on SVM and ANNs to detect eyesight problems and to help doctors decide appropriately and quickly. The ANNs is applied due to its known strength in checking and classifying sightedness of the eye. The developers trained the ANNs by using Backpropagation algorithm due to its high capability in maps classification especially nonlinear mapping. The results showed that with ANNs, the nonlinear SVMs specificity increased from 92% to 97%. These results prove the capabilities of ANNs as well as SVM as pattern recognition techniques (Kotsia & Pitas, 2007; Pontil & Verri, 1998) in detecting the state of the eyesight (Noaman, Muharram, & Alqubati, 2014).

A study conducted to meet two purposes; first is to test the reliability of an SVM-based algorithm for detecting corneal patterns objectively and automatically using a set of twenty-two parameters provided by Pentacam measurements, and second is to put the algorithm model outcomes in comparison with other popular keratoconus classifiers results. For the classification, 22 parameters were applied to an SVM algorithm developed as a piece of computer software based on machine-learning and called Weka. The researchers calculated and compared the accuracy through cross-validation tests for, KC versus normal, forme fruste versus normal and across all the 5 cohorts, with other outputs of some other popular classification models. In case of the KC vs. normal, the algorithm outstanding accuracy was 98.9%, while 99.1% recorded for sensitivity, and 98.5% obtained for the KC detection specificity. For the forme fruste versus normal eyes case, the results were 93.1% for accuracy, 79.1% for sensitivity, and 97.9% for specificity. For the classification of the five cohorts, 88.8% was recorded for accuracy and 89.0% for the average sensitivity, and 95.2% for specificity (Ruiz Hidalgo et al., 2016).

### 2.3. Decision Tree - Based Classifiers

In 2005, a study applied an automated classification system based on machine learning and decision tree induction to distinguish between keratoconus and normal corneal shapes quantitatively and objectively. They also set a comparison between the developed classifier and other popular classification systems. A model of seventh-order Zernike polynomial was applied to the corneal surface. The classifier was based on the C4.5 algorithm for the DT classifier. The outcomes of the automated classifier were compared

with several modified indices, the KISA%, the Cone Location and Magnitude, Keratoconus Prediction Index (KPI), Schwiegerling's Z3, and Rabinowitz–McDonnell– given that the standard thresholds of classification for every method were followed. For every method of classification, the researcher additionally examined the area under the ROC (receiver operator characteristic) curve. The performance of the developed classification method based on DT was greater than or equal to the other classification methods examined. The results showed 92% for accuracy, and 0.97 for the area under the curve of ROC. With 4 of 36 surface features of Zernike polynomial coefficients, the automated DT model decreased the required information for the discrimination between keratoconus and normal eye. The 4 classification attributes the study used were greater sagittal depth, inferior elevation, trefoil and oblique toricity. The research findings confirmed the capability of the decision tree-based classifier in distinguishing among various corneal shapes, it is also assured that it is a dependable quantitative classifier through Zernike polynomials and any device which can produce raw elevation data. They also assured that the framework is applicable to other problems of classification (Twa et al., 2005).

Kabari and Nwachukwu (2012) introduced a framework combining ANNs and DTs for various eyes diseases diagnosis such as myopia, hyperopia, and Astigmatism. It is distinguished from other related studies because it didn't exploit one algorithm only but two technologies are used for data mining ANNs and DTs. The researchers presented a hybrid model called NNDTEDDS (Neural Networks Decision Trees Eye Disease Diagnosing System). They asserted that both technologies have been tools for discovering knowledge and introducing a hybrid model, the researchers developed rules that show how eye diseases diagnosis is done based on the physical conditions of the eye and its related symptoms. These rules show how the knowledge obtained in neural networks after being trained from earlier samples of physical conditions of eye and symptoms. The findings of the study prove a considerable success of 92%, which reflects how combining the two technologies are efficient and effective in diagnosing eye diseases such as myopia, hyperopia, and astigmatism. Moreover, Kabari and Nwachukwu assured that the concluded rules are useful in teaching younger ophthalmologists (Kabari & Nwachukwu, 2012)

#### 2.4. Naive Bayes Based Classifiers

An experimental research introduced an automated, Naive Bayes-based system for detecting eye disease. The automated system uses CBR (Case-Based Reasoning) as an experience-based model for reasoning, and the Naïve Bayes as an eye diseases classifier by adopting the theorem of Bayes. The

findings showed that the accuracy of the automated model of Naïve Bayes was 82%. Accordingly, they concluded that an expert system based on a combination of Naïve Bayes and CBR can be a promising technology for eye diseases diagnosis. Naïve Bayes classifier is capable of introducing effective diagnoses for people but more improvements and enhancements were still required (Kurniawan, Yanti, & Ahmad Nazri, 2014).

#### 2.5. Studies comprised Several Classifiers

The focus of a study conducted in 2012 was to examine the effectiveness of an AI-based system in Keratoconus diagnosis. The researcher believed that the developed system would help device producers to enhance their production so that the medical system can assist experts and support the recognition phase automatically. They trained some classifiers, DT, SVM, ANNs, RBFNN, and Multi-Layer-Perceptron to recognize whether it is a Keratoconus or suspect to Keratoconus eye. The classifiers were trained with part of the dataset and tested by the rest. The output of the study indicated that the proposed algorithm 's accuracy was 91% in discrimination among KC, suspect to KC and normal eye. This accuracy assures the capability of the developed algorithm in automatic detection of KC or suspects to KC (Toutounchian, Shanbehzadeh, Khanlari, & Stage, 2012).

In 2014, a survey conducted to review several automated, computer-based systems for ocular diseases diagnosis. The researchers focused on three kinds of data comprising imaging, clinical, and genetic. According to the researchers, such types of data were the commonly used data types in the computer-aided diagnosis for ocular diseases such as Pathological Myopia, Macular Degeneration that is age-related, Glaucoma, and Diabetic Retinopathy. The researchers asserted that over the past years, ocular diseases diagnosis that is based on computer-aided models has shown substantial advancement. However, the significance of having fully automatic models which are capable of exploiting the clinical knowledge and incorporate mixed data sources still needed for future development (Zhang et al., 2014).

#### 2.6. Google's DeepMind Health Project

AI is extensively applied for tracing, normalizing, collecting, and storing data in healthcare. In this regard, Google's AI research branch announced its new project to investigate the extent to which technology could be efficient in analyzing eye scans through medical records data mining for the purposes of delivering faster and better health services and improving eye treatment. The project is called DeepMind Health. It is a cooperative project with Moorfields Eye Hospital NHS Foundation Trust launched in 2016. Although the project is still in the early stages of AI research in health, Deep Mind Health reassures that the undertaken

research will have practical benefits, and the outcomes of the study will be submitted to academic journals that are peer-reviewed after performing the strict clinical scrutiny (“Researching for tomorrow,” 2018).

### 3. Conclusion

It is beneficial and helpful to explore the plenty of attempts of applying AI to eye diseases diagnosis throughout the history. The exploration of related works over the past decades gives some insights as to what extent the AI-based technology supports the medical care in areas where much effort and time are required by specialists. Numerous studies have been conducted to test the accuracy and specificity of one or more of the artificial intelligence algorithms such as the ANNs, SVMs, DT, and Naive Bayes. Some researchers have tested one algorithm while others have concerned with combining two or more algorithms. All the reviewed research reported high efficiency and effectiveness of the tested algorithms.

### References

- Arbelaez, M. C., Versaci, F., Vestri, G., Barboni, P., & Savini, G. (2012). Use of a support vector machine for keratoconus and subclinical keratoconus detection by topographic and tomographic data. *Ophthalmology*, *119*(11), 2231–2238. <https://doi.org/10.1016/j.ophtha.2012.06.005>
- Bagherinia, H., Chen, X., Flachenecker, C., Angeles, R., Burger, D., Caroline, P., ... Reeder, K. (2008). Support Vector Machine (SVM)-Based Classification of Corneal Topography. *Investigative Ophthalmology & Visual Science*, *49*(13), 1023. Retrieved from <http://dx.doi.org/>
- Camarillo, T., Choi, K., Hamilton, G., Miles, M., Muller, K., Williams, K., ... Schrepel, P. (2002). Athletes as an Ideal Target Population for Orthokeratology Keratoconus: Improving Quality of Life Through Advancements in Detection and Treatment.
- Carvalho, L. A. (2005). Preliminary Results of Neural Networks and Zernike Polynomials for Classification of Videokeratography Maps: *Optometry and Vision Science*, *82*(2), 151–158. <https://doi.org/10.1097/01.OPX.0000153193.41554.A1>
- de Carvalho, L. A., & Barbosa, M. S. (2008). Neural networks and statistical analysis for classification of corneal videokeratography maps based on Zernike coefficients: a quantitative comparison. *Arquivos Brasileiros de Oftalmologia*, *71*(3), 337–341. Retrieved from <http://www.ncbi.nlm.nih.gov/pubmed/18641817>
- Kabari, L., & Nwachukwu, E. (2012). Neural Networks and Decision Trees For Eye Diseases Diagnosis. In P. Vizureanu (Ed.), *Advances in Expert Systems*. InTech.
- Kermany, D. S., Goldbaum, M., Cai, W., Valentim, C. C. S., Liang, H., Baxter, S. L., ... Zhang, K. (2018). Identifying Medical Diagnoses and Treatable Diseases by Image-Based Deep Learning. *Cell*, *172*(5), 1122–1131.e9. <https://doi.org/10.1016/j.cell.2018.02.010>
- Kotsia, I., & Pitas, I. (2007). Facial Expression Recognition in Image Sequences Using Geometric Deformation Features and Support Vector Machines. *IEEE Transactions on Image Processing*, *16*(1), 172–187. <https://doi.org/10.1109/TIP.2006.884954>
- Kurniawan, R., Yanti, N., & Ahmad Nazri, M. Z. (2014). Expert systems for self-diagnosing of eye diseases using Naïve Bayes (pp. 113–116). IEEE. <https://doi.org/10.1109/ICAICTA.2014.7005925>
- LaFee, S. (2018, February). Artificial Intelligence Quickly and Accurately Diagnoses Eye Diseases and Pneumonia.
- Maeda, N., Klyce, S. D., & Smolek, M. K. (1995). Neural network classification of corneal topography. Preliminary demonstration. *Investigative Ophthalmology & Visual Science*, *36*(7), 1327–1335.
- Mazen M Sinjab. (2015). *step by step Reading Pentacam Topography (basics and case study series)*.
- Netto, A. V. (2005). System Based on Computational Intelligence for Ophthalmology Image Understanding. *IEEE Latin America Transactions*, *3*(5), 14–22. <https://doi.org/10.1109/TLA.2005.1642434>
- Noaman, K. M., Muharram, A. A., & Alqubati, I. A. (2014). Diagnosis of Poor Eyesight based on Support Vector Machine and Artificial Neural Networks. *Journal of Emerging Trends in Computing and Information Sciences*, *5*(10).
- Pontil, M., & Verri, A. (1998). Support vector machines for 3D object recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *20*(6), 637–646. <https://doi.org/10.1109/34.683777>
- Researching for tomorrow [WWW Document], 2018. DeepMind.
- Ruiz Hidalgo, I., Rodriguez, P., Rozema, J. J., Ni Dhubghaill, S., Zakaria, N., Tassignon, M.-J., & Koppen, C. (2016). Evaluation of a Machine-Learning Classifier for Keratoconus Detection Based on Scheimpflug Tomography. *Cornea*, *35*(6), 827–832. <https://doi.org/10.1097/ICO.0000000000000834>
- Schwiegerling, J., & Greivenkamp, J. E. (1996). Keratoconus detection based on



videokeratographic height data. *Optometry and Vision Science: Official Publication of the American Academy of Optometry*, 73(12), 721–728.

- Smolek, M. K., & Klyce, S. D. (1997). Current keratoconus detection methods compared with a neural network approach. *Invest Ophthalmol. Vis. Sci.*, 38(0146–0404 (Print)), 2290–2299.
- Toutounchian, F., Shanbehzadeh, J., Khanlari, M., & Stage, A. R. (2012). Detection of Keratoconus and Suspect Keratoconus by Machine Vision. *International Multiconference of Engineers and Computer Scientists*, 1(March 2012), 14–16.
- Twa, M. D., Parthasarathy, S., Roberts, C., Mahmoud, A. M., Raasch, T. W., & Bullimore, M. A. (2005). Automated Decision Tree Classification of Corneal Shape: *Optometry and Vision Science*, 82(12), 1038–1046. <https://doi.org/10.1097/01.opx.0000192350.01045.6f>
- Zhang, Z., Srivastava, R., Liu, H., Chen, X., Duan, L., Kee Wong, D. W., ... Liu, J. (2014). A survey on computer aided diagnosis for ocular diseases. *BMC Medical Informatics and Decision Making*, 14. <https://doi.org/10.1186/1472-6947-14-80>



# Eta Correlation Coefficient Based Feature Selection Algorithm for Machine Learning: E-Score Feature Selection Algorithm

Muhammed Kürşad UÇAR<sup>1\*</sup>

<sup>1</sup>Sakarya University, Faculty of Engineering, Electrical-Electronics Engineering, 54187, Sakarya / Turkey  
mucar@sakarya.edu.tr

## Abstract

Feature selection algorithms are great importance in the field of machine learning. The primary function of feature selection algorithms is to select features in a meaningful way. Features Selection Algorithms methods are still being developed today. The reason for this is that data quantities are growing day by day. As the data increases, more advanced, better performance, feature selection algorithms are needed. In this study, Eta Correlation Coefficient based E-Score Feature selection algorithm was developed. Two versions were prepared for E-Score. We tested the performance of the E-Score method with three classifiers and compared with conventional F-Score Feature Selection Algorithm. According to the results, both versions of the E-Score feature selection algorithm have improved performance and is better than the F-Score. According to these results, it is thought that the E-Score Feature Selection Algorithm can be used in the field of machine learning.

**Keywords:** Eta Correlation Coefficient, E-Score Feature Selection Algorithm, Feature Selection Methods.

## Makine Öğrenmesi için Eta Korelasyon Katsayısı Tabanlı Özellik Seçme Algoritması: E-Score Özellik Seçme Algoritması

### Öz

Makine öğrenmesi alanında özellik seçme algoritmaları büyük öneme sahiptir. Çok büyük verilerin anlamlı bir şekilde azaltılması özellik seçme algoritmalarının temel işlevidir. Bu yöntemler günümüzde hala geliştirilmeye devam etmektedir. Bunun sebebi her geçen gün daha büyük verilerle çalışıyor olmasıdır. Veriler arttıkça daha gelişmiş, performansı daha iyi özellik seçme algoritmalarına ihtiyaç duyulacaktır. Bu çalışmada Eta Korelasyon Katsayısı tabanlı E-Score Özellik seçme algoritması geliştirilmiştir. Geliştirilen yöntem için iki farklı versiyon hazırlanmıştır. E-Score yönteminin performansı üç sınıflandırıcı ile test edilmiştir. Ayrıca literatürde bulunan F-Score Özellik Seçme Algoritması ile de kıyaslanmıştır. Elde edilen sonuçlara göre E-Score özellik seçme algoritmasının her iki versiyonu da performansı arttırmıştır. Ayrıca F-Score ile kıyaslandığında daha iyi başarı oranı elde etmiştir. Bu sonuçlara E-Score Özellik Seçme Algoritmasının makine öğrenmesi alanında kullanılabileceği düşünülmektedir.

**Anahtar Kelimeler:** Eta Korelasyon Katsayısı, E-Score Özellik Seçme Algoritması, Özellik Seçme Yöntemleri.

## 1. Introduction

In machine learning, datasets are the essential elements. Thanks to today's technology, the amount of collected data has reached enormous amounts. Massive

data sometimes have a negative impact on the machine learning process (Guan *et al.*, 2014). Nowadays, one of the most significant problems in machine learning is that significant data lengthens the process and reduces performance. The reason for the decrease in performance is that the irrelevant data is in the cluster.

\* Corresponding Author. Phone: +90 506 849 31 46  
E-mail: mucar@sakarya.edu.tr

Received : Dec 18, 2018  
Revision : Jan 9, 2019  
Accepted : Jan 17, 2019



To solve this problem, Polat has developed algorithms to select the related properties from datasets (Polat and Güneş, 2009; Kavsaoglu, Polat and Bozkurt, 2014). These algorithms are commonly called feature selection algorithms.

Feature selection algorithms aim to increase the performance of classification by selecting important features from datasets according to specific algorithms (Polat and Güneş, 2009; Guan *et al.*, 2014; Cai *et al.*, 2018). Training time, classification accuracy rate, data size, number of features selected affects performance. There are many different types of data in the datasets (Cai *et al.*, 2018). Therefore, a feature selection algorithm cannot be used in each dataset.

Feature selection algorithms can be used wherever machine learning is available. For example, it is used in many areas such as image processing, signal processing, classification problems and data mining (Khotanzad and Hong, 1990; Goltsev and Gritsenko, 2012). As the problems develop, new solutions are developed. Recently, the Ensemble Feature Selection algorithms have been developed (Li, Gao and Chen, 2012; Elghazel and Aussem, 2015). This method combines performance with different feature selection algorithms to improve performance.

The performance of the feature selection algorithms developed in the literature is generally assessed by classification algorithms such as k-Nearest Neighborhood Algorithm (kNN), Support Vector Machines (SVMs), Radial Basis Function (RBF) (Huang, 1999; Cai *et al.*, 2018). A useful feature selection algorithm has a high accuracy rate and fast operation (Cai *et al.*, 2018).

Many feature selection algorithms have been developed in the literature. These can be developed based on statistical or different basic principles (Tsang-Hsiang Cheng, Chih-Ping Wei and Tseng, 2006; Khoshgoftaar *et al.*, 2012). In the literature, feature selection algorithms use three different methods according to the learning method (Cai *et al.*, 2018). These are Filter, Wrapper, and Embedded Model. In the filter model, the selection is made by considering the relationship between the features and the class label (Cai *et al.*, 2018). The calculation workload is less than the Wrapper model (Cai *et al.*, 2018). The filter model makes the selection of features according to a specific criterion (Cai *et al.*, 2018). The embedded method selects the features in the training process (Cai *et al.*, 2018). All these algorithms still need to be improved regarding performance.

In this article, we have developed an Eta correlation coefficient-based feature selection algorithm like the filter model. The features were selected according to the correlation value between the features and the class label and their performances were tested with kNN, Probabilistic Neural Networks (PNN) and SVMs.

In order to reach the highest level of quality, authors should comply with the rules set out in this template.

The template will be returned to the author for the reorganization of the articles not prepared by the template. Returned articles must be returned after they have been arranged by the rules.

## 2. Materials and Methods

Figure 1 shows the operation steps in this article. First, the feature selection algorithms select the features in the datasets. Then, various classifiers classify features. Finally, the performances of the classifiers are calculated.

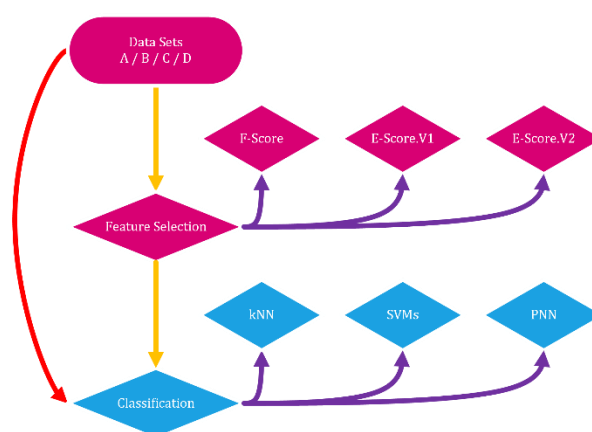


Figure 1. Flow diagram

### 2.1. Sample Datasets

Four datasets (A / B / C / D) were used to test the developed method (Table 1). These are downloaded from the UCI Machine Learning Repository (Andrzejak *et al.*, 2001; Andrzejak RG, Lehnertz K, Rieke C, Mormann F, David P, 2001). The data includes the Electroencephalography (EEG) signal features. Each dataset has two labels (Epilepsy(1)/Non-Epilepsy(2)). Each dataset has 178 properties.

Table 1. Sample datasets

Information	Datasets			
	A	B	C	D
<b>Epilepsy</b>	1150	1150	1150	1150
<b>Non-Epilepsy</b>	1150	1150	1150	1150
<b>Total</b>	2300	2300	2300	2300
<b>Number of Features</b>	178	178	178	178

### 2.2. Eta correlation coefficient

In the literature, there are many correlation calculation methods. However, each data group needs the appropriate unique correlation formula (Alpar, 2010). There are various types of data in the field of machine learning. Class labels are often Unordered Qualitative variables. Eta Correlation Coefficient ( $r_{pb}$ ) is used when calculating the correlation coefficient

between qualitative and continuous numerical variables (Equation 1) (Alpar, 2010). The method changes when the data type changes (Alpar, 2010).

$$r_{pb} = \frac{\bar{Y}_1 - \bar{Y}_0}{s_y} \sqrt{p_0 p_1} \quad (1)$$

In the equation,  $\bar{Y}_0$  and  $\bar{Y}_1$  are the average of the data in class 0 and 1 respectively.  $s_y$  is the standard deviation of all data in both classes (Equation 2).

$$s_y = \sqrt{\frac{\sum Y^2 - \frac{(\sum Y)^2}{n}}{n}} \quad (2)$$

$N$ ,  $N_0$  and  $N_1$  is the number of elements of the total, Class 0 and Class 1 respectively. Equation 3 shows  $p_0$  and  $p_1$ .

$$p_0 = \frac{N_0}{N}, p_1 = \frac{N_1}{N} \quad (3)$$

### 2.3. Feature selection based on Eta correlation coefficient: Eta-Score

In this study, we have developed the Eta correlation coefficient-based feature selection algorithm. The algorithm has two versions (E-Score.V1 - E-Score.V2).

#### 2.3.1. Selection Criteria 1 - E-Score.V1

Figure 2 shows the E-Score.V1 process steps. First, the Eta correlation coefficient ( $Eta$  or  $r_{pb}$ , Equation 1) for each feature is calculated. Second, the Eta threshold is determined ( $Eta$  or  $r_{pb}$ , Equation 1). If  $Eta > Eta_{mean}$ , that feature is selected.

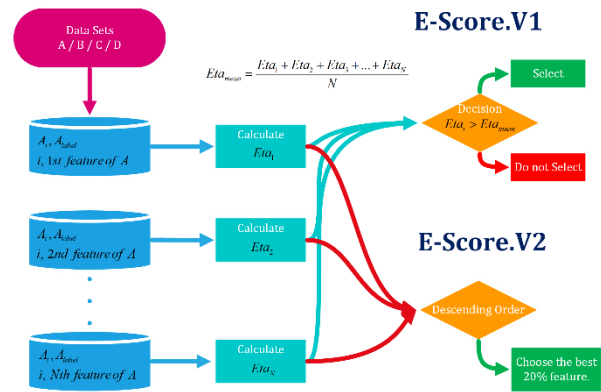


Figure 2. Flow diagram for E-Score

$$Eta_{mean} = \frac{Eta_1 + Eta_2 + Eta_3 + \dots + Eta_N}{N} \quad (4)$$

#### 2.3.2. Selection Criteria 2 - E-Score.V2

Figure 2 shows the E-Score.V2 process steps. First, the  $Eta$  value for the features is sorted in descending order. The first 20% of the features are selected. Eighty percent of the value of a book is hidden in 20 percent of the pages (Koch, 2014). The purpose of this study is to reduce the data size by 80% and to improve system performance. In addition to this aim, performance evaluations were made by selecting the features from 1% to 100% (Figure 3).

#### 2.3.3. Performance Evaluation

kNN, PNN, and SVMs were used to test the proposed algorithms. The performances of these classifiers were measured according to the following criteria. Accuracy rate, sensitivity, specificity, and the working time are performance evaluation criteria. Also, the working time of the algorithm was evaluated.

The datasets for the classification process are divided into two sets: Training (50%) and Test (50%) (Table 2). Besides, different Training / Test rates for E-Score.V1 have been tried, and the accuracy rate is shown graphically for each classifier and each data group (Figure 4).

Table 2. Datasets distribution for the test and training process

Class	For A dataset			For B dataset		
	Training (%50)	Test (%50)	Total	Training (%50)	Test (%50)	Total
Epilepsy	1150	1150	2300	1150	1150	2300
Non-Epilepsy	1150	1150	2300	1150	1150	2300
Total	2300	2300	4600	2300	2300	4600
Class	For C dataset			For D dataset		
	Training (%50)	Test (%50)	Total	Training (%50)	Test (%50)	Total
Epilepsy	1150	1150	2300	1150	1150	2300
Non-Epilepsy	1150	1150	2300	1150	1150	2300
Total	2300	2300	4600	2300	2300	4600

### 3. Results

This study aims to develop a new feature selection algorithm in the field of machine learning. For this, we established the Eta correlation coefficient-based E-Score Feature Selection Algorithm with two different versions (Section 2.3.). The improved method has been tested in different classifiers according to some

performance criteria (Section 2.3.3.). The E-Score was also compared with the F-Score Feature Selection algorithm available in the literature (Polat and Güneş, 2009).

The working time of the E-Score algorithm was measured for four different datasets (Table 3). Besides, the working time performance of the algorithm was compared with the F-Score feature selection algorithm (Table 3).

**Table 3.** Results of E-Score working time evaluation

Datasets	All Features	F-Score		Eta-Boost.V1		Eta-Boost.V2	
		Number	Time (sec)	Number	Time (sec)	Number	Time (sec)
A	178	60	0.019	79	0.171	36	0.170
B	178	68	0.020	85	0.166	36	0.165
C	178	62	0.019	77	0.166	36	0.165
D	178	54	0.019	73	0.166	36	0.164

sec: Second

kNN, PNN and SVMs classifiers evaluated the performance of the E-Score algorithm. According to the performance results, kNN classifier and for each dataset (A/B/C/D), E-Score.V2 is the best-performing feature selection algorithm among other algorithms (Table 4). Besides, E-Score.V1 has similar performance with F-Score (Table 4).

In the PNN classifier, the performance of feature selection algorithms depends on the datasets (A/B/C/D) (Table 4). E-Score.V1 is the best feature selection algorithm for SVMs (Table 4). When the feature selection algorithms examined the effects of the classifiers operating time, E-Score.V2 most successful feature selection algorithm (Table 4).

**Table 4.** Evaluation of the performance of the E-Score feature selection algorithm

A												
Classifier	kNN				PNN				SVMs			
Performance	Acc	Sen	Spe	T	Acc	Sen	Spe	T	Acc	Sen	Spe	T
All Features	86.78	0.74	1.00	0.33	93.17	0.87	1.00	2.56	99.61	0.99	1.00	0.12
F-Score	88.91	0.78	1.00	0.11	92.17	0.94	0.90	0.79	99.13	0.98	1.00	0.08
Eta-Boost.V1	87.35	0.75	1.00	0.13	91.91	0.90	0.94	1.04	99.26	0.99	1.00	0.08
Eta-Boost.V2	90.39	0.81	1.00	0.06	63.26	0.98	0.29	0.52	98.96	0.98	1.00	0.07
B												
Classifier	kNN				PNN				SVMs			
Performance	Acc	Sen	Spe	T	Acc	Sen	Spe	T	Acc	Sen	Spe	T
All Features	85.17	0.70	1.00	0.32	91.09	0.84	0.98	2.55	98.13	0.96	1.00	0.14
F-Score	87.30	0.75	1.00	0.11	78.57	0.87	0.70	0.96	96.74	0.95	0.98	0.28
Eta-Boost.V1	86.35	0.73	1.00	0.15	88.43	0.88	0.89	1.13	97.57	0.96	0.99	0.11
Eta-Boost.V2	88.43	0.77	1.00	0.07	73.61	0.93	0.54	0.52	95.74	0.94	0.98	0.24
C												
Classifier	kNN				PNN				SVMs			
Performance	Acc	Sen	Spe	T	Acc	Sen	Spe	T	Acc	Sen	Spe	T
All Features	83.09	0.66	1.00	0.32	89.13	0.80	0.98	2.65	97.91	0.97	0.99	0.14
F-Score	85.96	0.72	1.00	0.11	81.00	0.91	0.71	0.80	96.48	0.95	0.98	0.11
Eta-Boost.V1	85.09	0.70	1.00	0.12	75.83	0.88	0.63	1.02	96.78	0.96	0.98	0.10
Eta-Boost.V2	87.65	0.75	1.00	0.07	59.70	0.96	0.23	0.53	95.96	0.95	0.97	0.09
D												
Classifier	kNN				PNN				SVMs			
Performance	Acc	Sen	Spe	T	Acc	Sen	Spe	T	Acc	Sen	Spe	T
All Features	81.43	0.64	0.99	0.32	48.78	0.90	0.08	2.62	94.30	0.96	0.93	0.34
F-Score	83.78	0.69	0.98	0.08	48.91	0.96	0.02	0.75	93.39	0.93	0.93	0.25
Eta-Boost.V1	82.74	0.67	0.99	0.12	49.48	0.96	0.03	0.97	93.78	0.95	0.93	0.27
Eta-Boost.V2	85.26	0.73	0.97	0.07	49.43	0.98	0.01	0.52	92.70	0.93	0.92	0.24

Acc Accuracy Rate (%) , Sen Sensitivity, Spe Specificity, T Time (second)

E-Score.V2 selects only the first 20% of all features. The percentage change can increase performance (Figure 3). In kNN and SVMs, small performance

changes were observed due to the number of features (Figure 3). However, PNN performance is highly variable depending on the number of features (Figure 3).

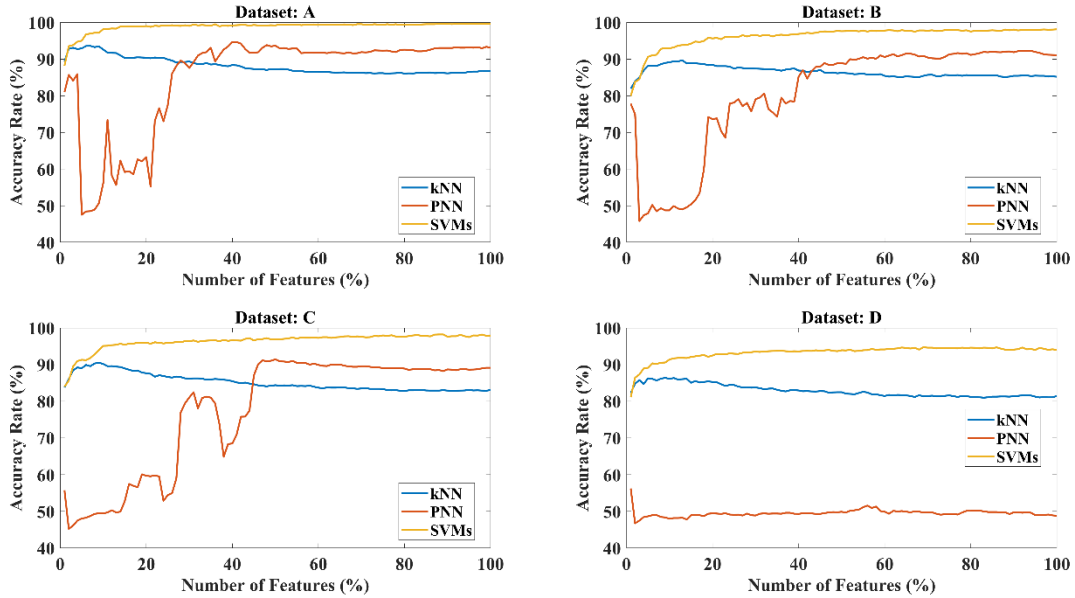


Figure 3. For E-Score.V2, Accuracy rates for selected properties in different percentages

Training and Test rates are 50% and 50% for classification. For the E-Score.V1 algorithm, the change of the test data was monitored from 5% to 95% (Figure

4). If the test data exceeds 65-70%, system performance decreases (Figure 4). When the test dataset is 50%, the system performance is maximum (Figure 4).

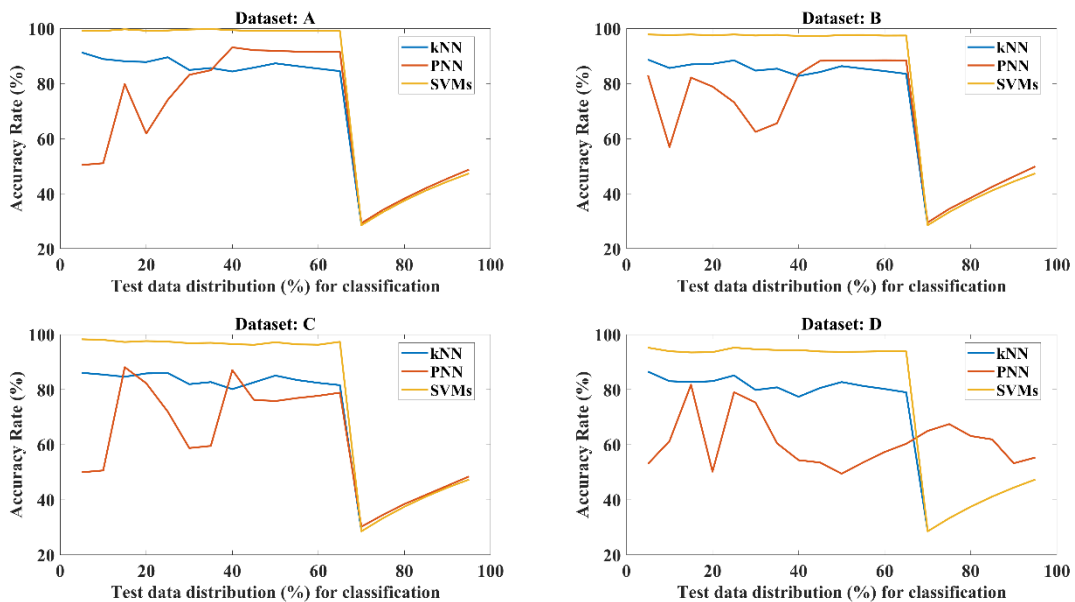


Figure 4. E-Score.V1 feature selection algorithm performance for different Training / Test distributions



#### 4. Discussion and Conclusions

A new feature selection algorithm has been developed with this study. Feature selection algorithms are an essential part of machine learning. These algorithms are required to shorten the duration of learning and to minimize the number of features (Polat and Güneş, 2009; Guan *et al.*, 2014; Kavsaoglu, Polat and Bozkurt, 2014; Cai *et al.*, 2018). The number of features selected by the E-Score method is between 20-40% compared to the total number of features. This reduces the workload considerably. Besides, E-Score increases the classification performance of the system. E-Score performance is quite good compared to the F-Score feature selection algorithm in the literature (Polat and Güneş, 2009). E-Score has reduced the workload and improved the performance of the system, such as feature selection algorithms in the literature (Polat and Güneş, 2009; Guan *et al.*, 2014; Kavsaoglu, Polat and Bozkurt, 2014; Cai *et al.*, 2018).

E-Score is a correlation-based feature selection algorithm. As the E-Score is statistical-based, the correlation between features and intergroup correlation can be accurately estimated (Alpar, 2010). However, the method can only be applied between qualitative and continuous numerical variables. For other data types, similar process with E-Score is recommended, but it is recommended to use the correlation formulas according to the data type.

According to the results obtained in the study, each feature selection algorithm does not adapt to each dataset. Performance has improved. However, there is no corresponding improvement in each dataset.

As a result, when the E-Score feature selection algorithm is examined regarding performance, it is considered to be a quality method that can be used in the field of machine learning.

#### Acknowledgment

Matlab-based codes for the E-Score Feature Selection Algorithm are available from [GitHub](#).

Access connection:  
[https://github.com/MKUCARE/E\\_Score\\_Feature\\_Selection.git](https://github.com/MKUCARE/E_Score_Feature_Selection.git)

Please refer to each publication you use the method or code.

#### References

Alpar, R. (2010) *Applied Statistic and Validation - Reliability*. Detay Publishing. Available at: [https://books.google.com.tr/books/about/Uygulamalı\\_istatistik\\_ve\\_geçerlik\\_güv.html?id=ITk1MwEACAAJ&pgis=1](https://books.google.com.tr/books/about/Uygulamalı_istatistik_ve_geçerlik_güv.html?id=ITk1MwEACAAJ&pgis=1) (Accessed: 11 January 2016).

Andrzejak, R. G. *et al.* (2001) 'Indications of nonlinear deterministic and finite-dimensional structures in time series of brain electrical activity: Dependence on recording region and brain state', *Physical Review E - Statistical Physics, Plasmas, Fluids, and Related Interdisciplinary*

*Topics*, 64(6), p. 8. doi: 10.1103/PhysRevE.64.061907.

Andrzejak RG, Lehnertz K, Rieke C, Mormann F, David P, E. C. (2001) *UCI Machine Learning Repository: Epileptic Seizure Recognition Data Set*, UCI. Available at: <https://archive.ics.uci.edu/ml/datasets/Epileptic+Seizure+Recognition> (Accessed: 14 August 2018).

Cai, J. *et al.* (2018) 'Feature selection in machine learning: A new perspective', *Neurocomputing*. Elsevier, 300, pp. 70–79. doi: 10.1016/J.NEUCOM.2017.11.077.

Elghazel, H. and Aussem, A. (2015) 'Unsupervised feature selection with ensemble learning', *Machine Learning*. Springer US, 98(1–2), pp. 157–180. doi: 10.1007/s10994-013-5337-8.

Goltsev, A. and Gritsenko, V. (2012) 'Investigation of efficient features for image recognition by neural networks', *Neural Networks*. Pergamon, 28, pp. 15–23. doi: 10.1016/J.NEUNET.2011.12.002.

Guan, D. *et al.* (2014) 'A Review of Ensemble Learning Based Feature Selection', *IETE Technical Review*, 31(3), pp. 190–198. doi: 10.1080/02564602.2014.906859.

Huang, D.-S. (1999) 'Radial Basis Probabilistic Neural Networks: Model and Application', *International Journal of Pattern Recognition and Artificial Intelligence*. World Scientific Publishing Company, 13(07), pp. 1083–1101. doi: 10.1142/S0218001499000604.

Kavsaoglu, A. R., Polat, K. and Bozkurt, M. R. (2014) 'A novel feature ranking algorithm for biometric recognition with PPG signals.', *Computers in biology and medicine*, 49, pp. 1–14. doi: 10.1016/j.combiomed.2014.03.005.

Khoshgoftaar, T. *et al.* (2012) 'First Order Statistics Based Feature Selection: A Diverse and Powerful Family of Feature Selection Techniques', in *2012 11th International Conference on Machine Learning and Applications*. IEEE, pp. 151–157. doi: 10.1109/ICMLA.2012.192.

Khotanzad, A. and Hong, Y. H. (1990) 'Rotation invariant image recognition using features selected via a systematic method', *Pattern Recognition*. Pergamon, 23(10), pp. 1089–1101. doi: 10.1016/0031-3203(90)90005-6.

Koch, R. (2014) *The 80/20 Principle and 92 Other Powerful Laws of Nature: The Science of Success*. Available at: <https://www.amazon.com/80-20-Principle-Secret-Achieving/dp/1486213421> (Accessed: 23 September 2018).

Li, Y., Gao, S.-Y. and Chen, S. (2012) 'Ensemble Feature Weighting Based on Local Learning and Diversity', *AAAI*. Available at: <https://www.semanticscholar.org/paper/Ensemble-Feature-Weighting-Based-on-Local-Learning-Li-Gao/733e4973aec6d9de139781a76ca2f6b3f05b293b> (Accessed: 24 September 2018).

Polat, K. and Güneş, S. (2009) 'A new feature selection method on classification of medical datasets: Kernel F-score feature selection', *Expert Systems with Applications*, 36(7), pp. 10367–10373. doi: 10.1016/j.eswa.2009.01.041.

Tsang-Hsiang Cheng, Chih-Ping Wei and Tseng, V. S. (2006) 'Feature Selection for Medical Data Mining: Comparisons of Expert Judgment and Automatic Approaches', in *19th IEEE Symposium on Computer-Based Medical Systems (CBMS'06)*. IEEE, pp. 165–170. doi: 10.1109/CBMS.2006.87.



# A Survey on Anomaly Detection and Diagnosis Problem in the Space System Operation

Seçil Taburoğlu\*

Senior Software Design Engineer Turkish Aerospace Industries Ankara, Türkiye

## Abstract

Spacecraft telemetry data is transferred from satellite to ground control station. The data contains not only information about health status of the satellite but also contains response messages to telecommand (telecommand data is send to spacecraft from ground control station) data. Telemetry data can indicate data error, communication link failure, sensor error, equipment and electronic devices failure. Safety and reliability are provided by telemetry and telecommand data. The most important subjects are safety and reliability for space mission. Therefore, telemetry data should be analyzed and take measures against to attack or unexpected situation. Various intelligent anomaly detection methods are proposed in the literature. Supervised/unsupervised (machine learning) anomaly detection approaches and data mining technology are the most used methods. This paper is a literature review for anomaly detection approaches in space system operation. Anomaly detection techniques have been investigated in the literature, but studies on space domain is quite limited. It is considered to contribute to literature in terms of that.

**Keywords:** Anomaly detection; Machine learning, Data mining; Spacecraft health.

## Uydu Sistemlerinde Anomali Tespiti ve Tanısı Sorununa Yönelik Yapılmış Bir


### Araştırma

#### Öz

Uydu uzölçüm verileri uydudan yer kontrol istasyonuna transfer edilir. Bu veri sadece uydu sağlık verilerini içermekle kalmayıp uzkomut (telekomut verisi, yer kontrol istasyonundan uyduya iletilir) mesajlarına yanıtları da içerir. Telemetri verileri; veri hatası, iletişim bağlantısı hatası, sensör hatası, ekipman ve elektronik cihaz arızası gibi bilgileri içerebilir. Güvenlik ve güvenilirlik, telemetri ve telekomut verilerine bakılarak sağlanılmaktadır. Uzay görevinde en önemli konular güvenlik ve güvenilirlik olduğu için telemetri verileri analiz edilmeli ve saldırılara veya beklenmeyen durumlara karşı önlemler alınmalıdır. Literatürde çeşitli akıllı anomali saptama yöntemleri önerilmiştir. Gözetimli/Gözetimsiz (makine öğrenmesi) anomali tespiti ve veri madenciliği yaklaşımları en çok kullanılan yöntemlerdir. Bu makale uzay sistemlerinde anomali tespiti hakkında genel bakış sunmaktadır.

**Anahtar Kelimeler:** Anomali tespiti; Makine öğrenmesi; Veri madenciliği; Uydu sağlığı.

\* Corresponding Author.

 E-mail: [secil.taburoglu@tai.com.tr](mailto:secil.taburoglu@tai.com.tr)

Received : Jan 7, 2019

Revision : Jan 29, 2019

Accepted : Jan 30, 2019

## 1. Introduction

### A. Anomaly detection

Anomaly detection (Chandola et al., 2009) is a technique used to identify dataset which does not conform to an expected behavior or other items in a dataset. In general, these unexpected behaviors are defined as attacks. Whereas these situations can be unexpected behaviors which are previously not known, rather than an attack.

The anomaly detection provides very significant and important information about the system. Also, anomaly detection helps prevent potential malfunctions and serious errors. This improves system security and reliability.

### B. Spacecraft anomaly detection

Anomaly detection is an important topic for space system, because so much money and time are spent. Also satellites have become incredibly useful, especially for meteorology, communication, and navigation and military. These domains are costly and safety critical. Therefore failures are not acceptable. The anomaly detection can help carry out fault diagnosis and prevent the occurrence of potential failures.

There are many anomaly detection methods exist, in general following steps (Gilmore et al., 2016) are done.

- Data Preprocessing and Feature extraction: A list of related parameters for components or subsystems of the spacecraft system are selected
- Model Generating: Model is generated on the normal or abnormal behavior of the spacecraft system.
- Detecting: Statistical based, knowledge based or machine learning and data mining algorithms are used.

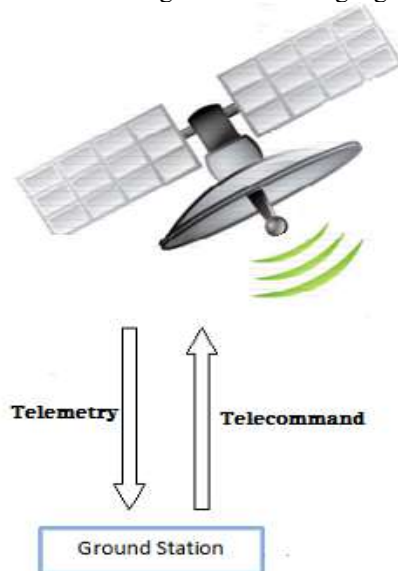


Figure 1 Telemetry – Telecommand Communications

Telemetry data include some sensor values, such as temperature, voltage, angular velocity and temperatures which have low and high limit value. That's why limit checking

(Machida et al., 2006) is one alternative for anomaly detection. But for some cases when limit values are normal, anomalies can exist, this means that some class of anomalies occur without violating the limits on the variables. In order to take measures before the situation occurs, these anomalies should be predicted. Also domain experts or operators always should monitor and review out of limit values. If the limit values are inappropriate, anomaly will detected, else false alarms will be generated and real anomalies may be missed. Therefore, to overcome these problems machine learning methods are used instead of limit checking. Also, Machida, K., (Machida et al., 2006) have developed a new method for limit checking. They have combined limit checking and Sparse Bayesian Learning (or Relevance Vector Machine). RVM is used to learn a model of high and low limit values from old normal telemetry. The resulting models are then used in later operations to detect anomalies for target variables online.

Data Mining Techniques (clustering and classification based) and hybrid approach are the most important methods. An important advantage of these approaches compared with the expert systems and model-based approach is that it does not require complete and accurate expert knowledge or models.

There are a large number of surveys in the literature about the detection of anomaly but there is very limited number of in terms of space domain. This survey aims to give a closer look on these enhancements and to summarize and categorize some articles presented in this field according to the various anomaly detection techniques. Also, in this review, the methods in the literature are shortly summarized for the industry (real world) which have just started to work on the anomalies in the spacecraft.

## 2. Methodology

There are two main approach for anomaly detection: knowledge-driven approach and data-driven approach (Yairi et al., 2017). In knowledge-driven approach, highly accurate results can be obtained, if knowledge is accurate. However, this method is costly because it require expensive expert knowledge and manual checking. Data-driven approach includes machine learning methods (Liu, D et al., 2017; Biswas et al., 2016; Gao et al., 2012a; Machida et al., 2006; Shi, et al., 2017; Azevedo et al., 2012) (classification (Liu et al., 2017; Machida et al., 2006; Gao et al., 2012), clustering (Biswas et al., 2016; Gao et al., 2012; Shi et al., 2017; Azevedo et al., 2012)). This approach is better than knowledge-driven in terms of cost. Also it is an automatic detection approach; however, most of the time highly accurate results are not available.

### Machine Learning Approaches (Data-driven approaches)

Reviewed methods are listed in **Table 1**. This table includes dataset and methods' pros and cons.

#### A. Classification approaches (supervised learning)

Labelled data is required for classification. Some of the data is used for training and labelled data is acquired on this method.

This method classifies data divided into two categories as normal and abnormal. This approach is not fully automated and expert knowledge is required. In general, following algorithms are used to anomaly detection:

**k-Nearest Neighbor (KNN) classification:** KNN algorithm classifies data sets based on their similarity with neighbors. D. Liu, J. Pang (Liu, et al., 2017) have used KNN classification with enhanced similarity measures on actual telemetry data. Telemetry data has been divided into time series. Mahalanobis Distance and DTW Distance(Liu, et al., 2017) algorithm are used on telemetry data to calculate distance measure. Distance measure is used to measure the similarity of the telemetry data. KNN algorithm classifies new telemetry data according to the similarity measures. Liu, J. Pang (Liu et al., 2017), have obtained different results for different dataset. In general, they can obtain satisfied results with Mahalanobis Distance and DTW Distance.

**Support Vector Machines (SVM):** SVM supports classification and regression, which is useful for statistical learning theory, and fully recognizes these factors. SVM can be defined as a field of pattern recognition. SVM classifies data into two classes. Yu Gao, Tianshe Yang, Nan Xing (Gao et al., 2012) have used binary SVM to detect anomalies. This method is not valid for fault diagnosis because fault diagnosis has to deal with multi-type malfunctions. So, they have proposed two approach; combining several binary SVMs and implementing multi-class classification. They have verified approach with actual satellite data and simulated by Matlab/Simulink. They have obtained good results for two cases (accuracy: 99.2%, 97.4%), and performance is also satisfactory. Also they used Principal Component Analysis (PCA) for feature extraction. PCA is a dimension-reduction mathematical technique. This technique is used to reduce large datasets to smaller sets. Feature identification and reduction is an important step for anomaly detection.

### B. Clustering approaches (unsupervised learning)

This method uses unlabeled telemetry data. Therefore there is no need for extra space domain knowledge. The object of unsupervised learning technique is to find similar objective data points and combines similar data points. Gao, Yu (Gao et al., 2012) have detected analogies in runtime by using unsupervised learning. Firstly, they automatically detects and removes abnormal data from the archived (historical) telemetry data to construct normal behavior model. Then, according to the normal behavior model. They have detected anomalies in real time data. Single-linkage clustering is used for detection. Single linkage is a type of hierarchical clustering methods. In this method, the similarity of the two clusters is the similarity of the most similar members. This study is important, because this proposed new approach does not require expert knowledge (low cost). Also, there is no cost of data training (labelled data). According to the results, when detection ratio increases, false positive ratio increases. Supervised learning results are higher than unsupervised learning results.

Azevedo, Denise Rotondi, Ana Maria Ambrósio, and Marco Vieira (Azevedo et al., 2012) have used K-means and Expectation Maximization(EM) for clustering. K-means is an unsupervised learning algorithm. This algorithm divides the data set with the number of groups represented by the variable K. EM algorithm can be thought of as an extension of the k-means algorithm. EM is an iterative algorithm for learning probabilistic categorization model from unlabeled telemetry data. They have calculated the dissimilarity indexes using both Euclidean and Manhattan distances, then they have used clustering algorithms.

Fuertes, Sylvain, Barbara Pilastre, and Stéphane D'Escrivan (Fuertes et al., 2018) have compared three unsupervised algorithms: One-Class Support Vector Machine (OC-SVM is a type of Support Vector Machines) (Rana, Divya, 2015), Density-based spatial clustering of applications with noise (DBSCAN) and k-Nearest Neighbors algorithm. They have obtained approximate results. They argue that the success of the method depends on the choice of features rather than machine learning methods.

### C. Hybrid approaches

Hybrid approach combines unsupervised and supervised methods. This approach is not fully automated, expert knowledge is required because of supervised learning techniques. The main goal is to develop more automated methods.

Biswas, Gautam (Biswas et al., 2016) have used unsupervised learning algorithm to cluster time series data. Their hypotheses is that the larger groups of clusters will show normal (routine telemetry data – maneuver, expected orbit propagation etc.) operations. On the other hand smaller groups will show anomalous situations. They have used hierarchical clustering method. Then they have verified results by a supervised approach of consulting domain experts. Although they achieve good results, this method is not automate and cheap.

Machida, K. (Machida et al., 2006) has used Dynamic Bayesian Network(DBN) to learn model from past telemetry data in normal operation by Expectation-Maximization algorithm. It is particularly suited for modeling hybrid systems involving both continuous variables and discrete variables. Also the DBN is well-suited for time-series space data.

### D. Other approaches

Auto-Regressive Integrated Moving-Average (ARIMA) is used by Ibrahim, Sara K (Ibrahim et al., 2018). ARIMA is a model for time series forecasting. They have applied different machine learning techniques and compared to each other. According to the results, ARIMA have highest prediction accuracy but in case of short-term forecasting models. Also they have used PCA for dimension reduction but they lost some important data, consequently they preferred domain expert knowledge.

**Table 1** Articles classification

Paper	Methods used	Methodology	Dataset	Pros and Cons
(Liu et al., 2017)	k-Nearest Neighbor (KNN) classification	k-Nearest Neighbor (KNN) classification is used and similarity measure is enhanced. Mahalanobis Distance and DTW Distance are used.	The Wafer data sets by Carnegie Mellon University (CMU) and ECG data sets, RobotFailure data sets by University of California at Irvine	(-) Labelled data as required (this is the meaning of expert knowledge is required) (-) Data set results have very different accuracy(RobotFailure dataset results have accuracy of 69.47%, while Wafer has accuracy of 98.66% ) (+) Mahalanobis distance and the DTW distance, have high accuracy
(Biswas, et al., 2016)	Mixed Method (Unsupervised and supervised methods)	Unsupervised learning are used to cluster large telemetry database of time series data. Then, experts opinion is used for some inputs(supervised method)	Telemetry data from the EPS of the LADEE spacecraft.	(+) Unsupervised learning approach does not required domain expert knowledge (-) Anomaly detection is not being done automatically. (Some data is analyzed by expert human) (-) Evaluated results are not given.
(Gao et al., 2012)	Unsupervised learning, Nearest Neighbor Algorithm	Firstly dataset is limited with Euclidean distance. Then k-Nearest Neighbors (kNN) algorithm is used to anomaly detection.	TM data of the power subsystem of in-orbit satellite.	(+) Unsupervised learning approach does not required domain expert knowledge (-) False positive rate is high when detection rate is high
(Machida et al., 2006)	Machine learning (ML, Supervised learning) and data mining (DM) technology.	Anomaly Detection from Telemetry with Kernel Principal Component Analysis, Dynamic Bayesian Networks	JAXA (Japan Aerospace Exploration Agency) and SCC (Space Communication Corporation)	(-) Semi-automatic therefore domain knowledge is required. Also system behavior models are required. (-) There are some problems in practical use (+) Satisfactory results have been obtained
(Gao et al., 2012)	Supervised learning	Principal Component Analysis (PCA) and Support Vector Machines (SVM).	Data is received from actual in-orbit satellite and simulated by matlab/Simulink.	(+) High accuracy (-) Training dataset should be generated, therefore domain knowledge is required.
(Shi et al., 2017)	Clustering (Unsupervised Learning)	Satellite telemetry time series are clustered with Special Points Series Segmentation	Open datasets	(+) Low computational time (+) Reduced noise effect (+) Unsupervised learning approach does not required domain expert knowledge
(Azevedo et al., 2012)	Clustering (Unsupervised Learning)	K-means algorithm and Expectation Maximization are used. Then the results are compared to each other.	Two different satellites data	(-) Feature selection method is not efficient (-) False negative results are obtained when satellite data changes related to satellite aging. (+) They have obtain satisfactory result if there is at least one out of limit value.
(Ibrahim et al., 2018)	Machine learning (ML) and data mining (DM) techniques	Autoregressive integrated moving average (ARIMA), Multi-Layer Perceptron (MLP), Recurrent Neural Network (RNN), Long Short-Term Memory Recurrent Neural Network (LSTM RNN), Deep Long Short-Term Memory Recurrent Neural Networks (DLSTM RNNs), Gated Recurrent Unit Recurrent Neural Network (GRU RNN), and Deep Gated Recurrent Unit Recurrent Neural Networks (DGRU RNNs) algorithms are used. Then the	Real telemetry data of Egyptsat-1 satellite	(+) Low computational time (especially for ARIMA and MLP models) (-) Training dataset should be generated, therefore domain knowledge is required. (+) High accuracy (they used small dataset and obtained highly accurate results for ARIMA)

		results are compared to each other.				
(Fuertes et al., 2018)	Clustering (Unsupervised Learning)	One-Class Machine Neighbors algorithm, DBSCAN	Support Vector (OC-SVM), k-Nearest Neighbors algorithm, DBSCAN	Vector	CNES operated satellite data	(+) High accuracy, but depends on the application context and feature selections.

### 3. Conclusions

The most critical and outstanding articles are reviewed for this survey. This survey shows the most used methods for anomaly detection in space domain. Through this literature review, several conclusions have been reached about anomaly detection and supervised/unsupervised learning algorithms for use in satellite anomaly detection system. Most of the proposed solutions are centered around data mining techniques. The work in fault analysis has mainly focused on supervised learning techniques. Naïve Bayes, Support Vector Machines and k-Nearest Neighbor (KNN) classification are the most frequently used machine learning algorithms. Nowadays supervised learning techniques are no longer used because they need knowledge of expertise. There is limited unsupervised study but the number of papers is increasing fast. Hybrid approaches are often used because they provide better results. However, there is a fact that abnormal detection systems still have high false positive rates and studies on this area still continues.

### References

- Azevedo, D.R., Ambrósio, A.M. and Vieira, M., 2012, May. Applying data mining for detecting anomalies in satellites. In 2012 Ninth European Dependable Computing Conference (pp. 212-217). IEEE..
- Biswas, G., Khorasgani, H., Stanje, G., Dubey, A., Deb, S. and Ghoshal, S., 2016. An application of data driven anomaly identification to spacecraft telemetry data. In Prognostics and Health Management Conference.
- Chandola, V., Banerjee, A. and Kumar, V., 2009. Anomaly detection: A survey. ACM computing surveys (CSUR), 41(3), p.15..
- Fuertes, S., Pilastre, B. and D'Escrivan, S., 2018. Performance assessment of NOSTRADAMUS & other machine learning-based telemetry monitoring systems on a spacecraft anomalies database. In 2018 SpaceOps Conference (p. 2559).
- Gao, Y., Yang, T., Xing, N. and Xu, M., 2012, July. Fault detection and diagnosis for spacecraft using principal component analysis and support vector machines. In Industrial Electronics and Applications (ICIEA), 2012 7th IEEE Conference on (pp. 1984-1988). IEEE.
- Gao, Y., Yang, T., Xu, M. and Xing, N., 2012, January. An unsupervised anomaly detection approach for spacecraft based on normal behavior clustering. In Intelligent Computation Technology and Automation (ICICTA), 2012 Fifth International Conference on (pp. 478-481). IEEE.
- Gilmore, C. and Haydaman, J., 2016, January. Anomaly Detection and Machine Learning Methods for Network Intrusion Detection: an Industrially Focused Literature Review. In Proceedings of the International Conference on Security and Management (SAM) (p. 292). The Steering Committee of The World Congress in Computer Science, Computer Engineering and Applied Computing (WorldComp).
- Ibrahim, S.K., Ahmed, A., Zeidan, M.A.E. and Ziedan, I., 2018. Machine Learning Methods for Spacecraft Telemetry Mining. IEEE Transactions on Aerospace and Electronic Systems.
- Liu, D., Pang, J., Xu, B., Liu, Z., Zhou, J. and Zhang, G., 2017, August. Satellite Telemetry Data Anomaly Detection with Hybrid Similarity Measures. In Sensing, Diagnostics, Prognostics, and Control (SDPC), 2017 International Conference on (pp. 591-596). IEEE.
- Machida, K., Fujimaki, R., Yairi, T., Kawahara, Y. and Sato, Y., 2006. Telemetry-mining: A machine Learning Approach to Anomaly detection and fault Diagnosis for space Systems. In 2nd IEEE International Conference on Space Mission Challenges for Information Technology, IEEE.
- Rana, Divya, 2015, One Class SVM Vs SVM Classification Divya Rana.
- Shi, X., Pang, J., Liu, D. and Peng, Y., 2017, July. Satellite telemetry time series clustering with improved key points series segmentation. In Prognostics and System Health Management Conference (PHM-Harbin), 2017 (pp. 1-7). IEEE.
- Yairi, T., Takeishi, N., Oda, T., Nakajima, Y., Nishimura, N. and Takata, N., 2017. A Data-Driven Health Monitoring Method for Satellite Housekeeping Data Based on Probabilistic Clustering and Dimensionality Reduction. IEEE Transactions on Aerospace and Electronic Systems, 53(3), pp.1384-1401.