Journal of Soft Computing and Artificial Intelligence

ISSN: 2717-8226

Journal of Soft Computing and Artificial Intelligence

Journal homepage: https://dergipark.org.tr/en/pub/jscai

International Open Access

Volume 04
Issue 01

June, 2023

Journal of Soft Computing and Artificial Intelligence (JSCAI) is an international peer-reviewed journal that publishes integrated research articles in all areas of soft computing and artificial intelligence. The aim of the JSCAI journal is to provide a platform for researchers, professionals, and academicians around the world to combine and exchange new developments and their applications in various areas of soft computing and artificial intelligence. Journal of Soft Computing and Artificial Intelligence (JSCAI) is an international peer-reviewed journal that publishes integrated research articles in all areas of soft computing and artificial intelligence. The journal covers all branches of engineering, including mechanics, computer science, electronics, energy, aerospace engineering, materials science, nuclear engineering, systems analysis, alternative technologies, etc.

JSCAI publication, which is open access, is free of charge. There is no article submission and processing charges (APCs).

JSCAI is indexed & abstracted in:

Crossref (Doi beginning: 10.55195/jscai..xxxxxx)

Directory of Research Journals Indexing (DRJI)

Google Scholar

Index Copernicus (ICI Journal Master List)

OpenAIRE

Asos Index

Directory of Open Access scholarly Resources (ROAD)

Authors are responsible from the copyrights of the figures and the contents of the manuscripts, accuracy of the references, quotations and proposed ideas and the Publication Ethics (https://dergipark.org.tr/en/pub/jscai/policy)

Journal of Soft Computing and Artificial Intelligence (JSCAI) allows the author(s) to hold the copyright of own articles.

©
JSCAI
25 June 2023

This work is licensed under a Creative Commons Attribution 4.0 International License.

Table of Contents

Journal of Soft Computing and Artificial Intelligence

Journal homepage: https://dergipark.org.tr/en/pub/jscai

International Open Access

Volume 04
Issue 01

June, 2023

Letter to Editor

A User and Entity Behavior Analysis for SIEM Systems: Preprocessing of The Computer Emergency and Response Team Dataset

Yasin Görmez [1], ***Halil Arslan***[2], ***Yunus Emre Işık***[1], ***İbrahim Ethem Dadaş***[3]

[1] [1]*Management Information Systems, Faculty of Economics and Administrative Sciences. Sivas Cumhuriyet University, 58050, Sivas, Turkiye*
[2]*Computer Engineering, Engineering Faculty. Sivas Cumhuriyet University, 58050, Sivas, Turkiye*
[3] *Principal Constultant. Detay Danışmanlık, 34674, İstanbul, Turkiye*

ARTICLE INFO

Article history:
Received **December 2, 2022**
Revised **February 22, 2023**
Accepted **March 8, 2023**

Keywords:
User and Entity Behavior Analysis
Preprocessing
Classification
CERT
Security Information and Event Management

ABSTRACT

A lot of work has been done to prevent attacks from external sources and a great deal of success has been achieved. However, studies to detect internal attacks aren't sufficient today. One of the most important studies for the detection of insider attacks is User and Entity Behavior Analysis (UEBA). In this letter, UEBA studies in the literature were reviewed and The Computer Emergency and Response Team Dataset was analyzed (CERT). For this purpose, preprocessing and feature extraction steps were applied on CERT datasets. Several log files combined with respect to user and for each user the number of activities in the specified time interval were obtained. The python code of these preprocessing and feature extraction steps were shared as open source in GitHub platform. In the final phase, future analysis was described and UEBA system planned to be designed was explained.

1. Introduction

Many of the companies have started to carry out their vital business processes with digital systems thanks to developments in technology. They store all important information such as employee, marketing strategy, application info and project documents in the computer systems. These systems have been made very resistant to external attacks by using applications such as intrusion detection system (IDS). However, IDS like applications do not take precautions against insider attacks.

Insider attacks represent the malicious action that performed by the employee of the organizations. These employees not only have authorization to access companies' own systems but also, they have authorization to access customers' systems. Therefore, insider attacks are very dangerous for the companies. Especially large companies suffer from insider attackers, and they try to control whole of their system using Security Information and Event Management (SIEM) solutions. SIEM solutions are the platform where all activities in a network are recorded and reported in real time. It is possible to make User and Entity Behavior Analysis (UEBA) with real-time data reported on SIEM platforms. UEBA is the best way to capture malicious action from insider employees. A UEBA integrated with SIEM systems is very helpful in early detection of internal attacks that are very dangerous for companies [1].

To date, many studies have been conducted about attack prediction of SIEM and UEBA. Anumol proposed an intrusion prediction system (IPS), which is called open-source security information management (OSSIM), for SIEM framework to perform event analysis using data mining techniques

[1] Corresponding author
e-mail: ysngrmz@gmail.com

DOI: 10.55195/jscai.1213782

[2]. Laue et al. developed an open source SIEM framework within the GLACIER projects. This system does not require any licensing fees and it contains advance algorithms such as artificial intelligence, data collection and anomaly detection. It is also possible to monitor all event using powerful user interface [3]. Asanger and Hutchison applied k-nearest neighbor algorithm on datasets that contain 15 million Windows security events from various perspectives to show performance of unsupervised anomaly detection systems [4]. Goldstein et al. applied six different algorithms on the same dataset using sliding window technique and they showed that the best algorithm is global k-nearest neighbor algorithm [5]. Lukashin et al. proposed novel UEBA architecture that integrated to SIEM platforms, and they achieved the 97.49%, 47.71% and 54.40% accuracy, precision and recall respectively [6]. Tian et al. proposed long short-term memories (LSTM) for UBEA, and they applied multimodal based system on the CERT dataset. Their system achieved 97% accuracy, 98.84% true positive rate and 14.81% false positive rate [7]. Lee and Zincir-Heywood compared decision tree and self-organizing maps on CERT dataset, and they showed that self-organizing maps is better than the decision tree. In addition to this, they performed these analyses on both daily and weekly data and it is seen that weekly data is better than the daily data to detect anomalies [8]. Sharma et al. extracted activity-based feature from CERT dataset, and they proposed long short-term memories based autoencoder model. They achieved 90.17% accuracy, 91.03% recall, 9.84% false positive rate and 90.15% true negative rate with proposed model [9]. Al-Shehari and Alsowail compared logistic regression, decision tree, random forest, k-nearest neighbor and kernel support vector machines and they showed that the best method is logistic regression for CERT dataset. In addition to this they addressed the imbalance problem of this dataset, and they prosed solutions for it [10]. Dosh tried to show effect of feature selection algorithms on CERT dataset by using random forest, Naïve bayes, and nearest neighbor algorithm and he shows that feature selection algorithm does not work properly on CERT dataset [11]. Shashanka et al. collected network traffic data using Niara platform from Nov 2015 to Jan 2016 and 1.315.895.522 raw data were recorded. They proposed a singular value decomposition model, and they analyzed the UEBA from different perspective [12]. Carlsson and Nabhani generated dataset using amount of traffic sent, the timing of sending packets, direction of the traffic and ports. They applied six different machine learning models on this dataset, and they showed that k-nearest neighbor and random forest were achieved highest accuracy [13].

Although UEBA has been studied in the literature by researchers, the application area has not reached the desired levels. Many events such as web site connections, temporary device usage, computer usage time, used applications or programs, email activities and file activities are needed to improve performance of insider attacker malicious activities prediction systems. However, logs of these activities are not shared publicly due to security and information privacy reasons. This lack of data affects the number of studies very much. The most common dataset for UEBA is The Computer Emergency and Response Team Dataset (CERT) that produced thanks to support of Carnegie Mellon University [14]. This dataset contains many raw data from device, email, file, http, and logon activities. In addition to this, Lightweight Directory Access Protocol (LDAP) information and result of OCEAN test are also included in these datasets. In the literature, most of UEBA have been done on this dataset, however many of them used different kind of feature extraction techniques. CERT dataset contains several files which contains raw event data. It takes a lot of effort to get this data ready for the feature extraction stage. None of the study in the literature has shared this process or dataset that is ready to extract feature. Because of this reason, CERT dataset preprocessed in this letter and these steps were explained. The python code of these process in shared in GitHub that is open-source web-based storage service for development projects [15]. In addition to this, future prospects of our team for UEBA and SIEM integrated anomaly detection systems were mentioned and the system we are considering designing and missing points of CERT datasets were argued.

2. The Computer Emergency and Response Team Dataset

CERT dataset is insider threat test datasets, which generated synthetically by The Computer Emergency and Response Team [16]. This dataset consists of daily activities from 1000 users for 500 days. Five different activity types, which are temporary device connection, incoming-outgoing email, file transfer, visited websites and computer usage, were tracked. Different files were generated for each activity types. In addition to this, monthly LDAP files from December 2009 to May 2011 with detailed personal information and result of OCEAN personality test [17] for each user were shared. CERT dataset has different version and version 4.2 was used in this letter, because it contains the largest number of malicious actions. Many actions in these datasets are standard actions and the Computer Emergency and Response Team determined three scenarios, which were listed below, to define malicious action. Therefore, they shared another three files which consist of the malicious actions.

- User who did not previously use removable drives or work after hours begins logging in after hours, using a removable drive, and uploading data to wikileaks.org. Leaves the organization shortly thereafter.
- User begins surfing job websites and soliciting employment from a competitor. Before leaving the company, they use a thumb drive (at markedly higher rates than their previous activity) to steal data.
- System administrator becomes disgruntled. Downloads a keylogger and uses a thumb drive to transfer it to his supervisor's machine. The next day, he uses the collected keylogs to log in as his supervisor and send out an alarming mass email, causing panic in the organization. He leaves the organization immediately.

Although version 4.2 of CERT dataset was used in this study, there were more recent version of it. This version was used, because it is more generalized version, and it is the version with highest number of samples belonging to malicious actions. Some of the other versions of CERT datasets contains five scenarios for malicious actions. However, these preprocessing steps can be applied on the other versions of CERT dataset with a few changes.

3. Preprocessing on Cert Dataset

Several feature extraction techniques such as daily, monthly and log-on-log-off based, were used on CERT dataset. It is seen that one of the best performances was achieved with log-on-log-off based. Thus, in this letter, session-based approach was applied during the feature engineering part. Generated action files in CERT dataset are action based and these actions are not separated with respect to user. For example, device file consists of temporary device connection for all users. However malicious action should be detected based on a user. Because of this reason, firstly an empty dictionary was created where each key of this dictionary represents employee of CERT dataset. Values of each key were consisting of employee's action where each action has four attributes: action id, action date, action personal computer and action type. Device, email, file, http and logon files were processed separately, and actions of user were added to dictionary. Action types are connected or disconnect for device, log-on or log-off for logon, file transfer for file, incoming or outgoing email for email and web browser activities for http files. After this phase, all user activities are sorted by date in themselves. A comma separated values (csv) file was generated for each user where a file consists of all action of user that are ordered by date (the first activity is shown first). As a result of these process a total of 32770222 activities were extracted and 1000 csv file was generated. Table 1 shows the number of events per activity types.

Table 1 Number of Events Per Activity Types

File Name	Activity Type	Number of Event
logon	Logon	470591
	Logoff	384268
device	Connect	203339
	Disconnect	202041
email	Incoming our Outgoing Email	2629979
file	File Transfer to Temporary Device	445581
http	Browser Activity	28434423

As can be seen in the table 1, the most frequent activity is browser activity, and the least frequent activity is device connect or disconnect. In CERT dataset, also there is answer files which contains the list of all malicious events. These files are separated according to the scenarios. 30, 30 and 12 employees committed malicious activity for scenario 1, 2 and 3 respectively. Total number of malicious events are 345 for scenario 1, 6765 for scenario 2 and 213 for scenario 3.

In this letter samples are generated with respect to sessions; thus, it is assumed that an activity sample is all events between a log on and a log off. Thus, for each user all events between a log on and log off were combined. As mentioned in earlier stage, malicious activities were based on event. In a sample, some events may be malicious while the others not. Thus, it is assumed that, if any of events during session was malicious, that session labeled as anomaly. As a result of this process, 384269 samples were generated where 69 of them were anomaly according to scenario 1, 693 of them were anomaly according to scenario 2, 33 of them were anomaly according to scenario 3 and 383474 of them were normal behavior. As can be seen from the number of samples, the CERT dataset has imbalance problem. This problem can be mitigated by combing all scenarios in one class which is anomaly. However, in anomaly detection systems scenarios were also important, because some scenarios may be more dangerous than the others. As can be seen from the literature review conducted throughout the letter, imbalance problem is not only problem for CERT dataset, but also it is one of the main problems in anomaly detection systems. Thus, the malicious actions were not combined in one class. However, in the future they can be combined easily. In this way, the effect of imbalance problem on the performance of anomaly detection systems also be analyzed using our preprocessed dataset.

Two types of values, which were numerical and action sequence based, were generated for each sample. In action sequence-based values all action during the session were collected in chronological order. In some samples, session duration was too low (for example signed out immediately after signing in) and some sessions had only logon or logoff (Action may not be caught due to some processes such as SIEM connection error). These types of data were not eliminated because it is thought that they can be also used in further process while feature extraction. For example, the combination of sessions also be a malicious event. Some time series model using regression or long short-term memories algorithm can be proposed to make anomaly detection system more complicated.

Numerical values were computed according to the information package of dataset and literature reviews [3], [9], [14], [16]. Firstly, actions are divided into two categories that are during the working hours and during the work of hours. For this purpose, it is assumed that working hours were between the 8:00 a.m. and 7:00 p.m. and also it is assumed that only weekdays are working hours. Using these information, session duration in work on, session duration in work off, number of email during the work on, number of email during the work off, number of file activity during the work on, number of file activity during the work off, number of browser activity during the work on, number of browser activity during the work off, number of temporary device activity during the work on and number of temporary device activity during the work off were computed for each sample. At the end of the this, process action sequence based, and numerical values are concatenated and team information, role information, openness value, conscientiousness value, extraversion value, agreeableness value and neuroticism value for employee were added to each sample. The python code of this preprocess steps was shared in GitHub platform [15]. Figure 1 summarizes the step of preprocessing. The python codes of the systems and the functions were developed with respect to these steps.

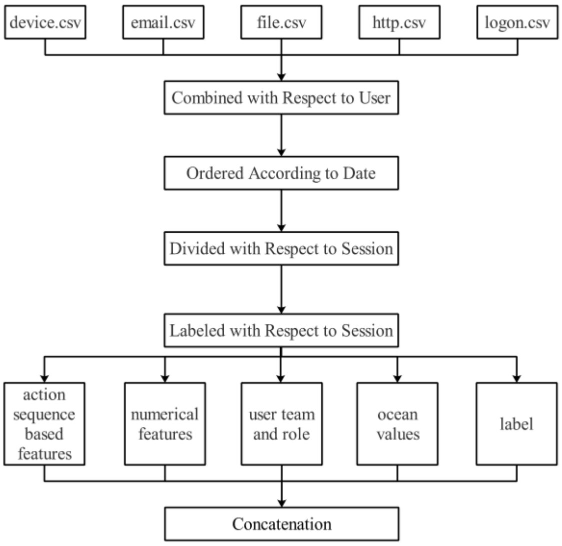

device.csv
email.csv
file.csv
http.csv
logon.csv
Combined with Respect to User
Ordered According to Date
Divided with Respect to Session
Labeled with Respect to Session
action sequence based features
numerical features
user team and role
ocean values
label
Concatenation

Figure 1 Summary of Preprocessing Steps.

4. Conclusions and Future Prospects

In this letter, importance of SIEM systems and UEBA were mentioned and some studies about anomaly detection on SIEM systems and effect of UEBA to detect insider malicious action were reviewed. Because of the security reason, data collected from SIEM systems are not shared publicly, thus designing an insider attack detection system is hard to research. In the literature, the most common used dataset is CERT [14] for UEBA. There are many studies that developed a machine learning algorithm on CERT dataset, however they did not share their preprocessing steps publicly. In this letter, CERT log files were preprocessed and a file, which is ready to extract feature, generated. It is also possible to use this file without feature extraction to design UEBA system.

In the future work, firstly several machine learning algorithms will be proposed using preprocessed dataset. For this context, multi class and binary class prediction systems will be developed separately. In multi class prediction systems each malicious actions that are generated using three scenarios will be predicted separately. In binary class prediction systems, actions will be labeled as malicious or not malicious. For this purpose, malicious actions generated using three scenarios will be assigned to insider attack. In addition to this, several feature extraction techniques will be applied to action sequence-based values to extract different kind of features. The effect of features will be analyzed thanks to this stage.

Deep learning techniques are more effective than the traditional machine learning algorithms in many problems including UEBA, especially if the number samples are large [9], [19]– [23]. Thus, instead of using only traditional methods, deep learning approaches will also be proposed. Different kind of layers such as convolutional neural networks, long short-term memories and graph convolutional networks will be used to generate deep models. In this model user, role or department-based approaches will be used. Samples will be grouped with respect to approaches and separate input layers will be generated for each of them. At the final phase all layers will be concatenated for classification layer.

In the final phase, we are planning to develop anomaly detection systems that is integrated to SIEM platforms. One of the most important modules of this system will be detection of insider attackers. As mentioned before, CERT dataset is the most common used dataset to detect insider attackers. It is very comprehensive dataset. Because it generated synthetically, it has some shortcomings. The first shortcoming is number of scenarios are not enough to cover some malicious event for large companies. Anomaly actions in version 4.2 of CERT dataset were produced using three different scenarios, however in real world applications there are many several scenarios. Thus, we are planning to collect new data that consist of more real-world scenarios. For this context, we will collect data from our employees, which is more than 700, using our SIEM platform. The second shortcoming is CERT dataset does not consist of some information from real world applications. The most important one is it does not collect connection information such as remote desktop or secure shell (SSH). In our example, these deficiencies will be identified, and data collected through extensive research. The other shortcoming is because CERT dataset generated synthetically some information in features are not real. For example, many connect of websites in the http request files are generated randomly (They are not real website). However, this information is crucial for anomaly detection, and it is vital to have real data. In addition to this, visited web site depends on location. For these reasons, frequently used websites will be identified and categorized for Türkiye.

It is thought that it can be done in different studies from the malicious action detection problem using these data. The first important analysis is identifying employees who will be fired using the visited web site information. The other important analysis is revealing work performance using visited website and log-on and log-off information.

References

[1] P. Slipenchuk and A. Epishkina, “Practical User and Entity Behavior Analytics Methods for Fraud Detection Systems in Online Banking: A Survey,” in Biologically Inspired Cognitive Architectures 2019, Cham, 2020, pp. 83–93. doi: 10.1007/978-3-030-25719-4_11.

[2] E. T. Anumol, “Use of Machine Learning Algorithms with SIEM for Attack Prediction,” in Intelligent Computing, Communication and Devices, New Delhi, 2015, pp. 231–235. doi: 10.1007/978-81-322-2012-1_24.

[3] T. Laue, T. Klecker, C. Kleiner, and K.-O. Detken, “A SIEM Architecture for Advanced Anomaly Detection,” vol. 6, no. 1, p. 17, 2022.

[4] S. Asanger and A. Hutchison, “Experiences and Challenges in Enhancing Security Information and Event Management Capability Using Unsupervised Anomaly Detection,” in 2013 International Conference on Availability, Reliability and Security, Sep. 2013, pp. 654–661. doi: 10.1109/ARES.2013.86.

[5] M. Goldstein, S. Asanger, M. Reif, and A. Hutchison, “Enhancing Security Event Management Systems with Unsupervised Anomaly Detection:,” in Proceedings of the 2nd International Conference on Pattern Recognition Applications and Methods, Barcelona, Spain, 2013, pp. 530–538. doi: 10.5220/0004230105300538.

[6] A. Lukashin, M. Popov, A. Bolshakov, and Y. Nikolashin, “Scalable Data Processing Approach and Anomaly Detection Method for User and Entity Behavior Analytics Platform,” in Intelligent Distributed Computing XIII, Cham, 2020, pp. 344–349. doi: 10.1007/978-3-030-32258-8_40.

[7] Z. Tian, C. Luo, H. Lu, S. Su, Y. Sun, and M. Zhang, “User and Entity Behavior Analysis under Urban Big Data,” ACMIMS Trans. Data Sci., vol. 1, no. 3, p. 16:1-16:19, Sep. 2020, doi: 10.1145/3374749.

[8] D. C. Le and A. N. Zincir-Heywood, “Evaluating Insider Threat Detection Workflow Using Supervised and Unsupervised Learning,” in 2018 IEEE Security and Privacy Workshops (SPW), May 2018, pp. 270–275. doi: 10.1109/SPW.2018.00043.

[9] B. Sharma, P. Pokharel, and B. Joshi, “User Behavior Analytics for Anomaly Detection Using LSTM Autoencoder - Insider Threat Detection,” in Proceedings of the 11th International Conference on Advances in Information Technology, New York, NY, USA, Jul. 2020, pp. 1–9. doi: 10.1145/3406601.3406610.

[10] T. Al-Shehari and R. A. Alsowail, “An Insider Data Leakage Detection Using One-Hot Encoding, Synthetic

Minority Oversampling and Machine Learning Techniques," Entropy, vol. 23, no. 10, Art. no. 10, Oct. 2021, doi: 10.3390/e23101258.

[11] M. Dosh, "Detecting insider threat within institutions using CERT dataset and different ML techniques," Period. Eng. Nat. Sci. PEN, vol. 9, no. 2, Art. no. 2, May 2021, doi: 10.21533/pen.v9i2.1911.

[12] M. Shashanka, M.-Y. Shen, and J. Wang, "User and entity behavior analytics for enterprise security," in 2016 IEEE International Conference on Big Data (Big Data), Dec. 2016, pp. 1867–1874. doi: 10.1109/BigData.2016.7840805.

[13] O. Carlsson and D. Nabhani, "User and Entity Behavior Anomaly Detection using Network Traffic," p. 52.

[14] "Insider Threat Test Dataset." Carnegie Mellon University, Sep. 30, 2020. doi: 10.1184/R1/12841247.v1.

[15] "Arge-Preprocessing-CERT." Detaysoft, Oct. 23, 2022. Accessed: Oct. 23, 2022. [Online]. Available: https://github.com/Detaysoft/Arge-Preprocessing-CERT

[16] W. R. Claycomb and A. Nicoll, "Insider Threats to Cloud Computing: Directions for New Research Challenges," in 2012 IEEE 36th Annual Computer Software and Applications Conference, Jul. 2012, pp. 387–394. doi: 10.1109/COMPSAC.2012.113.

[17] "Big Five personality traits," Wikipedia. Oct. 07, 2022. Accessed: Oct. 20, 2022. [Online]. Available: https://en.wikipedia.org/w/index.php?title=Big_Five_personality_traits&oldid=1114671408

[18] J. Glasser and B. Lindauer, "Bridging the Gap: A Pragmatic Approach to Generating Insider Threat Data," in 2013 IEEE Security and Privacy Workshops, May 2013, pp. 98–104. doi: 10.1109/SPW.2013.37.

[19] R. Abdulhammed, M. Faezipour, A. Abuzneid, and A. AbuMallouh, "Deep and Machine Learning Approaches for Anomaly-Based Intrusion Detection of Imbalanced Network Traffic," IEEE Sens. Lett., vol. 3, no. 1, pp. 1–4, Jan. 2019, doi: 10.1109/LSENS.2018.2879990.

[20] A. Hassan and A. Mahmood, "Deep Learning approach for sentiment analysis of short texts," in 2017 3rd International Conference on Control, Automation and Robotics (ICCAR), Apr. 2017, pp. 705–710. doi: 10.1109/ICCAR.2017.7942788.

[21] Y. Görmez, M. Sabzekar, and Z. Aydın, "IGPRED: Combination of convolutional neural and graph convolutional networks for protein secondary structure prediction," Proteins Struct. Funct. Bioinforma., vol. 89, no. 10, pp. 1277–1288, 2021, doi: 10.1002/prot.26149.

[22] X. Hou, T. Arslan, A. Juri, and F. Wang, "Indoor Localization for Bluetooth Low Energy Devices Using Weighted Off-set Triangulation Algorithm," presented at the Proceedings of the 29th International Technical Meeting of the Satellite Division of The Institute of Navigation (ION GNSS+ 2016), Sep. 2016, pp. 2286–2292. doi: 10.33012/2016.14720.

[23] Z. Wang, S. Sugaya, and D. P. T. Nguyen, "Salary Prediction using Bidirectional-GRU-CNN Model," p. 4, 2019.

Journal of Soft Computing and Artificial Intelligence

Journal homepage: https://dergipark.org.tr/en/pub/jscai

International Open Access

Volume 04
Issue 01

June, 2023

Research Article

Comparative Analysis of Globalisation Techniques for Medical Document Classification

***Bekir Parlak*[1], *Salih Berkan Aydemir*[1]**

[1]*Department of Computer Engineering, Amasya University, 0500, Amasya, Turkey*

ARTICLE INFO

Article history:
Received **December 9, 2022**
Revised **February 13, 2023**
Accepted **February 17, 2023**

Keywords:
Medical documents
Text Classification
Feature selection
Globalisation techniques

ABSTRACT

Medical document classification is one of the important topics of text mining. Globalisation techniques play a major role in feature selection stage. Therefore, globalization techniques affect text classification performance. Our aim in the study is to conduct a detailed analysis on two data sets with English and Turkish content by using medical text summaries of Turkish articles. These datasets consist of Turkish and English text summaries of the same articles. To observe how successful local feature selection methods in the field of text classification affect the classification performance on these two equivalent data sets by applying different globalisation techniques. The feature selection methods used are CHI2(chi-square), MI (mutual information), OR (odds ratio), WLLR (weighted log-likelihood ratio). Globalisation techniques are SUM (summation), AVG (average), MAX (maximum). Classifiers are MNB (multinomial naive bayes), DT (decision tree), and SVM (support vector machine). For the English Ohsumed data set, the highest Micro F score value of 95.48 was obtained in the max globalization method with the 2000-dimension CHI2 feature selection method and MNB classifier method. For the Turkish Ohsumed data set, the highest Micro F score value of 92.75 was obtained in the max globalization method with the 2000-dimension CHI2 feature selection method and MNB classifier method. In comparisons, it has been observed that the best classifier for Ohsumed datasets is MNB.

1. Introduction

With the rapid development of internet technologies in recent years, it can be seen that there is a huge increase in the number of electronic documents. The fact that the internet is more accessible to people and the increase in personal computers are among the reasons for this increase. Text classification methods play an important role in many documents on the internet. It can be used in solving various problems such as text classification [1], spam filtering [2], author identification [3], classification of web pages [4], classification of medical texts [5]. The importance of text classification increases the importance of databases where text classification is used. Documents in the database named MEDLINE are generally used for information access to medical texts and text classification studies. MEDLINE is a bibliographic database containing over 21 million documents from approximately 5.600 medical journals. This database can be queried with certain parameters over the internet, thanks to a search platform called PubMed [6]. In Turkey, there is TUBITAK's Medical Database created to facilitate access to information for experts working in the field of medicine. MEDLINE and ULAKBIM Medical Database are indexed by taking the relevant MeSH (Medical Subject Headings) terms with category information and selecting them manually by experts. Although an automated system is not used for indexing the

[1] Corresponding author
e-mail: salih.aydemir@amasya.edu.tr
DOI: 10.55195/jscai.1216800

MEDLINE database, there are automatic text classification studies on MEDLINE data in the literature. On the other hand, more useful and concise data are obtained by applying various methods on medical document data. These methods are featuring weighting, classification, feature selection, pre-processing and feature extraction. The text properties corresponding to a large number of documents are also quite high. Therefore, the dimensionality disaster will be greatly affected if size reduction is not made in the face of high dimensional text features. Feature extraction and feature weighting are the two main methods of reducing feature dimensionality. Feature weighting is a text classification phase that calculates the feature weight for each feature of documents. Feature extraction is a size reduction process in which an initial dataset is reduced to more manageable groups for processing.

In text categorization (TC), feature selection can be applied after feature extraction. Considering local feature selection methods, a globalisation policy is required to transform multiple local scores into a unique global score [7]. Globalisation techniques play an important role in TC. Considering the local scores, the global score can be calculated using various globalisation techniques. On the other hand, pre-processing is an important step for TC. Here are some pre-processing methods for text classification: lowercase conversion, removal of stop words, stemming and tokenization [8].

The motivation of this study is to choose the ideal feature selection, classifier and globalization techniques for Turkish and English Ohsumed datasets. All experiments were repeated in different dimensions, Macro F1 and Micro F1 values were calculated, and the results were reported.

Other parts of the work are organized as follows. In the second part, a detailed study area is examined. The basic methods used in the study are explained in the third part. In fourth part, experimental studies are given. In the last part, the conclusion part and future studies are mentioned.

2. Literature Review

Until now, various feature selection methods and classifiers have been applied on TC. In this part, feature selection methods used in TC are included.

Zheng et al. used information gain (IG), chi-square (CHI), correlation coefficient (CC) and odds ratios (OR) feature selection methods on imbalance data. Authors discussed feature selection methods in both one-sided (CC, OR) and two-sided (IG, CHI) metrics [9]. SVM produces effective results for TC. Taire and Haruno have investigated the effect of prior feature selection for Support Vector Machine (SVM) [10]. There are new feature selection methods proposed for TC in the literature. Gunal has proposed a novel hybrid feature selection which combine filter and wrapper methods for text classification [11]. In another study, Biricik et al. proposed a supervised feature extraction algorithm by combining the effect of input properties on classes. Their method is called abstract feature extraction [12].

Conventional TC algorithms consist of three main parts as handcrafted, nature-inspired and graph-based [13]. In the field of TC, many optimization-based feature selection methods have been proposed. The sine-cosine optimization algorithm, which has been proposed inspired by the sin and cos curves, has been developed and it is used as the feature selection method [14]. Feature selection was proposed using PSO. In addition, radial basis function neural networks are used as classifiers [15]. On the other hand, TC is applied using handcrafted features [13]. Some scholars have used traditional classifiers for the creation of feature sets and classification purposes, and they have proposed graph based feature selection methods [16].

Although feature selection and classification algorithms play an active role in a TC problem, globalisation techniques have strong effects on TCs. Some of the feature weighting methods in the literature generate a single global weighting score for each feature. However, local-based methods produce a different score for each class. There are some ways to get global scores from local scores: maximization, average, weighted average, and weighted maximum are popular globalisation techniques [17]. Parlak and Uysal have performed the impact of globalisation techniques on feature selection methods in TC [5]. In their studies, they used two successful classifiers, while they used four benchmark data sets. For Turkish Ohsumed dataset, the highest Micro-F1 and Macro-F1 scores are 92.75 and 82.82, respectively. It was obtained with the combination of CHI2 method, MAX globalisation technique, and MNB classifier using 2000 feature size. SVM classifier is the successor classifier for most cases. Also, CHI2 method is more successful than the other feature selection method in most cases for this data set. MI is the worst feature selection method for all situations.

3. Preliminaries

In this section, the basic techniques used in the study are mentioned.

3.1 Classifiers

The aim of text classification studies is to classifying uncategorized documents into predefined classes. In our experiments, three successful classifiers were employed to evaluate selected features by different globalisation techniques for each dataset. These classifiers are Multinational Naive Bayes (MNB), Decision Tree (DT), and Support Vector Machines (SVM). MNB is a form of naive Bayes classifier and very successful classifiers in text classification domain [37]. As classic Naive Bayes models a document with the occurrence and not occurrence of certain features, MNB clearly models it using feature counts. Multinational and multi-variate Bernoulli event models are widely utilized for text classification studies. While MNB takes into account term frequencies, multi-variate Bernoulli event model employs document frequencies. DT is one of the most efficient classifiers in text classification domain [38]. DT is a nonlinear classifier where classes are not accepted until a logical class is detected. SVM is one of the best classifiers in text classification studies. It has two versions which are linear and non-linear. In the experiments, we employed linear version of SVM classifier. The main subject of SVM classifier is the margin. LibSVM library is used for SVM classifier with linear kernel [39].

3.2. Feature Selection Methods

In our experiments, we employed four local feature selection algorithms. These are Chi-Square (CHI2), Mutual Information (MI), Odds Ratio (OR), and (WLLR).

CHI2: CHI2 is a successful feature selection method in text classification domain. The CHI2 method calculates the lack of independence between feature t and class C [17]. A and B events are assumed to be independent if

$$p(XY) = p(X)p(Y) \tag{1}$$

CHI2 method can be calculated as below:

$$CHI2(t_i, c_i) = \frac{N*(TP*TN-FP*FN)^2}{(TP+FN)*(TP+FP)*(FN+TN)*(FP+TN)} \tag{2}$$

MI: MI is a local method which computes the correlation between classes and features [18]. MI is computed as below:

$$MI(t_i, c_j) = log\frac{P(t_i|c_j)}{P(t_i)} \tag{3}$$

OR: OR is a supervised and local feature selection method which calculates the membership and non-membership to each class by utilizing nominator and denominator in Equation 4, respectively [19]. So, the OR method can produce both the negative and the positive scores. The method is computed as:

$$\boldsymbol{OR(t_i, c_j) = log\frac{P(t_i|c_j)(1-P(t_i|\bar{c}_j))}{(1-P(t_i|c_j))P(t_i|\bar{c}_j)}} \tag{4}$$

WLLR: WLLR is a supervised and local feature selection method which is proposed by Nigam et al. [20]. The WLLR method is calculated as below:

$$WLLR(t_i, c_j) = P(t_i|c_j)log\frac{P(t_i|c_j)}{P(t_i|\bar{c}_j)} \tag{5}$$

3.3. Globalisation Techniques

In our experiments, MAX, SUM, AVG globalisation techniques were utilized [5]. The reason we use these methods is to examine in detail how the same feature selection methods affect performance with different globalization techniques. These methods are generally used in the literature.

All of the scores are summed in SUM technique. The scores computed on each class are globalized by multiplying class probabilities in AVG technique. In MAX technique, the maximum of all scores is taken. Here, $f(t_i, C_j)$ corresponds to the score of the feature t_i in class C_j. These globalization techniques can be calculated as below:

$$SUM = \sum_{j=1}^{M} f(t_i, c_j) \tag{6}$$

$$AVG = \sum_{j=1}^{M} P(C_j) * f(t_i, c_j) \tag{7}$$

$$MAX = \max_{j=1}^{M} f(t_i, c_j) \tag{8}$$

4. Experimental Study

In our experiments, we used two data sets. Micro-F1 and Macro-F1 scores were utilized to analysing classification performance. Ten largest classes were included in the experimental works. The characteristics of the data sets used in the article are given in Table 1 and Table 2. Within the scope of the study, experiments were carried out using java programming and WEKA tool. The flow chart of the analyzes made in Figure 1 has been added.

10-fold cross-validation was employed for fair evaluation. Different number of features which were selected by each feature selection method were fed into MNB, SVM and DT classifiers. 100, 250, 500, 1000 and 2000 dimension was used as a feature size. Also, the total number of features are 8610, 14334 for English and Turkish data sets, respectively. Resulting Micro-F1 and Macro-F1 scores are showed in Tables 3-8. For English Ohsumed dataset, the highest Micro-F1 and Macro-F1 scores are 95.48 and 88.25, respectively.

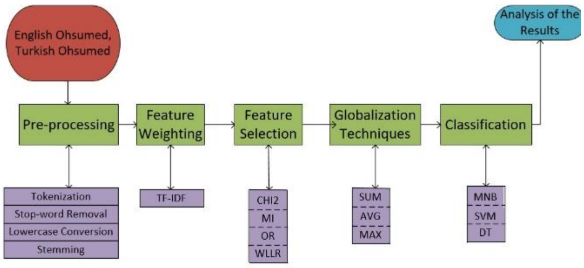


Figure 1 Flowchart of applied analyzes.

Table 1 Ohsumed Dataset for English

Class Number	**Disease Category**	***Number of Documents***
1	Bacterial Infections and Mycoses	631
2	Virus Diseases	249
3	Parasitic Diseases	183
4	Neoplasms	2513
5	Musculoskeletal Diseases	505
7	Stomatognathic Diseases	132
8	Respiratory Tract Diseases	634
10	Nervous System Diseases	1328
14	Female Genital Diseases and Pregnancy Complications	2876
23	Pathological Conditions, Signs and Symptoms	1924

Table 2 Ohsumed Dataset for Turkish

Class Number	**Disease Category**	***Number of Documents***
1	Bacterial Infections and Mycoses	284
2	Virus Diseases	44
3	Parasitic Diseases	116
4	Neoplasms	32
5	Musculoskeletal Diseases	140
7	Stomatognathic Diseases	39
8	Respiratory Tract Diseases	90
10	Nervous System Diseases	83
14	Female Genital Diseases and Pregnancy Complications	231
23	Pathological Conditions, Signs and Symptoms	73

The best score was obtained from the combination of CHI2 method, MAX method and MNB classifier using 2000 feature size. DT classifier is the second successful classifier. Also, CHI2 method is more successful than the other feature selection method in most cases for this dataset. MI is the worst feature selection method for all situations. Also, MAX globalisation is more efficient method than the other globalisation according to each feature selection method for most cases. For english dataset, in Table 3, in the classifications made with MNB, the best values were obtained with CHI2. In Table 4, in the classifications made with SVM, the best values were obtained with CHI2 in AVG. In Table 5, in the classifications made with DT, the score values were obtained with CHI. For turkish dataset, In Table 6, in the classifications made with MNB, the score values were obtained with CHI2 in MAX. In Table 7, in the classifications made with SVM, the best values were obtained with WLLR in AVG. In Table 8, in the classifications made with DT, the best values were obtained with OR in MAX.

Generally speaking, the performance increases as the number of dimesions increases in the datasets. While the highest scores in the English data set are obtained with the CHI2 method, the highest scores can be obtained with different methods in the Turkish data set.

5. Conclusion and Future Works

In this paper, we have comprehensively analysed two datasets consisting of Turkish and English abstracts extracted from Turkish medical journals. A comprehensive study on classification of two counterparts abstracts was showed by using three classifiers. Three different globalisation techniques and four local feature selection methods were used in performance analysis. Also, three pattern classifiers were used in classification stage. According to experimental studies, classification of English dataset containing medical abstracts is more successful than their counterparts in Turkish dataset. MNB classifier has gained more performance than SVM and DT classifiers on both data sets. The focal point of MNB classifier is the supposition of independence between terms. Also, MAX globalisation technique is the best method according to classification performance for most cases in both datasets. While CHI2 method is more successful than other methods, MI method is the worst method for most cases in both datasets. As a future work, a novel globalisation technique may be developed for medical domain. Also, the effect of different feature representation methods may be investigated in both languages. The analyzes made in the study were applied for two different data sets. However, more feature extraction methods and classification can be applied to different data. The article can be extended with the latest globalization techniques.

Table 3 Micro and Micro -F scores (%) obtained on English dataset with MNB

Micro-F	CHI2			MI			OR			WLLR		
Dimension	MAX	SUM	AVG	MAX	SUM	AVG	MAX	SUM	AVG	MAX	SUM	AVG
100	87.70	86.46	83.42	40.23	47.13	47.95	82.45	82.51	80.15	80.65	73.02	74.01
250	92.40	90.26	88.85	42.23	50.46	58.06	87.47	86.68	84.96	86.17	81.65	81.47
500	93.01	92.75	91.10	45.36	58.23	68.60	92.04	87.98	88.69	89.61	87.53	86.35
1000	94.89	94.70	93.81	48.46	63.77	77.56	94.40	89.94	90.15	91.88	90.26	89.67
2000	**95.48**	94.99	94.55	61.87	72.09	**85.83**	**95.14**	92.14	90.52	**93.16**	92.34	91.57
Macro-F	CHI2			MI			OR			WLLR		
Dimension	MAX	SUM	AVG	MAX	SUM	AVG	MAX	SUM	AVG	MAX	SUM	AVG
100	74.67	72.37	57.64	04.62	16.72	18.89	68.87	64.74	54.56	64.84	49.88	50.92
250	82.26	78.52	73.79	11.80	20.82	32.07	76.59	71.85	62.44	71.91	64.55	61.64
500	83.34	82.90	78.47	20.35	31.86	44.11	81.53	74.17	71.78	77.55	74.12	69.90
1000	87.71	87.20	84.04	26.88	40.86	54.56	86.73	77.08	76.38	81.45	78.43	76.70
2000	**88.25**	87.63	86.47	48.21	53.37	**66.68**	**87.42**	80.93	78.13	**84.15**	82.09	80.57

Table 4 Micro and Micro-F scores (%) obtained on English dataset with SVM

Micro-F	CHI2			MI			OR			WLLR		
Dimension	MAX	SUM	AVG	MAX	SUM	AVG	MAX	SUM	AVG	MAX	SUM	AVG
100	87.08	84.43	82.51	40.45	48.05	47.34	80.90	80.21	79.06	77.49	70.18	71.73
250	87.98	86.91	85.25	43.32	52.12	57.97	83.72	84.14	81.90	83.12	78.74	77.49
500	88.20	88.85	88.75	45.46	55.89	62.62	86.86	86.35	86.06	88.97	87.36	86.74
1000	91.73	92.19	91.62	48.15	58.41	68.68	87.59	88.14	88.20	91.52	91.10	91.52
2000	92.70	93.11	**93.26**	59.47	66.04	**79.45**	89.35	**91.21**	90.84	92.40	92.45	**92.60**
Macro-F	CHI2			MI			OR			WLLR		
Dimension	MAX	SUM	AVG	MAX	SUM	AVG	MAX	SUM	AVG	MAX	SUM	AVG
100	72.34	67.58	58.05	05.73	16.87	15.09	68.68	60.72	51.00	58.81	43.97	46.50
250	74.20	73.40	67.54	15.05	24.52	28.43	69.01	68.71	58.83	67.13	60.09	54.41
500	75.74	77.33	73.70	21.24	28.39	37.44	74.14	71.71	66.63	76.75	74.33	69.31
1000	80.87	81.93	80.47	29.06	34.16	42.38	72.41	74.30	73.65	80.55	79.41	79.57
2000	82.52	82.90	**84.64**	46.81	47.03	**56.94**	77.31	**79.73**	78.34	**82.35**	82.17	82.05

Table 5 Micro and Macro-F scores (%) obtained on English dataset with DT

Micro-F	CHI2			MI			OR			WLLR		
Dimension	MAX	SUM	AVG	MAX	SUM	AVG	MAX	SUM	AVG	MAX	SUM	AVG
100	**86.46**	84.02	81.90	40.11	45.46	48.87	79.25	80.21	77.02	76.76	65.88	69.21
250	85.02	83.60	85.13	41.23	45.78	55.88	82.39	80.84	81.47	82.02	77.69	73.80
500	85.31	85.71	85.13	41.57	47.75	62.29	85.71	81.90	81.40	83.78	82.57	80.91
1000	85.13	85.88	86.06	49.27	49.67	68.83	**85.77**	82.39	82.63	84.49	83.90	84.02
2000	85.77	85.60	85.37	57.34	59.20	**77.09**	85.19	85.19	84.31	**84.78**	84.49	84.61
Macro-F	CHI2			MI			OR			WLLR		
Dimension	MAX	SUM	AVG	MAX	SUM	AVG	MAX	SUM	AVG	MAX	SUM	AVG
100	**71.79**	67.91	57.78	04.01	11.91	15.55	66.87	62.84	49.02	59.43	39.49	42.08

250	68.40	66.20	68.00	08.12	12.43	23.44	67.34	62.46	56.26	64.74	58.50	45.93
500	69.93	70.13	68.87	09.23	20.08	33.77	**72.92**	64.91	57.48	67.06	64.66	60.19
1000	69.43	71.16	71.57	31.35	24.14	41.60	71.49	65.28	64.25	68.46	67.32	67.45
2000	70.88	70.78	70.35	44.88	34.17	**52.22**	70.26	69.45	67.07	**68.81**	67.87	68.41

Table 6 Micro and Macro-F scores (%) obtained on Turkish dataset with MNB

Micro-F	CHI2			MI			OR			WLLR		
Dimension	MAX	SUM	AVG	MAX	SUM	AVG	MAX	SUM	AVG	MAX	SUM	AVG
100	83.66	80.72	74.78	41.12	40.56	40.45	73.59	72.66	68.68	73.52	60.07	59.73
250	87.64	86.57	84.25	42.23	49.16	46.72	83.42	79.06	76.42	80.97	73.45	73.09
500	89.61	89.45	86.06	44.40	54.87	54.78	87.98	82.94	82.57	85.89	82.45	81.84
1000	91.41	91.67	89.99	48.26	62.62	54.22	90.09	87.36	85.89	89.29	88.14	86.57
2000	**92.75**	92.34	91.67	56.52	**79.25**	55.61	**91.42**	89.94	86.63	**92.08**	90.68	90.94
Macro-F	CHI2			MI			OR			WLLR		
Dimension	MAX	SUM	AVG	MAX	SUM	AVG	MAX	SUM	AVG	MAX	SUM	AVG
100	66.25	58.89	42.68	08.11	10.26	08.60	58.79	45.20	38.05	47.57	29.19	29.34
250	74.29	71.89	61.56	10.71	22.74	15.069	71.37	57.55	46.23	63.83	49.14	44.87
500	77.23	76.88	64.90	17.24	26.88	21.84	76.07	65.55	58.91	70.38	65.65	60.59
1000	80.97	81.23	74.61	24.88	35.79	22.14	77.43	71.17	62.42	76.49	74.92	70.30
2000	**82.82**	82.38	78.87	38.24	**55.80**	21.45	**79.87**	76.71	64.73	**80.70**	77.98	78.28

Table 7 Micro-F scores (%) obtained on Turkish dataset with SVM

Micro-F	CHI2			MI			OR			WLLR		
Dimension	MAX	SUM	AVG	MAX	SUM	AVG	MAX	SUM	AVG	MAX	SUM	AVG
100	81.84	79.25	73.24	40.90	41.12	46.40	74.29	72.02	69.74	71.58	59.12	60.68
250	83.54	83.30	80.47	42.01	49.47	52.51	80.15	74.22	76.76	74.01	66.98	67.21
500	83.30	82.76	81.40	43.97	51.74	55.15	80.84	77.42	78.02	79.19	76.28	75.81
1000	83.36	84.14	85.08	45.25	60.33	56.98	83.48	80.78	80.47	84.25	82.57	83.00
2000	84.78	**86.23**	85.83	52.60	**71.22**	29.33	81.03	**84.67**	83.42	86.69	87.03	**87.31**
Macro-F	CHI2			MI			OR			WLLR		
Dimension	MAX	SUM	AVG	MAX	SUM	AVG	MAX	SUM	AVG	MAX	SUM	AVG
100	62.91	57.94	42.99	08.26	10.27	09.30	54.75	46.29	37.40	45.18	27.94	30.00
250	64.83	65.33	58.20	11.36	20.12	18.07	62.02	49.55	51.62	54.20	42.60	42.00
500	65.30	64.30	59.19	16.39	23.27	24.34	62.58	56.64	52.08	60.90	57.23	53.52
1000	65.25	67.53	66.32	25.06	32.31	28.34	64.03	61.10	56.40	68.09	65.39	64.41
2000	66.41	68.18	**68.76**	35.26	46.37	**59.55**	60.37	**68.31**	62.41	71.10	71.52	**72.11**

Table 8 Micro-F scores (%) obtained on Turkish dataset with DT

Micro-F	CHI2			MI			OR			WLLR		
Dimension	MAX	SUM	AVG	MAX	SUM	AVG	MAX	SUM	AVG	MAX	SUM	AVG
100	79.64	76.76	73.59	39.89	40.56	44.51	70.18	68.83	69.51	66.67	52.80	51.15
250	**83.00**	81.34	78.61	40.11	45.46	52.60	77.82	73.59	75.95	76.08	63.29	64.59
500	81.40	79.96	80.08	42.23	46.51	53.75	80.59	74.22	76.15	77.89	74.50	72.23
1000	79.96	81.59	81.40	44.83	55.61	57.07	**83.66**	78.41	75.67	79.96	78.15	79.06
2000	82.27	81.09	81.53	47.34	**68.37**	59.73	81.96	79.57	77.02	**80.91**	79.56	79.96
Macro-F	CHI2			MI			OR			WLLR		
Dimension	MAX	SUM	AVG	MAX	SUM	AVG	MAX	SUM	AVG	MAX	SUM	AVG
100	59.16	54.14	44.47	04.26	08.22	07.13	48.02	44.27	38.21	40.68	25.10	25.13
250	**64.43**	60.76	53.38	05.37	14.85	16.96	59.73	49.09	49.50	55.67	37.28	36.19
500	62.20	59.30	58.12	11.26	17.17	20.79	62.52	50.59	50.52	55.14	52.20	46.45
1000	59.87	61.24	60.70	18.23	25.79	22.81	**64.75**	57.06	51.41	60.03	57.37	58.16
2000	62.24	60.59	61.21	28.93	**39.82**	24.26	61.33	57.51	53.09	**60.92**	60.23	59.87

References

[1] C. C. Aggarwal and C. Zhai, Mining text data. Springer Science & Business Media, 2012.

[2] G. V. Cormack, "Email spam filtering: A systematic review," 2008.

[3] S. J. Delany, M. Buckley, and D. Greene, "SMS spam filtering: Methods and data," Expert Systems with Applications, vol. 39, no. 10, pp. 9899-9908, 2012.

[4] M. Coulthard, "Author identification, idiolect, and linguistic uniqueness," Applied linguistics, vol. 25, no. 4, pp. 431-447, 2004.

[5] D. Madigan, A. Genkin, D. D. Lewis, S. Argamon, D. Fradkin, and L. Ye, "Author identification on the large scale," in Proceedings of the 2005 Meeting of the Classification Society of North America (CSNA), 2005.

[6] S. M. Zu Eissen and B. Stein, "Genre classification of web pages," in Annual Conference on Artificial Intelligence, 2004, pp. 256-269: Springer.

[7] B. Choi and X. Peng, "Dynamic and hierarchical classification of web pages," Online Information Review, 2004.

[8] N. Isaacson, "The "fetus-infant": Changing classifications of In Utero development in medical texts," in Sociological Forum, 1996, vol. 11, no. 3, pp. 457-480: Springer.

[9] B. Parlak and A. K. Uysal, "On classification of abstracts obtained from medical journals," Journal of Information Science, vol. 46, no. 5, pp. 648-663, 2020.

[10] N. L. Medicine. (2020, 2020-09-30). Available: https://pubmed.ncbi.nlm.nih.gov/

[11] C. A. Gonçalves, C. T. Gonçalves, R. Camacho, and E. C. Oliveira, "The Impact of Pre-processing on the Classification of MEDLINE Documents," in PRIS, 2010, pp. 53-61.

[12] M. Yetisgen-Yildiz and W. Pratt, "The effect of feature representation on MEDLINE document classification," in AMIA annual symposium proceedings, 2005, vol. 2005, p. 849: American Medical Informatics Association.

[13] A. Sarker and G. Gonzalez, "Portable automatic text classification for adverse drug reaction detection via multi-corpus training," Journal of biomedical informatics, vol. 53, pp. 196-207, 2015.

[14] J. G. Adeva, J. P. Atxa, M. U. Carrillo, and E. A. Zengotitabengoa, "Automatic text classification to support systematic reviews in medicine," Expert Systems with Applications, vol. 41, no. 4, pp. 1498-1508, 2014.

[15] N. Sánchez-Marono, A. Alonso-Betanzos, and M. Tombilla-Sanromán, "Filter methods for feature selection–a comparative study," in International Conference on Intelligent Data Engineering and Automated Learning, 2007, pp. 178-187: Springer.

[16] L. Talavera, "An evaluation of filter and wrapper methods for feature selection in categorical clustering," in International Symposium on Intelligent Data Analysis, 2005, pp. 440-451: Springer.

[17] T. N. Lal, O. Chapelle, J. Weston, and A. Elisseeff, "Embedded methods," in Feature extraction: Springer, 2006, pp. 137-165.

[18] A. K. Uysal and S. Gunal, "A novel probabilistic feature selection method for text classification," Knowledge-Based Systems, vol. 36, pp. 226-235, 2012.

[19] V. Srividhya and R. Anitha, "Evaluating preprocessing techniques in text categorization," International journal of computer science and application, vol. 47, no. 11, pp. 49-51, 2010.

[20] Z. Zheng, X. Wu, and R. Srihari, "Feature selection for text categorization on imbalanced data," ACM Sigkdd Explorations Newsletter, vol. 6, no. 1, pp. 80-89, 2004.

[21] H. Taira and M. Haruno, "Feature selection in SVM text categorization," in AAAI/IAAI, 1999, pp. 480-486.

[22] S. Günal, "Hybrid feature selection for text classification," Turkish Journal of Electrical Engineering and Computer Science, vol. 20, no. Sup. 2, pp. 1296-1311, 2012.

[23] G. BİRİCİK, B. Diri, and A. C. SÖNMEZ, "Abstract feature extraction for text classification," Turkish Journal of Electrical Engineering & Computer Sciences, vol. 20, no. Sup. 1, pp. 1137-1159, 2012.

[24] A. Dhar, H. Mukherjee, N. S. Dash, and K. Roy, "Text categorization: past and present," Artificial Intelligence Review, vol. 54, no. 4, pp. 3007-3054, 2021.

[25] M. Belazzoug, M. Touahria, F. Nouioua, and M. Brahimi, "An improved sine cosine algorithm to select features for text categorization," Journal of King Saud

University-Computer and Information Sciences, vol. 32, no. 4, pp. 454-464, 2020.

[26] H. Chantar, M. Mafarja, H. Alsawalqah, A. A. Heidari, I. Aljarah, and H. Faris, "Feature selection using binary grey wolf optimizer with elite-based crossover for Arabic text classification," Neural Computing and Applications, vol. 32, no. 16, pp. 12201-12220, 2020.

[27] B. M. Zahran and G. Kanaan, "Text feature selection using particle swarm optimization algorithm 1," 2009.

[28] F. Dong and Y. Zhang, "Automatic features for essay scoring–an empirical study," in Proceedings of the 2016 conference on empirical methods in natural language processing, 2016, pp. 1072-1077.

[29] V. Dasondi, M. Pathak, and N. P. Singh, "An implementation of graph based text classification technique for social media," in 2016 Symposium on Colossal Data Analysis and Networking (CDAN), 2016, pp. 1-7: IEEE.

[30] F. D. Malliaros and K. Skianis, "Graph-based term weighting for text categorization," in Proceedings of the 2015 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining 2015, 2015, pp. 1473-1479.

[31] L. Kumari, "Improved Graph Based K-NN Text Classification," Int J Eng Res Appl, vol. 3, pp. 928-931, 2013.

[32] C. Jiang, F. Coenen, R. Sanderson, and M. Zito, "Text classification using graph mining-based feature extraction," in Research and Development in Intelligent Systems XXVI: Springer, 2010, pp. 21-34.

[33] R. Liu, J. Zhou, and M. Liu, "Graph-based semi-supervised learning algorithm for web page classification," in Sixth International Conference on Intelligent Systems Design and Applications, 2006, vol. 2, pp. 856-860: IEEE.

[34] Y. Yang and J. O. Pedersen, "A comparative study on feature selection in text categorization," in Icml, 1997, vol. 97, no. 412-420, p. 35: Nashville, TN, USA.

[35] R. A. Calvo and H. A. Ceccatto, "Intelligent document classification," Intelligent Data Analysis, vol. 4, no. 5, pp. 411-420, 2000.

[36] T. DOĞAN and A. K. UYSAL, "The Effects of Globalisation Functions on Feature Weighting for Text Classification," in 2018 International Conference on Artificial Intelligence and Data Processing (IDAP), 2018, pp. 1-6: IEEE.

[37] B. Parlak and A. K. Uysal, "A novel filter feature selection method for text classification: Extensive Feature Selector," Journal of Information Science, p. 0165551521991037, 2021.

[38] S. Theodoridis, A. Pikrakis, K. Koutroumbas, and D. Cavouras, Introduction to pattern recognition: a matlab approach. Academic Press, 2010.

[39] T. Joachims, "Text categorization with support vector machines: Learning with many relevant features," in European conference on machine learning, 1998, pp. 137-142: Springer.

[40] G. Kou, P. Yang, Y. Peng, F. Xiao, Y. Chen, and F. E. Alsaadi, "Evaluation of feature selection methods for text classification with small datasets using multiple criteria decision-making methods," Applied Soft Computing, vol. 86, p. 105836, 2020.

[41] A. K. Uysal, "An improved global feature selection scheme for text classification," Expert systems with Applications, vol. 43, pp. 82-92, 2016.

[42] K. Nigam, A. K. McCallum, S. Thrun, and T. Mitchell, "Text classification from labeled and unlabeled documents using EM," Machine learning, vol. 39, no. 2, pp. 103-134, 2000.

Journal of Soft Computing and Artificial Intelligence

Journal homepage: https://dergipark.org.tr/en/pub/jscai

International Open Access

Volume 04
Issue 01

June, 2023

Review Article

A Review on Measurement of Particle Sizes by Image Processing Techniques

Vahit Tongur [1], *Ahmet Burçin Batıbay*[2], *Murat Karakoyun*[3*]

[1]*Software Engineering, Engineering and Natural Science Faculty, Konya Technical University, 42000, Konya, Turkey*
[2]*Metallurgical and Materials Engineering, Engineering Faculty, Necmettin Erbakan University, 42000, Konya, Turkey*
[3*]*Computer Engineering, Engineering Faculty, Necmettin Erbakan University, 42000, Konya, Turkey*

ARTICLE INFO

Article history:
Received December 13, 2022
Revised February 7, 2023
Accepted February 17, 2023

Keywords:
Image processing
Measurement
Particle size
Segmentation

ABSTRACT

This review is based on how to measure particle sizes with different image processing techniques. In addition, particle size significantly affects the material's mechanical properties. In material science, the material's structure is analyzed to understand that a material can provide specific standards, such as toughness and durability. Therefore, it is essential to make this measurement carefully and accurately. The segmentation approach, frequently used in image processing, aims to isolate objects in an image from the background. In this sense, separating particles from the background is a problem of image processing. In image processing applications, there are different approaches used in segmentation, such as histogram-based, clustering-based, region amplification, separation, and merging. In this review, a comparative analysis was examined by recent studies on particle size measurement.

1. Introduction

Developments in computer technology, data collection hardware, laser technology and automated imaging platforms have changed the field of biomedical research. Hardware systems that once needed manual intervention can now be programmed to run continuously for days or even weeks. In addition, high-content screening systems allow multiple experimental hypotheses to be automatically tested simultaneously [1].

Artificial Intelligence is abbreviated as AI technology. AI is a new branch of science that emerged in the 1950s. It not only studies technology, but also applies relevant technology to products and develops intelligent products. It is a technical discipline that resembles or partially resembles humans, used to help humans complete related activities and extend some human intelligence [2]. Breakthroughs in computer-aided diagnosis and AI will shortly transform how we diagnose diseases. The most encouraging development in AI is machine learning, the branch of science that allows computers to analyze and learn from data without human guidance. These technologies are often found in areas such as spell inspection and the development of self-driving cars, and are all performed by neural network algorithms [3]. Recently, modeling methods with artificial intelligence have gained significance. Parameters determined with many data and experiments can be estimated more practically and correctly with artificial intelligence methods. In addition to calculations, artificial intelligence methods are used to determine the cooling system parameters of nuclear reactors, fuel measurement, or any thermophysical properties [4].

Artificial intelligence technology is an interdisciplinary topic: with its development encompassing a wide range of content and intersecting with the fields of philosophy,

[*] Corresponding author

e-mail: mkarakoyun@erbakan.edu.tr

DOI: 10.55195/jscai.1218662

mathematics, statistics and so on, various more innovative theories and technologies have emerged. For now, the research and application directions of AI technology includes perception intelligence, thinking intelligence, learning intelligence and behavioral intelligence, etc. The AI algorithm mimics some laws; the algorithm is summarized by humans and then transformed into some algorithms to solve some problems. Under the background of the constant development of computer technology, the application of artificial intelligence algorithms obtains a better image processing effect to a certain extent, such as the application of several optimization algorithms in image processing [2].

As a medium containing a wealth of information, imagery is an essential resource for human beings to acquire and exchange information. In general, images are photographs, graphics, movies, videos, computed tomography (CT), magnetic resonance imaging (MRI), remote sensing and even two-dimensional or three-dimensional data. However, the image itself sometimes has some disadvantages. To extract reliable information from the image, it is necessary to process it.

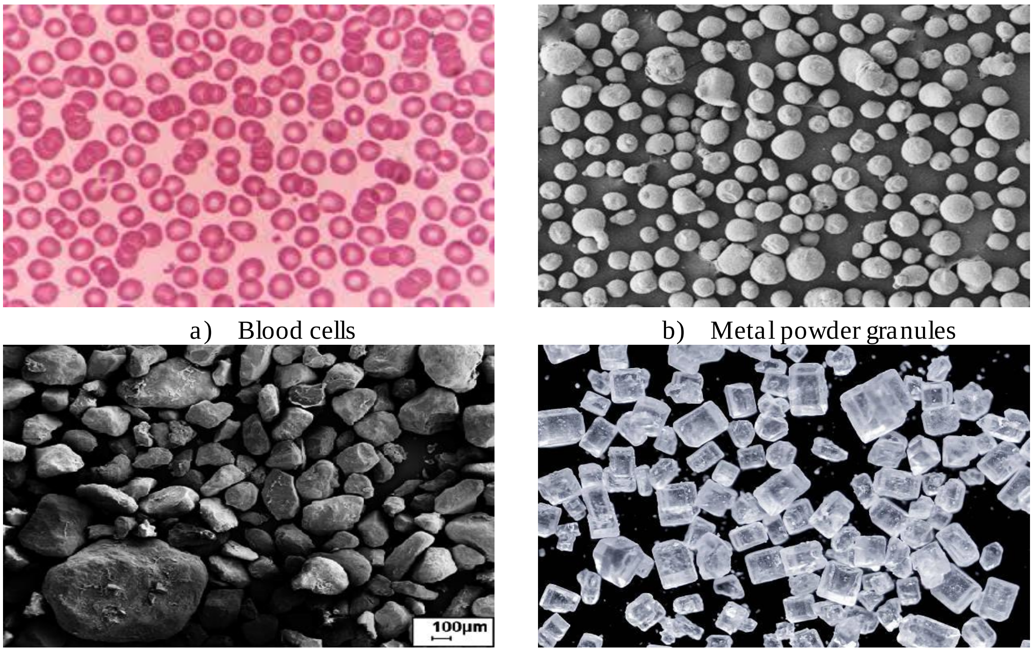
The main application areas of image processing include aerospace, terrain mapping, urban planning, medical research, combating product counterfeiting, engineering surface damage identification, real-time tracking, iris recognition and military, cultural, artistic and communicative aspects of human life, and business [2].

An image is equivalent to two dimensions for a machine and a computer. For the image to be processed by the computer, the image needs to be understood as a two-dimensional function, sampled, digitized and then processed. Therefore, image digitization is a necessary step for the computer to recognize images. After three steps of starting, developing, and popularizing, various disciplines have studied image processing, which has been widely used in various fields. Nowadays, with the fast development of science and technology, the science of image processing is making more and more progress both in theory and practice [2].

Particle size distribution of coarse-grain or fine-grain materials is figured out for experimental studies. Particle size distribution is a standard test used widely in different areas such as materials science, nanotechnology, biotechnology and metallurgy engineering to investigate the quality of materials. In recent decades, specific properties of particles are affected their manufacturing and application [5]. Size distribution is a crucial factor in increasing process performances. One of the earliest and most famous processes of size analysis was sieving or separating. Sieving is the cheapest method for measuring particle size. Particles are allocated according to their size and shape by particle size is defined as the diameter. Nonetheless, the action can be tedious and requires extra working time to process the fine and big particles. Accordingly, on-line particle size analysis was needed for the powder industry to be a critical impulse for particle sizing systems by imaging [6, 7]. Figure 1 presents sample images showing the particle distribution of materials in different areas.

There are several types of systems used for measuring particle sizes such as Dynamic Light Scattering (DLS) Analysis, X-Ray Diffraction (XRD), Scanning Electron Microscope (SEM) and light microscope. These systems are quite expensive and also are not accessible everywhere. The way to overcome this problem is to use an automatic image processing method. This method is fast and cheap for measuring particle size only needs a camera and a computer. And also, image processing methods can be used in biomedical and electronic industries. Besides using this approach, images are evaluated with little people interference and the images are more detailed and more specific than those achieved using a global one [5, 8].

Additionally, particle size distribution by using image processing has been investigated with shape and size characteristics in the past. This method precisely measures interparticle distance and aerodynamic size distribution from medical accessories. This method is more predictable than earlier theoretical models [5, 7].



a) Blood cells

b) Metal powder granules

c) Sand particles

d) Sugar crystals

Figure 1 Particle distribution of the different materials

One of the significant problems in practicing image processing and analysis methods for powder characterization is the specific selection of objects to allow accurate determination of their particular characteristics. Frequently, it cannot avoid touching or lapping over the particles in the images. This issue is demanding with fine without conducting powders due to electrostatic interaction [7, 9]. Particle images of different materials are presented in Fig. 1. Looking at the distribution of the particles in this example, it is seen that the particles in Fig. 1b are more discrete and separable, the particles in Fig. 1a, Fig. 1c and Fig. 1d are in contact or overlap. In this case, it can be said that the particle analysis of the image in Fig. 1b will be more accessible, on the contrary, the research in Fig. 1a, Fig. 1c and Fig. 1d will be more difficult. In summary, the state of the particle distribution in the images obtained from the materials can also affect the success of the application that analyzes and characterizes the material.

In general, there are five steps to analyze and characterize powder by using image processing. These steps include image acquisition, pre-processing, segmentation, extraction and representation of characteristic parameters. Image acquisition refers to digitalization as a crucial step for image processing techniques to convert images into numerical value for computers. The pre-processing step consists of fixing the image faults. This step includes different types of operations such as filters and edge detection. Segmentation is connected with extraction or discrimination. This step is to acquire a straight selection of the objects and also to uses specific algorithms such as Watershed [10]. Data extraction subjects to the applicable information to obtain specific parameters and shapes of individual particles for powder samples. Finally, representation is the last step to be processed the data by using the traditional statistic method and also shown by different types of graphs, diagrams and histograms [9, 11, 12]. Figure 2 shows the basic steps of an image processing application that performs particle size analysis.

This review's main aim is to research particle size measurement with different image processing techniques. With this review, which methods are used in the segmentation phase in image processing applications are examined in terms of shedding light on future studies.

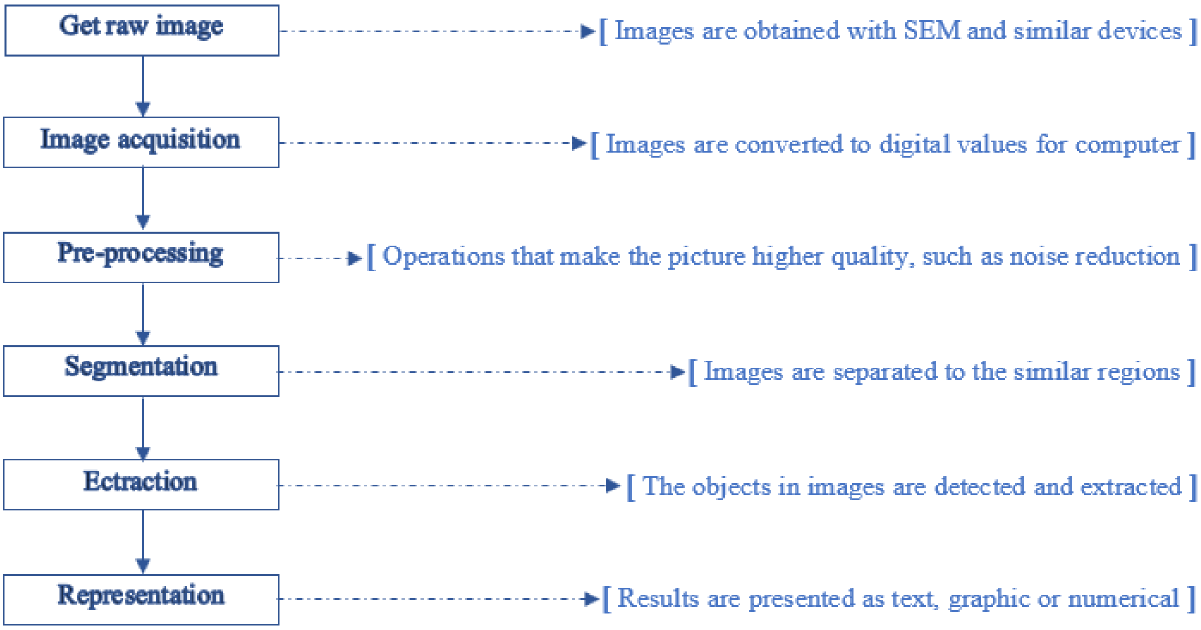

Get raw image
Image acquisition
Pre-processing
Segmentation
Ectraction
Representation
[Images are obtained with SEM and similar devices]
[Images are converted to digital values for computer]
[Operations that make the picture higher quality, such as noise reduction]
[Images are separated to the similar regions]
[The objects in images are detected and extracted]
[Results are presented as text, graphic or numerical]

Figure 2 The main steps of an image processing application for particle analyzing

2. Different Approaches

Particle size measurement has long been used in different fields and has attracted the attention of many researchers. Li et al. [13] proposed an algorithm to automate the intercept method for measuring the average particle size of metallic materials. The algorithm can extract continuous and closed particle regions using topological skeletons and measure average particle size by recognizing and classifying the intersections between selected test models and particle boundaries. They used image smoothing and thresholding methods as pre-process steps. They applied their algorithm on 200 microscope images includes different material microstructures. The stated that the proposed automatic algorithm achieved an overall accuracy that is more than %98.

Ro et al. [14] have developed a new microscopic image analysis to increase the accuracy of particle size estimation of very small fragmented minerals. In their work, they have used individual particle segmentation, shape factor, watershed and deep learning methods. For the accuracy of the results obtained, they have sieved three samples as ferruginous quartzite, coal and magnetite and compared them with the results found. Since the samples used were very small, they applied the image stitching method before the image processing. After, they have used both watershed segmentation and deep neural network methods to separate overlap.

Akkoyun and Ercetin [15] used some methods of the image processing includes thresholding, filtering and contouring to calculate the number and size of the particles in Mg alloys. The experiment images taken from SEM. They compared the results of the proposed method with the manual measurement results. They stated that the success of their automatic system is 94% with a 6% standard deviation between automatic and manual results.

Guo et al. [16] have used an improved watershed segmentation method to segment rock fragments. Rock images have been captured using tilt photogrammetry to reduce measurement errors of overlapping rock blocks. Then, image preprocessing techniques have been used to determine the target rock area. Then, morphological operations have been applied to the obtained binary image and the watershed segmentation method has been applied after the distance transformation has been calculated. The performance of the proposed method has been compared with the results of manual screening and the methods in the literature.

Hu et al. [17] stated that traditional watershed and threshold segmentation algorithms are insufficient and the total number of processed particles in the image processing process affects the result. In their work, a deep learning algorithm (DLA) is presented to segment roughly stacked batch images. Aggregate images with different sizes and mixtures have been segmented and the results analyzed. Test results have been compared with the traditional watershed algorithm. Experimental results showed that the deep learning algorithm overcomes the problem of over-partitioning and under-partitioning of the watershed algorithm. In addition, desired results have been obtained in the partitioning of aggregates with different particle sizes and aggregate materials.

An image segmentation method for coal particle size distribution analysis was investigated by Bai et al. [18]. To avoid over-segmentation, the technique used a gradient watershed for pre-segmentation and applied the k-nearest neighbor (KNN) algorithm for region merging. As a result, many images were automatically segmented using the proposed method and the manual method. The proposed method effectively estimated the particle size of coal particles with low dust and moisture content, as the difference in particle size distribution between the two ways was very small. Therefore, the particle size and the number of particle size ranges were used to estimate the size distribution of coal particles. In addition, because it is more automatic, it can reduce the labor of workers and provide real-time data to other equipment. Thus, it can improve the automation level of the whole coal preparation plant.

Yang et al. [19] suggested applying an image processing technique to measure coal pieces produced by uniaxial compression tests. The image processing method based on the MATLAB image processing toolbar was proposed in this paper. The watershed method was applied for fragment segmentation. The contrast between the images demonstrates the suitability of the image processing method proposed for coal seam measurement before and after image processing. The image's fragment was considered an ellipsoid characterized by major, minor, and minor axes. The image-processed cumulative distribution of coal samples was determined based on the image analysis results, the ellipsoid volume equation, and the center-lateral axis value relationship. Comparisons between image processing, manual sieving, and fractal-modeled cumulative distribution curves indicated that digital image processing is an effective and accurate tool for measuring the size distribution of coal fragments. In future applications, only a single photograph will need to be taken and processed, saving more time in the image acquisition process. Image analysis is based on MATLAB coding, and data will be stored automatically. However, code retrieval and data analysis are completed with human interaction. It is possible to make all these processes smarter with programming, artificial intelligence, and deep learning applications.

A variety of ore images and an annotated dataset were studied by Li et al. [20]. In order to solve the problem of over- and under-segmentation in the task of ore image segmentation, the neural network-based U-Net model and a marked watershed were used. The purpose of ore image segmentation in an outdoor environment was achieved by this method. The experimental results have shown that the proposed method has high speed, strong robustness, and high accuracy characteristics. It has great practical value for the actual task of statistical analysis of ore grains.

Wu et al. [21] worked to find a solution to the problem experienced in the measurement of overlapping particles in the particle measurement of iron green pellets. For this purpose, they developed an approach based on morphological operations in the first stage and circle scanning in the second stage. They applied their method at a local steel company, which was a real test environment. They stated that they compared their methods with the success of manual measurement and other methods and they achieved a success rate of 94.3%.

Watano and Miyanami developed a monitoring system for granule growth in the pharmaceutical industry [22]. They have prepared a special mechanism for the system they have developed. This mechanism used a charge-coupled device (CCD) camera with particle-adjusted illumination. Some preprocessing, such as filtering and noise removal, was done on the images obtained with this camera. After these processes, some granules overlapping were observed, and changes were applied on the CCD camera to separate them. After these changes, they tried to separate the particles using circle pattern matching and the eight-neighbor erosion methods on the images they obtained. However, it was observed that most of the particles did not separate. Therefore, these methods were applied twice. Thus, the separation success rate was increased. After this step, the Feter diameter of each object was calculated. The accuracy of the results obtained was compared with the values measured from the fluidized bed granulation system, and the comparison results were similar.

Mora and Kwan used digital image processing to measure coarse concrete's shape factor, sphericity, and convexity. The method used in this study is suitable for estimating both the volume and thickness of the particles [23]. They worked on forty-six rock samples with three different types taken from five sources. Once the sample image was obtained, the particles and the background were distinguished from each other by increasing the contrast between the particles and the background and drawing the particle edges. Then, geometric analysis was performed for size and shape measurement according to the determined particle boundaries.

Lin et al. presented a three-dimensional watershed algorithm for the Automatic Segmentation of Nuclei in Confocal Image Stacks. They performed their experiments on a small microscopic piece of the rat hippocampus. In the segmented region, it was observed connected many objects. For the segmentation to be calculated correctly, connected objects must be separated. The Watershed algorithm needs to provide information about objects' shapes and sizes. This algorithm is generally used to separate connected objects. However, two different image transformations are commonly used to determine the boundaries surrounding the basins. The first is the distance transform. This transformation is usually related to the shapes of objects. However, this method produces better results for regular geometric shapes. Gradient transformation usually causes over-segmentation. Therefore, the authors proposed a hybrid method called "gradient-weighted distance transform", which includes both methods. When preprocessing is done using this method, it is observed that the watershed algorithm also parses singular objects that are not connected. To overcome this, merge methods are implemented to separate singular objects. The results obtained were verified with a human observer and a success rate of 97% was achieved. In addition, a significant improvement was achieved compared to previous studies [24].

Tek et al., in their work, proposed an approach based on the watershed algorithm and Radon transformation (Radon, 1986) for segmenting blood cell images. The images used in this study are microscopic peripheral blood which may include red cells, parasites, and white cells. The proposed approach is a stepwise algorithm with several different steps. An initial rough segmentation has been done with the minimum watershed area (a modified version by authors) as first step. And then, Radon transformation has applied to the rough segmented image to extract markers. Finally, the watershed algorithm has been re-applied to the images generated after the Radon transformation to achieve the final segmented image. To evaluate the performance and accuracy of the proposed approach, the authors used a set of data having 20 microscopic blood cell images, including 2177 different cells (red blood cells, white blood cells, and infected cells with parasites) by using counting and locating criteria. In the experimental results, they stated that they were quite successful, with a high segmentation success of 95.4% [25].

Zelelew et al. proposed a stepwise image processing system named Volumetric-based Global Minimum to segment the asphalt concrete images. Their algorithm includes some pre-processing methods of image processing and thresholding phase and is applied to X-ray computed tomography images of asphalt concrete. The volumetric attributes of asphalt concrete were used to determine the border thresholds of the air-mastic and mastic-mass. So, the proposed algorithm was applied to characterize the microstructure of the asphalt concrete. They stated that the images obtained by their system had satisfied improvement compared to the raw images and that images were ready for the following processes. Thus, they specified that their algorithms are more successful and advanced than many methods used manually in this field [26].

Liao and Tarng developed an automatic optical inspection system based on image processing to measure the size distribution of rough particles. The proposed system works online and consists of four sub-modules: a particle partition module, an image-acquiring module, an image processing/analyzing module, and an electronic inspection module. The system was applied to particles that are non-uniform to measure some metrics of the particles, like the number of particles, particle size distribution, and accumulated weight percentages. They conducted a regression analysis between the results obtained by their system and the traditional net sieving system to measure the quality and accuracy of their work.

According to the results, they stated that the proposed system achieved satisfied precision and accuracy on rough particles and additionally added that their system can be a good alternative to measure the different coarse particles like dolomite, sinter, serpentine, limestone etc. [27].

Sharif et al. proposed a stepwise image processing system to segment the blood cell images. Their work involves some pre-processing steps consists of filtering, image enhancement, color conversion, and segmentation. Typically, the counting of the blood cell is done manually via a hematocytometer using a counting chamber. A complete blood count process includes white blood cells, hematocrit, hemoglobin, platelet, and red blood cell analyses. Each of these elements is very important for the body system and gives information about the body's capacity. The work aims to provide an automatic segmentation system that achieves complete cell count by using some pre-process steps and watershed algorithm [28].

Schorsch et al. proposed a study measuring the size distribution of particles during crystallization [29]. They prepared a special experimental mechanism for measurement and obtain images from it. Before particle size measurement, they performed some tests to minimize the overlap of particles and observed that the overlap was minimal under brighter light.

Then, they carried out image analysis on servers that could perform real-time and parallel analysis. As a result of these analyses, the particles were divided into three classes: sphere, needle, and cube. In their experimental studies, the capabilities of the proposed methods for the size and shape distributions of particles were tested. For this, the experimental results were compared with the Coulter multisizer [30], which is known as a standard for particle size measurement. The results obtained from both methods were observed very close. The authors stated that the device and the image analysis algorithm showed high performance in characterizing the particles' size and shape for different samples. They noted that these results would provide a better understanding of how crystal size and shape depend on the operating conditions of the process and how they can be controlled. They also noted that it led to the developing of a powerful technique for monitoring crystallization.

Bahrami and Honarvar, in their work, stated that raw cane sugar is a very significant raw stuff of white sugar in the sugar industry, and the morphological/physical features of the raw material can directly affect the quality of the final white sugar. So, they aimed to measure the morphological characteristics (perimeter, area size, squareness and crystal numbers) of the raw cane sugar using image processing tool Matlab on the crystal that generated by the flatbed scanner. A data set consisting of two groups of raw sugar cane, imported and domestic, was created to be used in the study. Based on the results they obtained from their studies, they stated that the digital image processing technique they used was useful in determining both the morphological and physical properties of different raw sugar crystals and could be used as an alternative [31].

Pavithra and Bagyamani [32] used watershed and Circular Hough Transform (CHT) [33] image processing techniques for white blood cell count in their study. The number of blood cells is an essential factor in treating of many diseases. A microscopic image obtained for the blood cell count contains around 100 red blood cells, while the white blood cell is only 1 or 3. Therefore, manual blood cell counting is unreliable and can give incorrect results. For all these reasons, they proposed a fully automated method for white blood cell counting in their study. First, preprocessing was applied to the image taken from the microscope. Since white blood cells are darker than red blood cells, the image is converted to a gray level according to a certain threshold value. Pixels below the threshold are black, while pixels above it are white. Thus, the noise and other blood cells in the image are cleared. After the image preprocessing, the median filtering image enhancement method was used to make the objects on the image more apparent. Then, the gradient magnitude technique was used for image segmentation. The Sobel edge detection algorithm was used to make the border of white blood cells more specific. In addition, the watershed algorithm was applied to prevent over-segmentation. Then morphological operations were used. Morphological operations help to determine the shape of the object. Next, erosion and morphological dilation methods were applied to smooth the image and remove unwanted small pixels. Then, opening and closing operations were applied. After all these procedures, the CHT method was used for white blood cell count.

This method calculates the number of white blood cells by calculating the center point and radius of circular shapes on the image. In experimental studies on 20 images obtained from European Leukemia Net, the success rate was observed between 74% and 100%.

In their work, Cai and Su performed a step-by-step image processing study to segment particle images with missing illumination, blurred, and cohesive/adjacent structure. Various improvements have been made in the basic levels of image processing to segment particles consisting of uneven illumination, focus blur, and contact each other automatically and accurately and bypass the deficiencies of existing methods in the literature. As the first step, the grade of focus blur is reduced by analyzing and determining the optimal viewing plane of the particles. Then, the effect of uneven lighting is also eliminated using an adaptive thresholding method. Finally, with the help of the image reconstruction method, changing the result of the Euclidean distance transform of the binary image and coupled with the watershed transform, the touching particles are segmented efficiently. They stated that: experimental results and error analyses show that the proposed method can more accurately segment the contacting particles and also effectively avoid the over-partitioning problem [34].

Jagadev and Virani proposed a study that identified leukemia patients using image processing and machine learning algorithms. They used three different methods in their studies. These are k-means, Marker controlled watershed and Hue-Saturation-Value (HSV) color-based segmentation algorithms. Two hundred images obtained from the Goa Medical College online database were studied. These images include images of patients with leukemia and healthy individuals. First, the data in Cyan-Magenta-Yellow-Key (CMYK) was converted to Red-Green-Blue (RGB). The k-means clustering algorithm was used for the first image segmentation. Here k is taken as 3. These are nuclei (high saturation), background (high brightness and low saturation), and other cells. The distance between the two classes was used as the Euclidean distance. Then each pixel in the cluster is labeled as a cluster index. Second, the Marked controlled watershed algorithm was used. After the foreground objects and background positions were marked separately, the watershed algorithm was applied. Thus, over-segmentation is prevented. Finally, HSV color-based segmentation method was used [35]. Finally the support vector machine (SVM) [36] method was applied to determine whether the samples had leukemia or not.

Meng et al. proposed an algorithm for automated particle size distribution measurement [37]. The proposed algorithm used local adaptive canny edge detection and modified circular Hough transform methods. The images in their studies were obtained from SEM and transmission electron microscopy (TEM) imaging devices. Median filtering [38] method was used to remove noise in the image before the image processing steps. Edges of objects are an important feature in image processing. Various edge detection algorithms have been developed such as Roberts [39], Prewitt [40], Kirsch [41], Sobel [42], Robinson [43] and Canny [44]. Canny edge detection includes a multi-step algorithm with two thresholds, Th and Tl, for detecting and connecting edges. It is not easy to manually select these two threshold values. Therefore, the authors proposed a local adaptive Canny edge detection method. Tl was set as 0.4, while Th was obtained by the Otsu [45] algorithm. The Hough transform (HT) is used to detect geometric shapes. The basic Hough transform was designed for line detection but has been later extended to other shapes. Circular HT (CHT) is a special HT that has proven itself to detect circular objects. The basic principle of CHT is to convert geometric coordinates such as (x, y) into Hough parameter space. The authors proposed a modified CHT method. This method separates the basic CHT into two stages. These are to find the circle center and to determine the circle radius. Experimental studies were shown that modified CHT could detect more circles on overlapping images. The size distributions of the circles obtained by this method were determined and compared with the manual measurement values. Experimental studies were shown that the results obtained from both methods are very close. This proved the success of the proposed method.

Yang et al. used image processing to measure the size of the produced sand grains because they knew that the concrete quality was significantly affected by the grain structure of the sand it contained [46]. The authors developed a sand casting (dispersion) system based on the characteristics of the sand particle contours and an extraction mechanism to overcome the lack of the traditional vibratory screening method. Their work includes both hardware and software parts. Their work includes both hardware and software parts. A sand dispersal vehicle has been designed and developed as a hardware part that can completely disperse the falling sand. On the other hand, a segmentation method was used for the system's software. The Otsu's maximum variance between clusters approach was used as a segmentation method. The segmentation process separated the filtered sand particles from the background. In the study, some settings about particle size are done according to the JTG E42-2005 [47] standard. As a result of the study, they stated: Their system performs non-contact measurements of the produced sand without damaging the particles. It is faster and cheaper than classical methods. This system can also be applied to different material particles, not limited to the sand produced.

Laucka et al. used image processing techniques to measure the size of powder granules used in the chemical industry. In their study, they established an experimental setup to capture the image of the granules. Images of granule particles in the samples were obtained with a line scan camera. They used two different examples in their study. These are monocalcium phosphate and ammonium nitrate. The results needed a correction due to the different compositions of the particle measurement. Therefore, an artificial neural network model was used to increase the efficiency of the particle correction algorithm. In addition, SVM was used for image classification. The main disadvantage of the proposed method is that it gives much better results when the particles are in a certain ratio of circular shapes. The experimental setup was also set up in a fertilizer production factory, and measurements were made every 5 to 7 minutes during actual production [48].

Patmonoaji et al. tested the watershed method for pore-throat identification in unconsolidated porous media with various particle sizes and shapes. One of the most important steps in the watershed method is to find the parameter value that may cause over-segmentation and under-segmentation. If this parameter is below the optimum value, over-segmentation occurs, and if it is above, under-segmentation occurs. Therefore, a value close to the local peak was chosen. Three types of regular spherical particles were modeled to validate the results from the watershed model [49].

Biswal et al. estimated the grain boundaries of the images obtained from the optical microscope using image processing methods. A Median filter was applied to remove noise from images. Histogram Equalization was used for this. Edge detection methods such as Sobel, Robert, Prewitt, and Canny were used to detect granules. The results obtained from these methods were compared, and the canny edge detection method was chosen. Using the canny edge detection method, a watershed transform was applied with the marker-controlled approach on an image. The marker was used to control over-segmentation. After the watershed transformation, the average particle size was found for all the samples used in the experimental studies [50].

Cohn et al. used instance segmentation, a useful and advanced tool in computer vision [51]. The Mask region-based convolutional neural network (R-CNN) [52] which Facebook presented in 2017, was used for the instance segmentation process. The authors aimed to show the utility of instance segmentation in material science using the Mask R-CNN. The images of the metal powder particles generated by the SEM were used for the segmentation. They trained two independent models for segmenting the powder particles and satellites in each image. Training steps improved the models with five particle and ten satellite image instances, respectively. Precision and recall measurement metrics were used to analyze the results of the segmentation of the powder particle images. They stated that false positives occurred because of the separation of large pieces formed by the fusion of particles. On the other hand, according to the authors, the cause of the false negatives is incomplete small, or extremely clogged particles that do not contain a strong visual signal. That Mask R-CNN was also used on the spheroid particles taken from the Ultra High Carbon Steel database.

3. Discussion

This study collects papers related to particle measurement (blood cells, metal powder, sand, sugar, etc.) in the literature. The objects and methods used in these papers and the result of the study are summarized in Table 1. According to Table 1, the Watershed and Thresholding methods are used more frequently for particle measurement. In addition, Edge Detection and Hough Transform methods are also widely used. The examined materials are usually microscopic objects. For this reason, it has been seen that image processing techniques are used for the size of the materials. First, images of the materials to be measured are taken by devices such as SEM. Afterward, the material particles' size is calculated using image processing techniques. In Table 1, it is seen that particle measurements are made in many areas, from blood cells to metal powders, from cane sugar to granules, and from asphalt concrete to sand particles. Although there are different types of materials, it can be seen that measurement techniques are similar. Table 1 shows that the success of the Canny and Hough Transform methods are close to each other, but the edge detection methods are more successful in larger objects. Although the results obtained from the studies on soil particles in the Watershed method could be more satisfactory, it is observed that 97% success is achieved in the studies performed on rat cells. This indicates that a suitable method should be selected for the study material. In addition, it is understood from the studies that the watershed method is an appropriate method for segmentation (especially for contacted particles). It can be seen in the studies that hybrid methods are also used instead of a single image processing method. Hybrid methods have increased the success of the results obtained.

Article	Data	Techniques	Results
[5]	Blood cells	Edge detection and morphological operator	This method is more accurate for larger objects and high-resolution images.
[6]	Limestone, coal, rounded stone, pyrite and iron ore	Watershed	Results showed some enhancements when applying the proposed approach to non-overlapped particles.
[7]	Quartz sands	Split desktop tool	Without wasting time and practical to other method.
[8]	Nanosilver powder	Thresholding	Results show that the suggested image-processing technique from SEM micrographs could measure nanoparticles that cannot be detected by DLS analysis.
[9]	Zirconia powder and adipic acid particles	Watershed and morphological operator	Segmentation process results are shown to be convenient to enable accurate edge detection for heterogeneous particles.
[11]	Soil particles	Watershed	Obtained results of soil samples by image analysis are not satisfied according to the expected results.
[12]	Nanomaterial powder	Thresholding, filtering and morphological operator	The proposed approach is convenient for analyzing the powder particle size.
[13]	Iron and stainless steel	Thresholding and morphological operator	The algorithm can extract continuous and closed particle regions and measure average particle sizes.
[14]	Ferruginous quartzite, coal and magnetite	Watershed and deep learning	Watershed segmentation and deep neural network were successfully applied to separate the overlaps.
[15]	Mg alloys	Thresholding, filtering, contouring	A success rate of 94% was observed between automatic and manual measurement.
[16]	Rocks	Morphological operators and Watershed	The proposed method has been successful in segmenting rock images.
[17]	Aggregate materials	Deep learning	Desired results have been obtained in the partitioning of aggregates with different particle sizes and aggregate materials.
[18]	Coal particles	Watershed and KNN	The proposed method effectively estimated the particle size of coal particles with low dust and moisture content.

[19]	Coal pieces	Watershed	The results showed indicated that digital image processing is an effective and accurate tool for measuring the size distribution of coal fragments.
[20]	Ore grains	NN-based U-Net and Watershed	The proposed method has high speed, strong robustness, and high accuracy characteristics.
[21]	Iron green pellets	Morphological operator and circle-scan	The proposed approach achieved 94.3% success in measuring overlapping particles.
[22]	Granules	Thresholding and morphological operator	Image processing results could detect both granule growth in the low and rapid granule growth in the high moisture with high accuracy.
[24]	Rat hippocampus	Watershed	The results obtained were verified with a human observer, and a success rate of 97% was achieved.
[25]	Blood cells	Watershed and Radon transformation	Results are quite successful, with a high segmentation accuracy of 95.4%.
[26]	Asphalt concrete	Enhancement thresholding	The proposed thresholding algorithm greatly enhances the traditional techniques used in the literature.
[27]	Rough particles	Thresholding	The proposed system achieved satisfied precision and accuracy on rough particles and can be applied to different rough particles such as dolomite, sinter, serpentine, limestone etc.
[28]	Blood cells	Watershed	This work illustrates an automatic segmentation system that reaches a complete cell count using pre-processing steps and a watershed algorithm.
[31]	Cane sugar	Matlab tool	Results show that the digital image processing technique can be useful in determining the morphological and physical properties of different raw sugar crystals as an alternative.
[32]	Blood cells	Watershed and Hough transform	The results' success rate is between 74% and 100%.
[34]	Noisy particles	Thresholding and watershed	Results illustrate that the proposed method can be more accurate by segmenting contacted particles.
[35]	Blood Cells	K-means, watershed and HSV	Many features are extracted to make the detection process more exact and detailed.
[35]	Carbon nanospheres	Canny and Hough transform	Results illustrate that both methods are very close. This proves the success of the proposed method.
[46]	Sand particles	Thresholding	The proposed system applies non-contact measurement of the sand without damaging the particles and is faster and cheaper than classical methods.
[48]	Powder granules	ANN and SVM	The most detailed results of measurements with sieves are reached for correction by using an ANN.
[50]	Aluminum based hybrid composite	Watershed and Canny	Results show that it can be a new image processing technique for grain boundary analysis.
[51]	Metal powders	Mask R-CNN	Results show that Mask R-CNN can be useful for automating image segmentation in material science.

4. Conclusion

Different studies in the literature are examined in Table 1. In these reviews, a summary of various image processing techniques used in many different fields is given. It has been seen that image processing techniques are used to count the blood cell in the health area, investigate different particles (metal powder, sand, cement, etc.) in materials science, and measurement of sugar particles in the food industry. Among these techniques, edge detection and heuristic algorithms, morphological operators, transformation approaches, thresholding, and tools that can be ready have been used in some studies. As it can be understood from the studies in the literature, it has been stated that different methods are recommended for different data. The same method has been observed to give different results on different data. The research concluded that any method was not successful in every data and that the method suitable for the data should be meticulously determined. Considering the past to the present, it is seen that the subject is up-to-date, existing methods and new techniques have continued to be researched, and many recent studies have been carried out in different fields.

Acknowledgments

The authors acknowledge the financial support from Necmettin Erbakan University Scientific Research Project under grant number 211219011.

References

[1] G. González and C. L. Evans, "Biomedical Image Processing with Containers and Deep Learning: An Automated Analysis Pipeline: Data architecture, artificial intelligence, automated processing, containerization, and clusters orchestration ease the transition from data acquisition to insights in medium-to-large datasets," *BioEssays*, vol. 41, no. 6, p. 1900004, 2019.

[2] X. Zhang and W. Dahu, "Application of artificial intelligence algorithms in image processing," *Journal of Visual Communication and Image Representation*, vol. 61, pp. 42-49, 2019.

[3] S. Robertson, H. Azizpour, K. Smith, and J. Hartman, "Digital image analysis in breast pathology—from image processing techniques to artificial intelligence," *Translational Research*, vol. 194, pp. 19-35, 2018.

[4] O. E. Akay and M. Das, "Modeling the total heat transfer coefficient of a nuclear research reactor cooling system by different methods," *Case Studies in Thermal Engineering*, vol. 25, p. 100914, 2021.

[5] S. V. Seyedin, S. H. Seyedin, and A. S. Seyedin, "Designing and programming an efficient software for sizing and counting various particles using image processing technique," *BRAIN. Broad Research in Artificial Intelligence and Neuroscience*, vol. 10, no. 2, pp. 103-118, 2019.

[6] S. Al-Thyabat and N. Miles, "An improved estimation of size distribution from particle profile measurements," *Powder Technology*, vol. 166, no. 3, pp. 152-160, 2006.

[7] I. Kursun, "Particle size and shape characteristics of kemerburgaz quartz sands obtained by sieving, laser diffraction, and digital image processing methods," *Mineral Processing & Extractive Metallurgy Review*, vol. 30, no. 4, pp. 346-360, 2009.

[8] E. Emil Kaya, O. Kaya, G. Alkan, S. Gürmen, S. Stopic, and B. J. M. Friedrich, "New proposal for size and size-distribution evaluation of nanoparticles synthesized via ultrasonic spray pyrolysis using search algorithm based on image-processing technique," vol. 13, no. 1, p. 38, 2020.

[9] A. Nazar, F. Silva, and J. Ammann, "Image processing for particle characterization," *Materials characterization*, vol. 36, no. 4-5, pp. 165-173, 1996.

[10] S. Beucher, "Use of watersheds in contour detection," in *Proceedings of the International Workshop on Image Processing*, 1979: CCETT.

[11] K. S. Naphade, "Soil characterization using digital image processing," Lehigh University, 1999.

[12] L. C. Wu and C. Yu, "Powder particle size measurement with digital image processing using Matlab," in *Advanced Materials Research*, 2012, vol. 443, pp. 589-593: Trans Tech Publ.

[13] X. Li *et al.*, "Automation of intercept method for grain size measurement: A topological skeleton approach," *Materials & Design*, vol. 224, p. 111358, 2022.

[14] S. Ro, J. Jon, and K. Ryu, "A method for improving the estimation accuracy of the particle size distribution of the minerals using image analysis," *Computational Particle Mechanics*, pp. 1-13, 2022.

[15] F. Akkoyun and A. Ercetin, "Automated Grain Counting for the Microstructure of Mg Alloys Using an Image Processing Method," *Journal of Materials Engineering and Performance*, vol. 31, no. 4, pp. 2870-2877, 2022.

[16] Q. Guo, Y. Wang, S. Yang, and Z. Xiang, "A method of blasted rock image segmentation based on improved watershed algorithm," *Scientific Reports*, vol. 12, no. 1, p. 7143, 2022.

[17] X. Hu, H. Fang, J. Yang, L. Fan, W. Lin, and J. Li, "Online measurement and segmentation algorithm of coarse aggregate based on deep learning and experimental comparison," *Construction and Building Materials*, vol. 327, p. 127033, 2022.

[18] F. Bai, M. Fan, H. Yang, and L. Dong, "Image segmentation method for coal particle size distribution analysis," *Particuology*, vol. 56, pp. 163-170, 2021.

[19] X. Yang, T. Ren, and L. Tan, "Size distribution measurement of coal fragments using digital imaging

processing," *Measurement*, vol. 160, p. 107867, 2020.

[20] H. Li, C. Pan, Z. Chen, A. Wulamu, and A. Yang, "Ore Method Based on U-Net and Watershed," *Computer, Materials & Continua*, vol. 65, no. 1, pp. 563-578, 2020.

[21] X. Wu, X.-Y. Liu, W. Sun, C.-G. Mao, and C. Yu, "An image-based method for online measurement of the size distribution of iron green pellets using dual morphological reconstruction and circle-scan," *Powder Technology*, vol. 347, pp. 186-198, 2019.

[22] S. Watano and K. Miyanami, "Image processing for on-line monitoring of granule size distribution and shape in fluidized bed granulation," *Powder technology*, vol. 83, no. 1, pp. 55-60, 1995.

[23] C. Mora and A. Kwan, "Sphericity, shape factor, and convexity measurement of coarse aggregate for concrete using digital image processing," *Cement and concrete research*, vol. 30, no. 3, pp. 351-358, 2000.

[24] G. Lin, U. Adiga, K. Olson, J. F. Guzowski, C. A. Barnes, and B. Roysam, "A hybrid 3D watershed algorithm incorporating gradient cues and object models for automatic segmentation of nuclei in confocal image stacks," *Cytometry Part A: the journal of the International Society for Analytical Cytology*, vol. 56, no. 1, pp. 23-36, 2003.

[25] F. B. Tek, A. G. Dempster, and I. Kale, "Blood cell segmentation using minimum area watershed and circle radon transformations," in *Mathematical morphology: 40 years on*: Springer, 2005, pp. 441-454.

[26] H. Zelelew, A. Papagiannakis, and E. Masad, "Application of digital image processing techniques for asphalt concrete mixture images," in *The 12th International Conference of International Association for Computer Methods and Advances in Geomechanics (IACMAG)*, 2008, pp. 119-124: Citeseer.

[27] C. Liao and Y. Tarng, "On-line automatic optical inspection system for coarse particle size distribution," *Powder Technology*, vol. 189, no. 3, pp. 508-513, 2009.

[28] J. M. Sharif, M. Miswan, M. Ngadi, M. S. H. Salam, and M. M. bin Abdul Jamil, "Red blood cell segmentation using masking and watershed algorithm: A preliminary study," in *2012 international conference on biomedical engineering (ICoBE)*, 2012, pp. 258-262: IEEE.

[29] S. Schorsch, T. Vetter, and M. Mazzotti, "Measuring multidimensional particle size distributions during crystallization," *Chemical Engineering Science*, vol. 77, pp. 130-142, 2012.

[30] W. R. Hogg and W. H. Coulter, "Apparatus and method for measuring a dividing particle size of a particulate system," ed: Google Patents, 1971.

[31] M. Bahrami and M. Honarvar, "Measurement of morphological characteristics of raw cane sugar crystals using digital image analysis," 2015.

[32] S. Pavithra and J. Bagyamani, "White blood cell analysis using watershed and circular hough transform technique," *Int. J. Comput. Intell. Inform*, vol. 5, no. 2, pp. 114-123, 2015.

[33] H. Rhody, "Lecture 10: Hough circle transform," *Chester F. Carlson Center for Imaging Science, Rochester Institute of Technology*, 2005.

[34] Y. Cai and M. Su, "An improved image processing method for particle measurement," in *2017 IEEE International Conference on Imaging Systems and Techniques (IST)*, 2017, pp. 1-6: IEEE.

[35] P. Jagadev and H. Virani, "Detection of leukemia and its types using image processing and machine learning," in *2017 International Conference on Trends in Electronics and Informatics (ICEI)*, 2017, pp. 522-526: IEEE.

[36] R. Dietrich, M. Opper, and H. Sompolinsky, "Statistical mechanics of support vector networks," *Physical review letters*, vol. 82, no. 14, p. 2975, 1999.

[37] Y. Meng, Z. Zhang, H. Yin, and T. Ma, "Automatic detection of particle size distribution by image analysis based on local adaptive canny edge detection and modified circular Hough transform," *Micron*, vol. 106, pp. 34-41, 2018.

[38] T. Huang, G. Yang, and G. Tang, "A fast two-dimensional median filtering algorithm," *IEEE transactions on acoustics, speech, and signal processing*, vol. 27, no. 1, pp. 13-18, 1979.

[39] L. G. Roberts, "Machine perception of three-dimensional solids," Massachusetts Institute of Technology, 1963.

[40] J. M. Prewitt, "Object enhancement and extraction," *Picture processing and Psychopictorics*, vol. 10, no. 1, pp. 15-19, 1970.

[41] R. A. Kirsch, "Computer determination of the constituent structure of biological images," *Computers and biomedical research*, vol. 4, no. 3, pp. 315-328, 1971.

[42] I. Sobel and G. Feldman, "A 3x3 isotropic gradient operator for image processing," *a talk at the Stanford Artificial Project in*, pp. 271-272, 1968.

[43] G. S. Robinson, "Edge detection by compass gradient masks," *Computer graphics and image processing*, vol. 6, no. 5, pp. 492-501, 1977.

[44] J. Canny, "A computational approach to edge detection," *IEEE Transactions on pattern analysis and machine intelligence*, no. 6, pp. 679-698, 1986.

[45] N. Otsu, "A threshold selection method from gray-level histograms," *IEEE transactions on systems, man, and cybernetics*, vol. 9, no. 1, pp. 62-66, 1979.

[46] J. Yang, W. Yu, H.-y. Fang, X.-y. Huang, and S.-j. Chen, "Detection of size of manufactured sand particles based on digital image processing," *PloS one*, vol. 13, no. 12, p. e0206135, 2018.

[47] P. China, "Ministry of Communications. JTGE 42-2005 test methods of aggregate for highway engineering [S]," ed: Beijing: China Communications Press, 2005.

[48] A. Laucka, V. Adaskeviciute, and D. Andriukaitis, "Research of the equipment self-calibration methods for different shape fertilizers particles distribution by size using image processing measurement method," *Symmetry*, vol. 11, no. 7, p. 838, 2019.

[49] A. Patmonoaji, K. Tsuji, and T. Suekane, "Pore-throat characterization of unconsolidated porous media using watershed-segmentation algorithm," *Powder Technology,* vol. 362, pp. 635-644, 2020.

[50] S. R. Biswal, T. Sahoo, and S. Sahoo, "Prediction of grain boundary of a composite microstructure using digital image processing: a comparative study," *Materials Today: Proceedings,* vol. 41, pp. 357-362, 2021.

[51] R. Cohn, I. Anderson, T. Prost, J. Tiarks, E. White, and E. Holm, "Instance segmentation for direct measurements of satellites in metal powders and automated microstructural characterization from image data," *JOM,* vol. 73, no. 7, pp. 2159-2172, 2021.

[52] K. He, G. Gkioxari, P. Dollár, and R. Girshick, "Mask r-cnn," in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 2961-2969.

Journal of Soft Computing and Artificial Intelligence

Journal homepage: https://dergipark.org.tr/en/pub/jscai

International Open Access

Volume 04
Issue 01

June, 2023

Research Article

Classification of News Texts from Different Languages with Machine Learning Algorithms

Sidar Ağduk , Emrah Aydemir[2] , Ayfer Polat[3]

[1] *Management Information Systems, Faculty of Economics and Administrative Sciences. Tarsus University, 33000, Tarsus, Turkey*
[2] *Management Information Systems, Institute of Business. Sakarya University, 54000, Sakarya, Turkey*
[3] *SAP Applications Business Analyst, BIZE Project Development Inc., 16000, Bursa, Turkey*

ARTICLE INFO

Article history:
Received **June 8, 2023**
Revised **June 12, 2023**
Accepted **June 19, 2023**

Keywords:
Text Mining
Text Classification
Machine Learning
Weka
Naive Bayes

ABSTRACT

As a result of the developments in technology, the internet is accepted as one of the most important sources of information today. Although it is possible to access a large number of data in a short time thanks to the Internet, it is critical to analyze this data correctly. The need for text mining is increasing day by day by processing and analyzing the increasingly irregular text type data in the digital environment and classifying them in a meaningful way. In this study, news texts obtained from online German, Spanish, English and Turkish news sites were separated according to predetermined world, sports, economy and politics categories. The data set consisting of 4000 news texts was classified using 41 different machine learning algorithms in the Weka program. The highest successful classification was obtained with Naive Bayes Multinominal and Naive Bayes Multinominal Updateable algorithms, and 93.5% for German news texts, 93.3% for English news texts, 82.8% for Spanish news texts and 88.8% for Turkish news texts.

1. Introduction

Thanks to the advancements in internet and information technologies, access to information has become significantly easier [1, 2]. In particular, the increasing use of the internet has led to the vast expansion of accessible data [3]. Therefore, in the age of information and technology we find ourselves in, it is crucial to be able to quickly access the desired accurate data [4]. Data mining techniques, which vary depending on the type of data stack, are used to extract meaningful information, process, and analyze the complex array of data found online [5]. Text mining, one of the types of data mining, is used to extract meaningful information from textually stored data such as emails, web pages, reports, articles, and official documents [6]. Text mining, resulting from the combined use of natural language processing and data mining techniques, uncovers the hidden meanings in textual data stacks with the help of computer systems [7]. In text mining applications, textual data is classified and categorized using natural language processing or data mining methods, and a model is created. Text mining performs prediction when encountering a new text that is not included in the dataset, based on the established model [8]. Text classification, used for the purpose of classification in text mining, enables the assignment of categories to newly encountered texts from existing categories [9].

The rest of the study is organized as follows: the second section provides information about text

[1] Corresponding author
e-mail: sidaragduk@tarsus.edu.tr
DOI: 10.55195/jscai.1311380

classification and the third section presents the studies on text classification found in the literature. The fourth section discusses the purpose of the study, the dataset and preprocessing stages, feature extraction, performance evaluation criteria, and the classification methods used, under the heading of methodology. The fifth section presents the findings of the study, while the final section includes evaluations of the results.

2. Text Classification

Due to the advancements in information and communication technologies, the number of documents created in the online environment has been increasing every day [10]. While the increase in accessible information brings many benefits, it also presents some challenges [2]. The classification of texts found in the online environment is among these challenges. Simply put, text classification is the process of determining to which previously defined category or categories a given text data belongs. In other words, text classification involves determining whether the textual data in set $B=\{b_1,b_2,...,b_n\}$ belongs to the classes in set $S=\{s_1,s_2,...,s_m\}$ that have been predetermined. Therefore, it is necessary to generate a value, true or false, for $(b_j, s_i) \in B \times S$. A function g can be represented as $g : D \times C \rightarrow \{true, false\}$, where g produces the actual results, i.e., true if the j_{th} document belongs to the i_{th} class, and false otherwise. Accordingly, a similar function f that operates in a similar manner can be created using machine learning methods, represented as $f : D \times C \rightarrow \{true, false\}$. The aim is for the results produced by the generated *f* function to be as similar as possible to the results of the *g* function. A model is created using machine learning methods, and an *f* function (classifier) that operates similarly to the g function is implemented. Finally, the similarity between the *f* function and the actual results, represented by *g*, is compared [11].

With the advancements in technology and the increasing use of the internet, there is a growing need for data analysis and categorization [1]. News agencies, one of the most important sources of information today, have incorporated the online environment into their publishing activities as a result of technological developments. Proper classification, labeling, and presentation of the content offered to readers are of critical importance in enabling access to accurate news texts [12, 13]. Text classification, which involves automatically separating documents into specific semantic categories, effectively utilizes machine learning techniques. Documents consisting of textual data can be uncategorized or composed of content belonging to one or more categories. In order to classify texts automatically using machine learning, textual data needs to be transformed into numerical form using various approaches. TF-IDF, Word2Vec, and FastText methods are among the approaches used to extract vector models of texts [10]. Upon reviewing the conducted studies, it can be observed that various classification algorithms such as Random Forest [1, 2, 4, 10, 14-16], K-Nearest Neighbor [12, 16-18], Naive Bayes [2, 4, 10, 12, 14-16], Support Vector Machines [4, 10, 12, 15, 16, 19], [15], C4.5 [12, 14], Artificial Neural Networks [10, 20-25], and Logistic Regression [10] are utilized in the classification of textual data.

3. Relevant Literature

Aydemir et al. classified 2248 news texts from a Turkish-language news website into eight different categories using Multinomial Naive Bayes Algorithm (MNBA) and Random Forest (RF) algorithms based on predefined news categories. The study achieved a classification accuracy rate of 95.24% with MNBA and 99.86% with RF algorithm [2]. Başkaya and Aydın classified a total of 80 news texts, consisting of four different categories and 20 news texts for each category, from different news websites and newspapers using Naive Bayes (NB), J48 Decision Trees, Support Vector Machine (SVM), and RF. The highest successful result was achieved with the Random Forest algorithm, with a success rate of 100% in all four classification types [1]. Uslu and Akyol performed text classification using 4900 Turkish news texts. The news texts consisted of seven different categories, with 700 news texts in each category. SVM, RF, and NB algorithms were used for the classification of Turkish news text contents in the study. The analysis results showed a successful classification rate of 89% with SVM, 87% with RF, and 91% with NB [4]. Acı and Çırak used the widely used Turkish Text Classification 3600 dataset for the classification of Turkish news contents. The dataset consists of 3600 news data, with 600 news texts in each of the six different categories. Convolutional Neural Networks and Word2Vec method were used for the text classification process, resulting in a success rate of 93.3% [3]. Çelik and Koç performed text classification on a dataset of 12,000 data samples from different Turkish news sources belonging to six different categories. The news texts, vectorized using Tfidfvectorizer, Word2Vec, and FastText methods, were then classified using DVM, NB, LR, RF, and ANN methods. The study achieved a highest success rate of 95.75%, obtained by classifying FastText vectorized news texts with DVM [10]. Şimşek and Aydemir classified 1017 emails obtained from 20 different Gmail and Hotmail accounts as spam or legitimate emails using 45 different classification algorithms. The study yielded the highest accurate classification rate of 94.78% with Naive Bayes Multinomial and Naive Bayes Multinomial Updateable algorithms [26]. Table 1 provides information about the studies included in the literature related to text classification.

Table 1 Related Literature

Study & Author	Data Size	Classification Method	Success Rate (%)
[1] Başkaya & Aydın (2017)	80	Naive Bayes Support Vector Machine C4.5 Algorithm Random Forest	90 95 65 100
[2] Aydemir et al. (2021)	2248	Multinomial Naive Bayes Random Forest	95.24 99.6
[3] Acı & Çırak (2019)	3600	Convolutional Neural Networks	93.3
[4] Uslu & Akyol (2019)	4900	Support Vector Machine Random Forest Naive Bayes	89 87 91
[15] Cusmuliuc et al. (2018)	10000	Naive Bayes Support Vector Machine Random Forest	92.43 95 95.93
[17] Aşlıyan & Günel (2010)	250	k-Nearest Neighbors	76.8
[27] Dilrukshi et al. (2013)	3569	Support Vector Machine	75
[28] Deniz et al. (2019)	799	Logistic Regression Naive Bayes Decision Tree Random Forest Support Vector Machine k-Nearest Neighbors	73.12 78.12 59.37 60.62 78.75 78.12
[29] Sel et al. (2019)	18878	MaxEnt Classification	94.54
[30] Jehad & Yousif (2020)	20800	C4.5 Algorithm Random Forest	89.11 84.97
[31]	4964	Support Vector Machine Artificial Neural Network	74.62 72.99
Shahi & Pant (2018)		Naive Bayes	68.31

4. Method

4.1. Research Objective

Despite the significant advancements in technology that bring great convenience to our lives, they also come with certain drawbacks. Particularly, the widespread use of the internet as the primary tool for accessing information leads to the generation of a large volume of data in real-time in the online environment. News agencies, being one of the most important sources of information, have also incorporated the online platform into their publishing activities due to these technological developments. In the face of increasing data on the internet, proper classification, labeling, and presentation of content to readers have become critically important for ensuring access to accurate news articles [12, 13]. In line with this, this study aims to successfully classify German, Turkish, Spanish, and English news texts into predefined categories such as world, economy, politics, and sports using various classification algorithms. The flowchart of the study is presented in Figure 1.

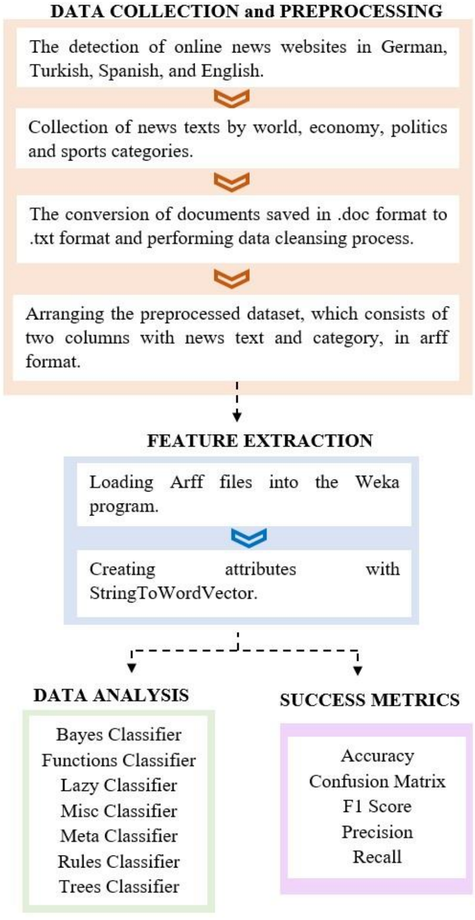

DATA COLLECTION and PREPROCESSING
The detection of online news websites in German, Turkish, Spanish, and English.
Collection of news texts by world, economy, politics and sports categories.
The conversion of documents saved in .doc format to .txt format and performing data cleansing process.
Arranging the preprocessed dataset, which consists of two columns with news text and category, in arff format.
FEATURE EXTRACTION
Loading Arff files into the Weka program.
Creating attributes with StringToWordVector.
DATA ANALYSIS
Bayes Classifier
Functions Classifier
Lazy Classifier
Misc Classifier
Meta Classifier
Rules Classifier
Trees Classifier
SUCCESS METRICS
Accuracy
Confusion Matrix
F1 Score
Precision
Recall

Figure 1 Study Flowchart

4.2. Data Set

In the study, a total of 4,000 news articles were obtained from online news websites publishing in German, English, Spanish, and Turkish languages, covering the categories of world, economy, politics, and sports. Each category consists of 250 news articles. The dataset used in the study has been publicly published on Kaggle [33]. Detailed information about the dataset used in the study is presented in Table 2.

Table 2 Data Set

Language of News	**Category Type**			
	World	***Economy***	***Politics***	***Sports***
German	250	250	250	250
Spanish	250	250	250	250
English	250	250	250	250
Turkish	250	250	250	250
Total	1000	1000	1000	1000

4.3. Feature Extraction

In order for machine learning classification algorithms to understand the dataset consisting of news texts, the texts need to be converted into numerical format. For this purpose, the StringToWordVector filter in the Weka program is used, which employs techniques such as TF-IDF and n-grams to transform the texts into numerical vectors [26]. Firstly, in the study, the "RegExpFromFile" command available in the Weka program is selected to determine whether a word is a stopword or not. In this step, the "WordTokenizer" command is also chosen to tokenize the words for the vectorization process. The preprocessed .arff format news data, which have gone through various preprocessing stages, are loaded into the Weka program, and then the attributes are extracted using the "StringToWordVector" filter. The "StringToWordVector" filter, which covers all words, generates attributes in numerical values indicating the frequencies of the words [32]. The vectorized form of the word frequencies is used in the classification phase. In the study, 2257 word vectors are extracted as features for the English news dataset, 2088 for the Spanish news dataset, 2257 for the German news dataset, and a total of 2572 for the Turkish news dataset. The parameter provided for the "StringToWordVector" function is as follows:

- weka.filters.unsupervised.attribute.StringToWordVector -R first -W 1000 -prune-rate -1.0 -N 0 -stemmer weka.core.stemmers.NullStemmer -stopwords-handler "weka.core.stopwords.RegExpFromFile -stopwords \"C:\\\\Program Files\\\\Weka-3-8-6\"" -M 1 -tokenizer "weka.core.tokenizers.WordTokenizer -delimiters \" \\r\\n\\t.,;:\\\'\\\"()?!\""

4.4. Classification Method

The data consisting of news texts were analyzed using the Weka program. The Weka program, which takes its name from the initials of "Waikato Environment for Knowledge Analysis," was developed at Waikato University in New Zealand. This program, which is free, open-source, and Java-based, enables various operations such as Classification, Clustering, Association, Data Preprocessing, and Visualization. The program includes commonly used machine learning algorithms [32]. In this study, 41 different classification methods belonging to the Bayes Classifiers, Tree Algorithms, Rule-Based Classifiers, Function Classifiers, Lazy Algorithms, Various Classifiers, and Meta-Learning Algorithms were used in the Classify tab of the Weka program for the analysis of the news data. Before the classification process, the dataset needs to be split into training and test sets. The main goal of machine learning algorithms is to generate models that make accurate predictions on the separated training dataset and evaluate the accuracy of the model on new data. The data used to test the accuracy of the model constitute the test dataset. The simplest approach used to split the dataset for training and testing is to randomly assign a percentage, for example, 80% for training and 20% for testing. Splitting the data percentage-wise may introduce some errors in determining the training and test data based on data distribution. To overcome this issue, the cross-validation method was used to split all data into training and test sets within themselves. With this method, the data is initially divided into 10 separate groups, and one group is used for testing while the remaining nine groups are used for training, repeated 10 times. Then, the average of classification performances in each iteration is calculated to obtain the final success rate. This process is visually explained in Figure 2 below.

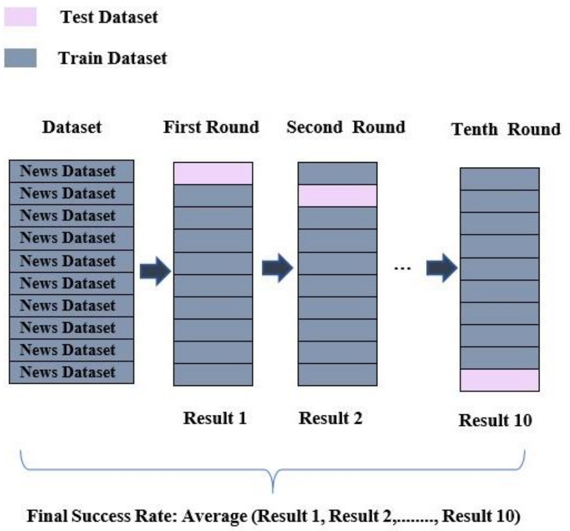


Figure 2 K-Fold Cross Validation

4.5. Performance Measures

In machine learning, the "Confusion Matrix," also known as the "Error Matrix," is used to compare the predicted and true values and interpret the performance of classification models. This matrix provides information about the correct or incorrect placement of test data into classes [32]. Along with the confusion matrix presented in Table 3 below, the following performance measures are obtained:

- Accuracy
- Recall
- Precision
- F1 score

Table 3 Confusion Matrix

		True Value	
		True	**False**
Prediction Value	**True**	True Positive (TP)	False Positive (FP)
	False	False Negative (FN)	True Negative (TN)

The definitions related to the confusion matrix in Table 3 are provided below:

- True Positive (TP): The instances that are correctly predicted as positive when the true value is positive.
- False Negative (FN): The instances that are incorrectly predicted as negative when the true value is positive.
- False Positive (FP): The instances that are incorrectly predicted as positive when the true value is negative.
- True Negative (TN): The instances that are correctly predicted as negative when the true value is negative.

In addition to these categorical values, precision (1), recall (2), F-score (3), and accuracy rate (4) are used when predicting categorical values. These values are calculated using the formulas provided below.

$$\frac{TP}{TP + FP} \qquad (1)$$

$$\frac{TP}{TP + FN} \qquad (2)$$

$$2 * \frac{Precision * Recall}{Precision + Recall} \qquad (3)$$

$$\frac{TP + TN}{TP + TN + FP + FN} \qquad (4)$$

5. Results

The features of the dataset consisting of news texts in different languages were determined using the "StringToWordVector" function in the Preprocess tab of the Weka program. The news data with extracted features were then tested with 41 different machine learning algorithms using the widely accepted 10-fold cross-validation method in the Classify tab, and the findings are presented in the following tables.

Table 4 Confusion Matrix

	Algorithm	**Success Rate by Languages (%)**			
		German	English	Spanish	Turkish
BAYES	Bayes Net	85.2	85.1	73.3	84.2
	Naive Bayes	86.7	86.7	72.7	84
	Naive Bayes Multinominal	93.5	93.3	82.8	88.8
	Naive Bayes Multinominal Text	25	25	25	25
	Naive Bayes Multinominal Updateable	93.5	93.3	82.8	88.8
	Naive Bayes Updateable	86.7	86.7	72.7	84
TREE	Decision Stump	38.6	38.6	39.5	33.9
	Hoeffding Tree	25	25	25	25
	J48	80.6	80.7	66.8	77.2
	LMT	90.8	90.9	78.6	86.8
	Random Forest	88.3	88.8	77.9	87.3
	Random Tree	64.6	64.6	50.3	59.2
	REP Tree	76.2	76.7	64.5	73.3
RULES	Decision Table	71.8	71.8	65.6	63.8
	JRip	78.1	79.5	65.9	72.5
	OneR	40.3	40.3	41.8	36.4
	PART	82.5	82.5	65.6	77.8
	ZeroR	25	25	25	25
	Simple Logistic	90.8	90.9	78.9	86.6

	SMO	93.3	93.2	80.5	87.4
LAZY	IBk	68	68.1	49.1	64.7
	KStar	69.4	69.3	51.3	66.5
	LWL	41.7	41.7	41.9	40.3
MISC	Input Mapped Classifier	25	25	25	25
META	AdaBoostM1	38.6	38.6	39.5	33.9
	Attribute Selected Classifier	81.1	80.8	68.3	80.5
	Bagging	83.3	83.9	73.4	76.9
	Classication Via Regression	80.9	80.1	67.8	66.5
	CV Parameter Selection	25	25	25	25
	Filtered Classifier	82.7	82.5	68.2	82.2
	Iterative Classifier Optimizer	86.6	86.2	73.2	82.2
	Logit Boost	86.6	86.2	73.1	82.2
	Multi Class Classifier	87.1	87.4	50.7	77.7
	Multi Class Classifier Updateable	92.4	92.4	76.5	85.7
	Random Committee	82	83.6	68.8	81.9
	Randomizable Filtered Classifier	42.3	42.3	38.4	35.3
	Random Sub Space	84.6	84.7	74.5	82.7
	Stacking	25	25	25	25
	Vote	25	25	25	25
	Weighted Instances Handler Wrapper	25	25	25	25
	Multi Scheme	25	25	25	25

When examining Table 4, it is observed that the highest classification results for German, English, Spanish, and Turkish news data belong to the Naive Bayes Multinomial and Naive Bayes Multinomial Updateable classifiers. The analysis results for these algorithms regarding German, English, Spanish, and Turkish languages, including values such as True Positive Rate (TP), False Positive Rate (FP), and F-Score, are presented in Table 5. The confusion matrices for each language are shown in Figures 3, 4, 5, and 6, respectively.

Table 5 Other performance metrics for the top classification algorithms

News Dataset	Algorithm	Accuracy Rate (%)	Precision	Recall	F-Score	TP	FP
German	Naive Bayes Multinominal	93.5	0.937	0.935	0.935	0.935	0.022
	Naive Bayes Multinominal Updateable	93.5	0.937	0.935	0.935	0.935	0.022
English	Naive Bayes Multinominal	93.3	0.935	0.933	0.933	0.933	0.022
	Naive Bayes Multinominal Updateable	93.3	0.935	0.933	0.933	0.933	0.022
Spanish	Naive Bayes Multinominal	82.8	0.830	0.828	0.829	0.828	0.057
	Naive Bayes Multinominal Updateable	82.8	0.830	0.828	0.829	0.828	0.057
Turkish	Naive Bayes Multinominal	88.8	0.890	0.888	0.889	0.888	0.037
	Naive Bayes Multinominal Updateable	88.8	0.890	0.888	0.889	0.888	0.037

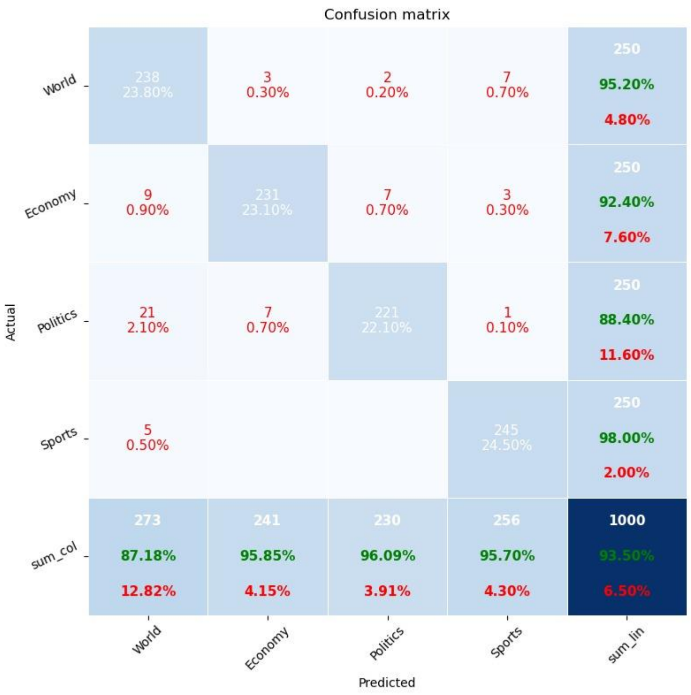


Figure 3 Confusion Matrix for German News Dataset

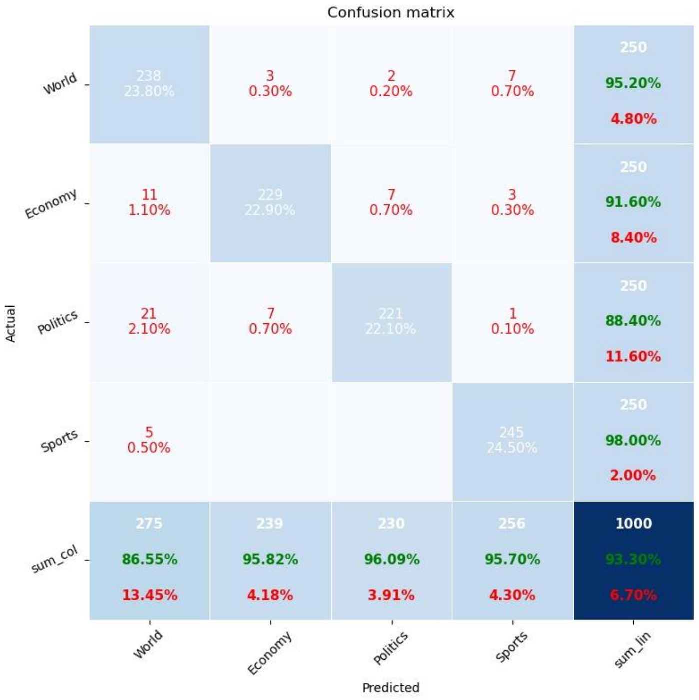


Figure 4 Confusion Matrix for English News Dataset

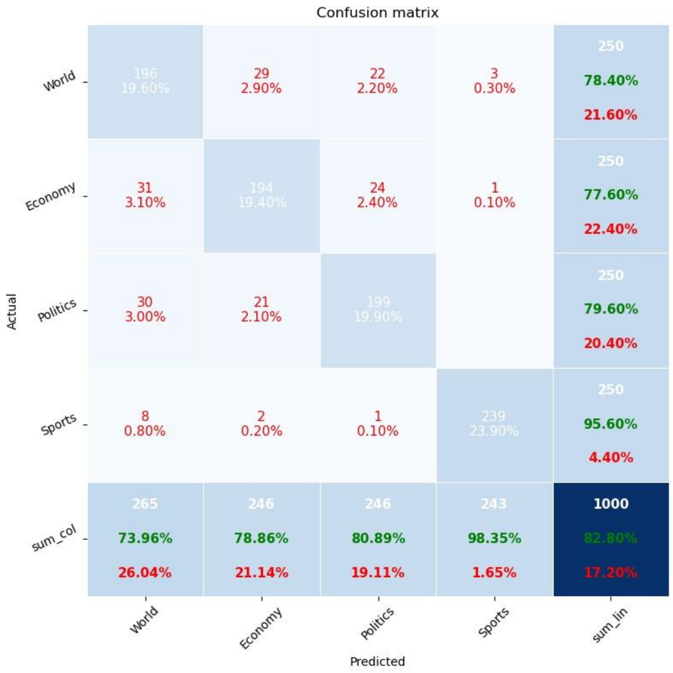


Figure 5 Confusion Matrix for Spanish News Dataset

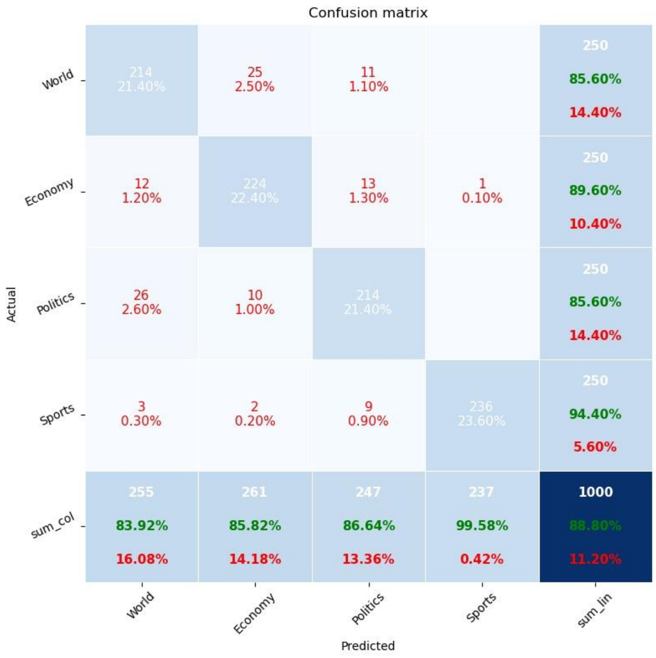


Figure 6 Confusion Matrix for Turkish News Dataset

6. Discussion and Conclusion

With the rapid increase of a large amount of textual data, particularly news articles, in online platforms and other sources, it has become increasingly important to effectively analyze and comprehend this data. Text classification of news articles serves as a fundamental step in categorizing and extracting meaningful information from these data. It assists many individuals in understanding news articles within large datasets, identifying trends, and making informed decisions. Therefore, accurate classification of news articles facilitates easy access to information, saves time, and plays an effective role in information management.

Aydemir et al. classified 2248 news texts from a

Turkish-language news website into eight different categories using Multinomial Naive Bayes Algorithm (MNBA) and Random Forest (RF) algorithms based on predefined news categories. The study achieved a classification accuracy rate of 95.24% with MNBA and 99.86% with RF algorithm [2]. Uslu and Akyol performed text classification using 4900 Turkish news texts. The news texts consisted of seven different categories, with 700 news texts in each category. SVM, RF, and NB algorithms were used for the classification of Turkish news text contents in the study. The analysis results showed a successful classification rate of 89% with SVM, 87% with RF, and 91% with NB [4].

In this study, German, English, Spanish, and Turkish news texts were classified according to the categories of world, economy, politics, and sports. A dataset consisting of 4000 news texts was tested using 41 classification algorithms in the Weka program. As a result, the highest classification performance was achieved with the Naive Bayes Multinomial and Naive Bayes Multinomial Updateable algorithms, belonging to the Bayes classifier, for all news texts. The success rates were determined as 93.5% for German news texts, 93.3% for English news texts, 82.8% for Spanish news texts, and 88.8% for Turkish news texts. Additionally, among other successful classification algorithms, it was observed that the SMO algorithm of the Functions classifier and the Multi Class Classifier Updateable algorithm of the Meta classifier were prominent. For the SMO algorithm, success rates of 93.3%, 93.2%, 80.5%, and 87.4% were obtained for German, Spanish, English, and Turkish news texts, respectively. For the Multi Class Classifier Updateable algorithm, success rates of 92.4%, 92.4%, 76.5%, and 85.7% were obtained for the same languages. Finally, it was determined that the Naive Bayes Multinomial Text, Hoeffding Tree, ZeroR, Input Mapped Classifier, CV Parameter Selection, Stacking, Vote, Weighted Instances Handler Wrapper, and Multi Scheme algorithms had the lowest success rates, indicating that the news texts were classified only into one category. In conclusion, this study demonstrates that Naive Bayes Multinomial and Naive Bayes Multinomial Updateable algorithms achieve high success rates when compared to similar studies in the literature. Furthermore, considering the success rates of other classification algorithms, it can be said that this study makes a significant contribution in terms of classification performance.

References

[1]. Başkaya, F., & Aydın, İ. Haber Metinlerinin Farklı Metin Madenciliği Yöntemleriyle Sınıflandırılması, In 2017 International Artificial Intelligence and Data Processing Symposium (IDAP), 2017, pp. 1-5. IEEE.

[2]. Aydemir, E. , Işık, M. & Tuncer, T. Türkçe Haber Metinlerinin Çok Terimli Naive Bayes Algoritması Kullanılarak Sınıflandırılması, Fırat Üniversitesi Mühendislik Bilimleri Dergisi, 2021, 33(2), pp. 519-526. doi: 10.35234/fumbd.871986

[3]. Acı, Ç. & Çırak, A. Türkçe Haber Metinlerinin Konvolüsyonel Sinir Ağları ve Word2Vec Kullanılarak Sınıflandırılması, Bilişim Teknolojileri Dergisi, 2019, 12(3), pp. 219-228. doi: 10.17671/gazibtd.457917.

[4]. Uslu, O., & Akyol, S. Türkçe Haber Metinlerinin Makine Öğrenmesi Yöntemleri Kullanılarak Sınıflandırılması, ESTUDAM Bilişim Dergisi, 2019, 2(1), pp. 15-20.

[5]. Doğan, K., & Arslantekin, S. Büyük Veri: Önemi, Yapısı Ve Günümüzdeki Durum, Ankara Üniversitesi Dil ve Tarih-Coğrafya Fakültesi Dergisi, 2016, 56(1), pp.15-36.

[6]. Bach, M. P., Krstić, Ž., Seljan, S., & Turulja, L. Text mining for big data analysis in financial sector: A literature review, Sustainability, 2019, 11(5), pp. 1-27.

[7]. Tan, A. H. Text mining: The state of the art and the challenges, In Proceedings of the pakdd 1999 workshop on knowledge disocovery from advanced databases, 1999, pp. 65-70.

[8]. Coşkun, C., & Baykal, A. Veri Madenciliğinde Sınıflandırma Algoritmalarının Bir Örnek Üzerinde Karşılaştırılması. Akademik Bilişim, 2011, 11, pp. 51-58.

[9]. Dalal, M. K., & Zaveri, M. A. Automatic Text Classification: A Technical Review, International Journal of Computer Applications, 2011, 28(2), pp. 37-40.

[10]. Çelik, Ö., & Koç, B. C. TF-IDF, Word2vec ve Fasttext Vektör Model Yöntemleri ile Türkçe Haber Metinlerinin Sınıflandırılması, Dokuz Eylül Üniversitesi Mühendislik Fakültesi Fen ve Mühendislik Dergisi, 2021, 23(67), pp. 121-127.

[11]. Tantuğ, A. C. Metin Sınıflandırma, Türkiye Bilişim Vakfı Bilgisayar Bilimleri ve Mühendisliği Dergisi, 2016, 5(2).

[12]. Toraman, C., Can, F., & Koçberber, S. Developing A Text Categorization Template For Turkish News

Portals, In 2011 International Symposium on Innovations in Intelligent Systems and Applications, 2011, pp. 379-383. IEEE.

[13]. Yıldırım, S., & Yıldız, T. Türkçe İçin Karşılaştırmalı Metin Sınıflandırma Analizi, Pamukkale Üniversitesi Mühendislik Bilimleri Dergisi, 2018, 24(5), pp. 879-886.

[14]. Amasyalı, M. F., Diri, B., & Türkoğlu, F. Farklı Özellik Vektörleri İle Türkçe Dokümanların Yazarlarının Belirlenmesi, In The Fifteenth Turkish Symposium on Artificial Intelligence and Neural Networks (TAINN'2006), 2006, pp. 4.

[15]. Cusmuliuc, C. G., Coca, L. G. and Iftene, A. Identifying Fake News on Twitter using Naive Bayes, SVM and Random Forest Distributed Algorithms, In Proceedings of The 13th Edition of the International Conference on Linguistic Resources and Tools for Processing Romanian Language, 2018, pp.177-188.

[16]. Doğan, S., & Diri, B. Türkçe Dokümanlar için N-Gram Tabanlı Yeni Bir Sınıflandırma (Ng-İnd): Yazar, Tür ve Cinsiyet, Türkiye Bilişim Vakfı Bilgisayar Bilimleri ve Mühendisliği Dergisi, 2010, 3(1), pp. 11-19.

[17]. Aşlıyan, R., & Günel, K. Metin İçerikli Türkçe Dokümanların Sınıflandırılması, Akademik Bilişim Konferansı, 2010, pp. 659-665.

[18]. Soucy, P., & Mineau, G. W. A Simple KNN Algorithm For Text Categorization, In Proceedings 2001 IEEE international conference on data mining, 2001, pp. 647-648. IEEE.

[19]. Joachims, T. Text Categorization With Support Vector Machines: Learning With Many Relevant Features, In European conference on machine learning, 1998, pp. 137-142.

[20]. Ma, L., Shepherd, J., & Zhang, Y. Enhancing Text Classification Using Synopses Extraction, In Proceedings of the Fourth International Conference on Web Information Systems Engineering, 2003, pp. 115-124. IEEE.

[21]. Lam, S. L., & Lee, D. L. Feature Reduction For Neural Network Based Text Categorization, In Proceedings. 6th international conference on advanced systems for advanced applications, 1999, pp. 195-202. IEEE.

[22]. Ng, H. T., Goh, W. B., & Low, K. L. Feature Selection, Perceptron Learning, And A Usability Case Study For Text Categorization, In Proceedings Of The 20th Annual International ACM SIGIR Conference On Research And Development In Information Retrieval, 1997, pp. 67-73.

[23]. Nakayama, M., & Shimizu, Y. Subject Categorization for Web Educational Resources using MLP, In ESANN, 2003, pp. 9-14.

[24]. Srinivasan, P., & Ruiz, M. E. Automatic Text Categorization Using Neural Network, In Proceedings of the 8th ASIS SIG/CR Workshop on Classification Research, 1998, pp. 59-72.

[25]. Ma, S., & Ji, C. A Unified Approach on Fast Training of Feedforward and Recurrent Networks Using EM Algorithm, IEEE transactions on signal processing, 1998, 46(8), pp. 2270-2274. IEEE.

[26]. Şimşek, H. & Aydemir, E. Classification of Unwanted E-Mails (Spam) with Turkish Text by Different Algorithms in Weka Program, Journal of Soft Computing and Artificial Intelligence, 2022, 3(1) , pp. 1-10. doi: 10.55195/jscai.1104694

[27]. Dilrukshi, I., De Zoysa, K., & Caldera, A. Twitter News Classification Using SVM, In 2013 8th International Conference on Computer Science & Education, 2013, pp. 287-291. IEEE.

[28]. Deniz, E., Erbay, H., & Coşar, M. Classification Of Turkish E-Mails With Doc2Vec, In 2019 1st International Informatics and Software Engineering Conference (UBMYK), 2019, pp. 1-4. IEEE.

[29]. Sel, İ., Karci, A., & Hanbay, D. Feature Selection for Text Classification Using Mutual Information, In 2019 International Artificial Intelligence and Data Processing Symposium (IDAP), 2019, pp. 1-4. IEEE.

[30]. Jehad, R., & Yousif, S. A. Fake News Classification Using Random Forest and Decision Tree (J48), Al-Nahrain Journal of Science, 2020, 23(4), pp. 49-55.

[31]. Shahi, T. B., & Pant, A. K. Nepali News Classification Using Naïve Bayes, Support Vector Machines and Neural Networks, In 2018 International Conference on Communication Information and Computing Technology (ICCICT), 2018, pp. 1-5. IEEE.

[32]. Aydemir, E. Weka İle Yapay Zeka. Seçkin Yayınevi, 2018, Ankara.

[33]. Ağduk, S., Aydemir, E. & Polat, A. (2022). News Texts by Category in Different Languages [Data set]. Kaggle. https://doi.org/10.34740/KAGGLE/DSV/3572093

Journal of Soft Computing and Artificial Intelligence

ISSN: 2717-8226

Journal of Soft Computing and Artificial Intelligence

Journal homepage: https://dergipark.org.tr/en/pub/jscai

International Open Access

Volume 04
Issue 01

June, 2023

Research Article

LabVIEW-based fire extinguisher model based on acoustic airflow vibrations

Mahmut Dirik [1]

[1] *Computer Engineering Department, Sirnak University, Şırnak, 73000, Türkiye*

ARTICLE INFO

Article history:
Received **June 6, 2023**
Revised **June 21, 2023**
Accepted **June 24, 2023**

Keywords:
Flame Classification
Fire Safety
Machine Learning
Sound Wave
Fire Extinguishing System

ABSTRACT

In recent years, soundwave-based fire extinguishing systems have emerged as a promising avenue for fire safety measures. Despite this potential, the challenge is to determine the exact operating parameters for efficient performance. To address this gap, we present an artificial intelligence (AI)-enhanced decision support model that aims to improve the effectiveness of soundwave-based fire suppression systems. Our model uses advanced machine learning methods, including artificial neural networks, support vector machines (SVM) and logistic regression, to classify the extinguishing and non-extinguishing states of a flame. The classification is influenced by several input parameters, including the type of fuel, the size of the flame, the decibel level, the frequency, the airflow, and the distance to the flame. Our AI model was developed and implemented in LabVIEW for practical use.

The performance of these machine learning models was thoroughly evaluated using key performance metrics: Accuracy, Precision, Recognition and F1 Score. The results show a superior classification accuracy of 90.893% for the artificial neural network model, closely followed by the logistic regression and SVM models with 86.836% and 86.728% accuracy, respectively. With this study, we highlight the potential of AI in optimizing acoustic fire suppression systems and offer valuable insights for future development and implementation. These insights could lead to a more efficient and effective use of acoustic fire extinguishing systems, potentially revolutionizing the practice of fire safety management.

1. Introduction

Fire-related disasters, both natural and human-induced, pose significant threats to life, property, and the environment. Thus, the development of effective preventive measures is crucial in mitigating these risks [1] [2]. Conventional firefighting methods, which often entail the use of chemicals or heavy equipment, might inadvertently inflict further harm on infrastructure, natural resources, or the residents of the affected area [3], [4]. Hence, understanding the specific characteristics of the fire and the burning materials is of paramount importance for identifying the most suitable extinguishing technique [5]–[7].

In this context, the potential of sound waves as a means of fire suppression has garnered significant attention. This unique method could not only ensure the safety of people and the environment, but also present a cost-effective, environmentally friendly option [8], [9]. Sound wave-based fire extinguishing systems generate pressure waves that disrupt the combustion process and extinguish the fire, offering a safe, non-toxic, and non-caustic solution [10]. However, this technology is still in the research and development phase, necessitating further studies to optimize its effectiveness and efficiency.

Prior research indicates that both low-frequency sound waves (30 Hz to 50 Hz) and high-frequency

[1] Corresponding author
e-mail: *mhmd.dirik@gmail.com*
DOI: 10.55195/jscai.1310837

sound waves (60 Hz to 90 Hz) can effectively disrupt combustion [11]–[17]. The efficacy of these acoustic waves depends on several factors including wave amplitude and source distance, both of which significantly influence flame dynamics. Moreover, other variables such as sound frequency, atmospheric conditions, and flame properties also play crucial roles [18], [19].

Considering the complexity of fire dynamics, there is a need for effective fire detection and suppression systems that can balance sensitivity, reliability, extinguishing efficiency, safety, and cost-effectiveness [20]–[23]. In this vein, the application of machine learning techniques and statistical analysis to study the characteristics of sound waves produced by flames offers promising advancements [24]. Coupled with the use of various sensors, cameras, and thermal imagers, data-driven approaches can provide a comprehensive understanding of fire behavior [25]–[30]. Such understanding can, in turn, contribute to the improvement of acoustic fire extinguishing systems [31]–[33].

In this study, we utilize a dataset comprised of 17,442 samples from experimental studies [34]–[37]. Our approach distinguishes itself from previous works through its innovative user interface and dynamic system. We propose a new model using LabVIEW, employing machine learning algorithms such as artificial neural networks, support vector machines, and logistic regression to predict flame extinguishing outcomes based on variables like fuel type, flame size, decibel level, frequency, airflow, and distance. The aim is to provide an decision-support system for sound wave fire extinguishing [34]–[37].

The paper is organized as follows: Section 2 elucidates the dataset, classification algorithms, and performance metrics used in our study. Section 3 presents the experimental results. Finally, Section 4 provides the conclusions drawn from this study.

2. Material and methods

In this section, we explain our systematic methodology, which covers the stages of data collection up to the culmination of the analysis. The process of data collection, the technical specifications of the collected dataset, and its dissemination are described in detail.

For the task of distinguishing between the extinguishing and non-extinguishing states of a flame, we used classification methods including artificial neural networks, support vector machines (SVM), and logistic regression. The rationale for these selected techniques and their relevance to our study are presented.

The efficiency of the classifiers was evaluated by applying performance metrics, namely accuracy, precision, recall, and F1-score. These metrics are briefly explained to allow an unbiased comparison of the performance of the classifiers used.

2.1. Data Acquisition

This research study utilized a dataset, derived from references [34]–[37], encompassing data aggregated from tests conducted on a fire extinguisher using four distinct fuel flames. The system framework is comprised of four subwoofers, two amplifiers, a control unit, and a computer employed as frequency sources. Ancillary instruments such as an anemometer, a decibel meter, a camera, and an infrared thermometer were engaged in measuring various parameters throughout the extinguishing process.

An expansive total of 17,442 experimental trials were executed utilizing this specified experimental apparatus. These trials were conducted within a fire chamber, explicitly engineered to function in conjunction with a sound-wave fire extinguishing system. The aggregated data were subsequently utilized to construct models capable of predicting the output characteristic (extinguishing or not) predicated on six input characteristics.

During model development, it is of paramount importance to critically review fundamental statistical properties of the data, as they can provide indispensable insights into the data distribution and variability. This allows for the identification of potential data anomalies such as outliers or missing values, which may have an impact on model performance. Furthermore, examining the statistical measures of individual variables aids in ensuring data accuracy and consistency with the expected values for that respective variable.

Table 1 presents a succinct statistical summary for all variables encompassed in the dataset, including minimum, maximum, mean values, and standard deviations. It is important to note that certain variables, such as 'fuel', are categorical and hence do not possess a significant mean or standard deviation. The 'Minimum' and 'Maximum' columns for these variables signify the classes of least and most frequent categorical variables respectively. This information assists in understanding the distribution of the categorical variables within the dataset.

Figure 1 illustrates the distribution of the variable 'fuel' among four categories: Petrol, Thinner, Kerosene, and LPG. The category demonstrating the highest frequency is 'Petrol' with a prevalence of 29.418%. Conversely, the category 'LPG' exhibits the lowest frequency, standing at 11.7647%.

Table 1 Data statistics table

	Minimum	Maximum	Mean	Deviation
Size	1	7	3.41	1.75
Fuel	-	-	-	-
Distance	10	190	100	54.8
Decibel	72	113	96.4	8.16
Airflow	0	17	6.98	4.74
Frequency	1	75	31.6	20.9
Status	0	1	0.498	0.5

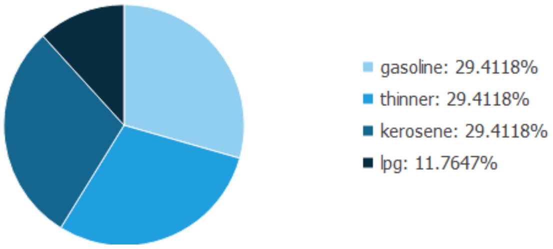


Figure 1 FUEL distribution pie chart

A correlation analysis serves as a robust empirical method for quantifying the dependencies that exist between constituent variables within a given data set. It is characterized by a numerical value, the correlation coefficient, which ranges from -1 to +1. A coefficient that tends towards the upper limit of +1 indicates a strong positive correlation. This means that an increase in one variable is usually accompanied by a corresponding increase in another. A correlation coefficient that approaches the lower limit of -1, on the other hand, indicates a strong negative correlation and signals an inverse relationship in which an increase in one variable generally triggers a decrease in the other. A correlation coefficient approaching zero, on the other hand, indicates that there is no or negligible linear correlation between the two variables under study.

In Table 2, each cell represents the calculated correlation coefficient, which makes a quantitative statement about the extent of the relationship between the corresponding pair of variables within the data set. This matrix highlights the inherent interdependence structure of the data set and promotes the formulation of insightful and rigorous inferential analyses.

Table 2 Inputs correlations

	Size	Fuel	Distance	Decibel	Airflow	Frequency
Size	1	0.431	-3.68e-11	6.8e-11	3.21e-11	6.49e-11
Fuel	0.431	1	0.176	0.176	0.176	0.176
Distance	-3.68e-11	0.176	1	-0.239	-0.707	-2.08e-15
Decibel	6.8e-11	0.176	-0.239	1	0.377	0.733
Airflow	3.21e-11	0.176	-0.707	0.377	1	-0.212
Frequency	6.49e-11	0.176	-2.08e-15	0.733	-0.212	1

A "feature trend" describes the course of the development of a certain variable over time. This progression is visually represented in Figure 2. The analysis of temporal data and the recognition of patterns facilitate the identification of trends, a crucial facet of comprehensive data analysis.

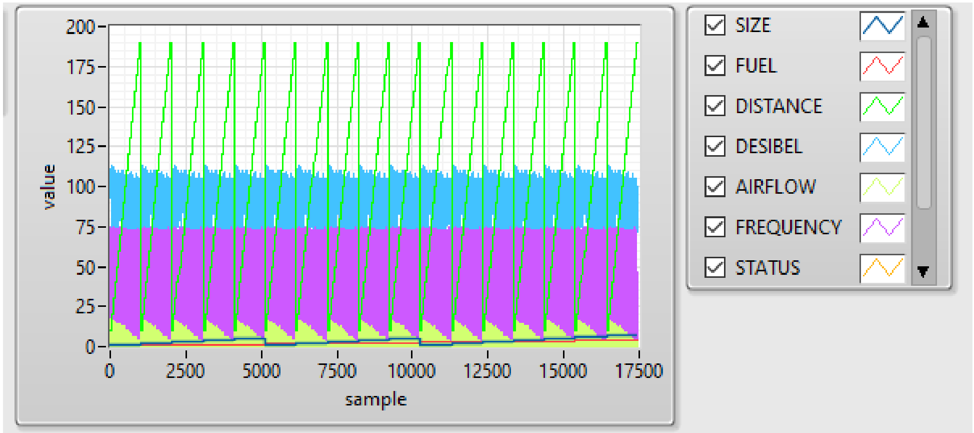


Figure 2 Feature trend

Numerous literary sources on the subject were consulted in the course of this study. These offer comprehensive insights into the processes of data collection and various other relevant scenarios. However, the focus of our work is on the development of a LabVIEW user interface and the evaluation of the effectiveness of learning algorithms.

Therefore, the specific machine learning classification techniques used in this study to analyze the above data are described in detail.

2.2. LabVIEW based Machine Learning Classifier

The software application called "LabVIEW-based Machine Learning Classifier" facilitates the creation of machine learning models in the LabVIEW programming environment. The graphically programmed interface allows users to quickly formulate and evaluate a variety of machine learning methods and algorithms. The program includes a number of pre-built machine learning classifiers that can be tailored to different scenarios, including classification, regression, and clustering. This section presents a model that uses the classification techniques of the developed LabVIEW-based machine learning classifier.

2.2.1. Artificial Neural Network

Artificial neural networks (ANNs), a sub-discipline of machine learning, draw inspiration from the structural and functional aspects of the human brain. The theoretical foundations for ANNs were laid in 1943 by McCulloch and Pitts [38], who constructed a mathematical model describing the neuronal activities of the brain. Subsequently, Hebb [39] proposed a mechanism of reinforcement-based learning to explain the learning processes of the human brain. Subsequently, Rosenblatt [40], [41] presented a computational model for the processing elements of the brain, which he called 'perceptrons', and thus provided the impetus for a thorough investigation of ANNs.

The aim of ANN's research is to develop machine learning systems based on a biological model of the brain, focusing in particular on the bioelectrical activity of the brain's neurons. This paves the way for the development of systems that are able to learn and adapt to new situations, much like the human brain. ANNs have applications in a variety of fields, including image recognition, natural language processing, speech recognition, and decision-making systems. For a graphical representation of the structure of an artificial neural network, see Figure 2.

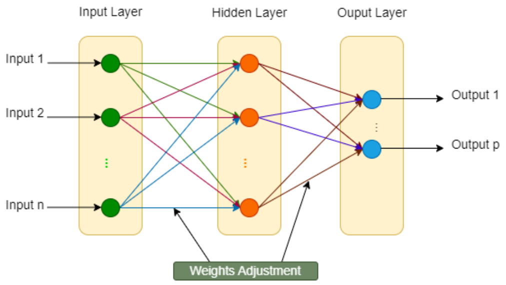


Figure 3 Basic structure of neural network.

Figure 3 contains the following integral parts:

Input layer: the first layer of the network that receives the input data and passes it on to the subsequent layer. The number of neurons in this layer corresponds to the number of features contained in the input data.

Hidden layers: These layers house the computations of the network. They consist of a collection of artificial neurons that process the input data and generate intermediate results. The number of neurons in each hidden layer, as well as the number of hidden layers themselves, can vary depending on the complexity of the problem under consideration.

Output layer: This is the last layer that produces the output of the network. It uses the intermediate results from the hidden layers and processes them further to produce the final output. The number of neurons in this layer reflects the number of classes inherent in the problem or the number of output features.

Weights: The connections between the layers, called 'weights', are critical to the learning process of the network. These weights are adjusted throughout the training phase of the network to optimize its performance and precision.

Activation function: This is a mathematical function applied to the output of each neuron that influences the final output of the neuron and consequently the output of the network.

Each layer hosts a large number of artificial neurons that process the input data to produce the final output. The architecture of the network, including the number of layers and neurons, as well as the activation function used, can be adjusted to achieve better results.

2.2.2. Support Vector Machines (SVM)

Support vector machines (SVMs) [42]–[44], a well-known category of machine learning algorithms, are mainly used for classification and regression tasks. SVMs can be roughly divided into three main categories: linear support vector machines, nonlinear SVMs, and multiclass SVMs. Linear SVMs are constructed in such a way that the instance groups of different classes separated by a hyperplane are equidistant, which allows for optimal delineation of the data. However, linear SVMs cannot handle datasets that are not linearly separable, necessitating the use of non-linear SVMs. Non-linear SVMs use kernel functions to classify non-linearly separable data. These kernel functions map the data to a higher-dimensional space and transform it into a linearly separable form. The resulting optimal hyperplane in this transformed space ensures a maximum span between the different classes. The data points, or 'vectors," closest to this hyperplane, called 'support vectors', determine the separation distance. Multiclass SVM, as the name suggests, is used to split data into multiple classes. This can be achieved by training multiple binary classifiers and merging their outputs, or by using a single classifier with multiple output values [45]–[48].

2.2.3. Logistic Regression

Logistic regression [49]–[51] is a statistical method for analyzing and modeling the relationship between a binary dependent variable and one or more independent variables. In logistic regression, the logistic function is used to estimate the probability that the outcome is 1, given a set of independent variables. The function assigns a value between 0 and 1 to each input value, which can be interpreted as the probability that the outcome is 1. The logistic regression model is trained on a set of labeled data, where each data point has a set of independent variables and a binary outcome. The model learns the relationship between the independent variables and the outcome by adjusting the parameters of the model so that the predicted probabilities match the actual outcomes as closely as possible.

The logistic function is represented by an S-shaped curve, the so-called sigmoid curve, which is defined as follows:

P(x) = 1 / (1 + e^ (-b0 - b1*x))

where P(x) is the probability that the outcome is 1 given the argument x, b0 and b1 are the parameters of the model and e is the base of the natural logarithm. The following figure shows a logistic regression model based on a sigmoid function.

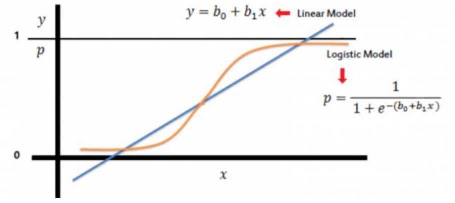


Figure 4 Logistic regression model based on the sigmoid function.

A logistic regression model is visually represented as an S-shaped curve with the probability of the dependent variable being '1' on the y-axis and the corresponding independent variable(s) on the x-axis. The curve starts at '0' on the left, moves through an inflexion point (the point of maximum slope), and ends at '1' on the right.

As the independent variable(s) increase, the curve becomes steeper, and this curve is symmetrical about the inflexion point. Therefore, the visual representation of logistic regression forms an S-shaped curve that illustrates the relationship between the independent variable(s) and the probability that the dependent variable has the value '1'.

This graphical interpretation helps to decipher the predictions of the model and understand the relationships between the variables, making it an indispensable tool for the insights gained from logistic regression.

2.3. Performance Metrics

There are several performance metrics that are used to evaluate the performance of a machine learning model [52]–[56]. In this study, the performance of the proposed system is evaluated using accuracy, precision, recall, and the F1-score. These metrics are commonly used to evaluate the performance of classification models. Accuracy is a measure of how well the system correctly predicts the class of instances. It is calculated as the ratio of correctly classified instances to the total number of instances. Precision is a measure of how well the system avoids false positives. It is calculated as the ratio of true positives to the total number of predicted positives. Recall, also known as "sensitivity", is a measure of how well the system finds all positive instances. It is calculated as the ratio of true positives to the total number of actual positive instances. The F1 score is a measure that combines both precision and recall. It is calculated as the harmonic mean of precision and recall. Using multiple metrics provides a better understanding of system performance. Understanding how well the system performs in terms of accuracy, precision, recall, and F1 score will help you identify the strengths and weaknesses of the proposed system.

Table 3 Performance metrics [57]–[60]

Abbreviation	*Description*	*Formula*
$\boldsymbol{ACC}$	$Accuracy$	$ACC = \frac{TP + TN}{TP + FP + TN + FN}$
$\boldsymbol{RCL}$	$Sensitivity\ (Recall)$	$RCL = \frac{TP}{TP + FN}$
$\boldsymbol{PRE}$	Precision	$PRE = \frac{TP}{TP + FP}$
FSC	F-1 Score	$\mathrm{FSC} = 2 * \frac{PRE.RCL}{PRE + RCL}$

The equations in Table 3 [61], [62] allow the calculation of the metrics for accuracy, precision, recall, and F1 score using the values of true positives (TP), false positives (FP), true negatives (TN) and false negatives (FN) from the confusion matrix (See table 4).

The Confusion Matrix [63], [64] is a tabular analysis tool that explicitly gives the number of true positives, true negatives, false positives and false negatives, all critical metrics for evaluating the performance of a binary classification model. This matrix essentially facilitates the accurate quantification of true and false predictions, allowing for a more nuanced assessment of the classifier's performance than simply assessing accuracy.

Table 4. Confusion matrix

Predicted Class		True Class	
		Positive	Negative
	Positive	TP	FP
	Negative	FN	TN

3. Experimental Results

In this study, machine learning algorithms implemented in LabVIEW are used to develop a decision support system for a sound wave-based fire extinguishing system. The system is designed to model fires caused by burning fuels using input parameters such as fuel type, flame size, decibel level, frequency, airflow, and distance. The aim of the study is to develop a system that can accurately predict the extinguishing and non-extinguishing states of a flame based on these parameters to enable more efficient use of the sound wave-based fire extinguishing system. Figure 5 show the block diagram perspectives of the proposed LabVIEW-based model. These images show the decision support system for the sound wave-based fire extinguishing system. The figures help to understand the planned design and operation of the system.

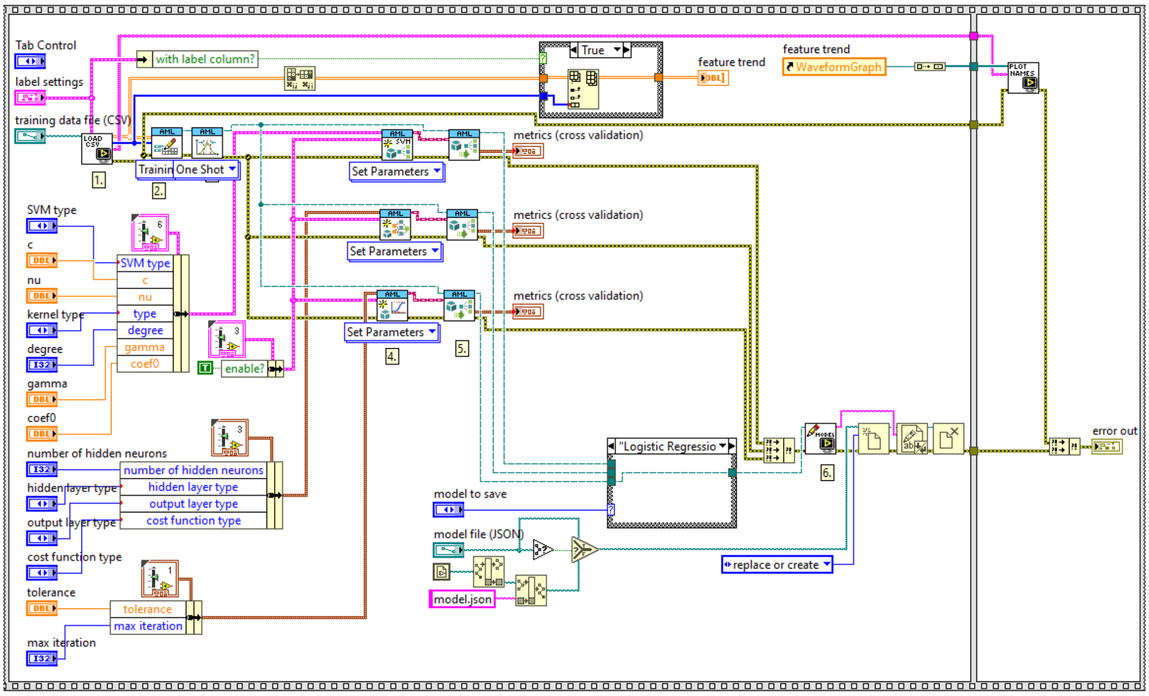


Figure 5 The block diagram of LabVIEW-based fire extinguisher model

Table 5 provides a summary of the algorithms and parameters used in the proposed model, such as the type of machine learning algorithm (e.g., neural network, SVM, logistic regression). This information can provide insight into how the model was constructed and how the different parameters were selected and used in the analysis. It can also give you an idea of how the model was trained and what factors were considered in the classification.

Table 5 Parameters setting LabVIEW-based fire extinguisher model.

Algorithm	Parameters	Values/types
SVM	SVM type	C_SVM
	Kernel type	Linear
	c	1
	nu	0,5
	degree	3
	gamma	0,5
	Coef0	0
Neural Network	Number of hidden layers	5
	Hidden layer type	sigmoid
	Output layer type	sigmoid
	Cost function type	Quadratic
Logistic Regression	Tolerance	0,001
	Max iteration	1000

Based on the values in Tables 4 and 5, various performance metrics such as accuracy, precision, detection, and F1 score were calculated. These performance metrics are a measure of the model's ability to correctly classify instances into positive and negative categories. Accuracy is the proportion of correctly classified instances out of the total number of instances. Precision is the proportion of correctly classified positive instances to the total number of predicted positive instances. Recall is the proportion of correctly classified positive instances out of the total number of actual positive instances. The F1 score is a measure of the trade-off between precision and recall. These performance measures give an overview of the performance of the model and how well it is able to classify instances into positive and negative categories. The results of these performance measures are shown in Table 6, which allows a comparison of the performance of the different algorithms used in the study.

Table 6 Performance metrics of learning algorithms

	ACC	*RCL*	*PRE*	*FSC*
SVM	0.86728	0.86096	0.86709	0.86716
ANN	**0.90893**	0.90874	0.90881	0.90885
LR	0.86836	0.86831	0.86829	0.86801

According to the developed models, the highest classification accuracy belongs to the model ANN with a value of 90.893%. The RLC, PRE, and FSC values of this model also seem to be higher than those of the other models listed in Table 6. According to this Table, the highest classification accuracy was achieved with the model ANN, with a value of 90.893%. The classification accuracy of the SVM and logistic regression models are 86.728% and 86.836%, respectively. This shows that the ANN model performs better than the other models in classifying the data correctly. These values may indicate that the ANN model correctly classifies the data and minimizes false positives and false negatives. However, it is important to note that accuracy is not always the best metric to evaluate the performance of a model. Other metrics such as precision, recall, and F1 score should also be considered.

4. Conclusion

The study presents the development of a sound wave-based fire extinguishing model using AI methods such as artificial neural networks, support vector machines, and logistic regression, implemented in LabVIEW. The model was able to classify the extinguishing and non-extinguishing states of a flame based on input parameters such as fuel type, flame size, decibel, frequency, airflow, and distance. The performance of the developed machine learning methods was analyzed and compared using performance metrics such as accuracy, precision, recall, and F1 score. The results of this study show that the highest classification accuracy of 90.893% was achieved by the neural network model, while it was 86.728% and 86.836% for the SVM and logistic regression models, respectively. This indicates that the neural network model performed best in classifying the extinguishing and non-extinguishing states of a flame. Furthermore, the use of sound wave-based models can provide a cost-effective and non-invasive alternative to traditional fire extinguishing methods. In summary, this study provides valuable insight into the potential of AI-based methods for solving fire extinguishing problems and can serve as a basis for future research in this area. The results show that the use of sound wave-based models can be an efficient and cost-effective alternative to traditional firefighting methods. Furthermore, the effectiveness of the model can be evaluated using various performance metrics. Overall, this study highlights the potential of AI-based methods in solving firefighting problems and shows how they can be a valuable tool for decision-making in firefighting systems.

Declarations

Funding The author would also like to thank "The Scientific Projects Coordinatorship of Şırnak University" for institutional support (project number: 2022.FNAP.06.01.01).

Competing interests, The authors declare no competing interests.

References

[1] C. A. Matticks, J. J. Westwater, H. N. Himel, R. F. Morgan, and R. F. Edlich, ‘Health Risks to Fire Fighters’, *J Burn Care Rehabil*, vol. 13, no. 2, pp. 223–235, Mar. 1992, doi: 10.1097/00004630-199203000-00010.

[2] A. B. Morgan and J. W. Gilman, ‘An overview of flame retardancy of polymeric materials: Application, technology, and future directions’, *Fire Mater*, vol. 37, no. 4, pp. 259–279, Jun. 2013, doi: 10.1002/FAM.2128.

[3] M. Shokouhi, K. Nasiriani, H. Khankeh, H. Fallahzadeh, and D. Khorasani-Zavareh, ‘Exploring barriers and challenges in protecting residential fire-related injuries: a qualitative study’, *J Inj Violence Res*, vol. 11, no. 1, p. 81, 2019, doi: 10.5249/JIVR.V11I1.1059.

[4] R. Olawoyin, ‘Nanotechnology: The future of fire safety’, 2018, doi: 10.1016/j.ssci.2018.08.016.

[5] Y. Awad, M. Kohail, M. A. Khalaf, and Y. A. Ali, ‘Effect of fire extinguishing techniques on the strength of RC columns’, *Asian Journal of Civil Engineering*, vol. 23, no. 1, pp. 113–123, Jan. 2022, doi: 10.1007/S42107-021-00414-8/FIGURES/12.

[6] F. Dubocq *et al.*, ‘Organic contaminants formed during fire extinguishing using different firefighting methods assessed by nontarget analysis’, *Environmental Pollution*, vol. 265, p. 114834, Oct. 2020, doi: 10.1016/J.ENVPOL.2020.114834.

[7] Y. Zhou, R. Bu, J. Gong, X. Zhang, C. Fan, and X. Wang, ‘Assessment of a clean and efficient fire-extinguishing technique: Continuous and cycling discharge water mist system’, *J Clean Prod*, vol. 182, pp. 682–693, May 2018, doi: 10.1016/J.JCLEPRO.2018.02.046.

[8] M. Rajczyk *et al.*, ‘Application of acoustic oscillations in flame extinction in a presence of obstacle’, *J Phys Conf Ser*, vol. 1101, no. 1, p. 012023, Oct. 2018, doi: 10.1088/1742-6596/1101/1/012023.

[9] A. B. Morgan and J. W. Gilman, ‘An overview of flame retardancy of polymeric materials: application, technology, and future directions’, *Fire Mater*, vol. 37, no. 4, pp. 259–279, Jun. 2013, doi: 10.1002/FAM.2128.

[10] V. Sharifi, A. M. Kempf, and C. Beck, ‘Large-Eddy Simulation of Acoustic Flame Response to High-Frequency Transverse Excitations’, *https://doi.org/10.2514/1.J056818*, vol. 57, no. 1, pp. 327–340, Nov. 2018, doi: 10.2514/1.J056818.

[11] Y. S. Taspinar, M. Koklu, and M. Altin, ‘Acoustic-Driven Airflow Flame Extinguishing System Design and Analysis of Capabilities of Low Frequency in Different Fuels’, *Fire Technol*, vol. 58, no. 3, pp.

1579–1597, May 2022, doi: 10.1007/S10694-021-01208-9/TABLES/4.

[12] A. N. Friedman and S. I. Stoliarov, 'Acoustic extinction of laminar line-flames', *Fire Saf J*, vol. 93, pp. 102–113, Oct. 2017, doi: 10.1016/J.FIRESAF.2017.09.002.

[13] X. Shi, Y. Zhang, X. Chen, Y. Zhang, Q. Ma, and G. Lin, 'The response of an ethanol pool fire to transverse acoustic waves', *Fire Saf J*, vol. 125, p. 103416, Oct. 2021, doi: 10.1016/J.FIRESAF.2021.103416.

[14] C. Xiong, Y. Liu, C. Xu, and X. Huang, 'Acoustical Extinction of Flame on Moving Firebrand for the Fire Protection in Wildland–Urban Interface', *Fire Technol*, vol. 57, no. 3, pp. 1365–1380, May 2021, doi: 10.1007/S10694-020-01059-W/FIGURES/11.

[15] J. O'Connor, V. Acharya, and T. Lieuwen, 'Transverse combustion instabilities: acoustic, fluid mechanic, and flame processes', *Prog Energy Combust Sci*, vol. 49, pp. 1–39, Aug. 2015, doi: 10.1016/j.pecs.2015.01.001.

[16] A. N. Friedman and S. I. Stoliarov, 'Acoustic extinction of laminar line-flames', *Fire Saf J*, vol. 93, pp. 102–113, Oct. 2017, doi: 10.1016/j.firesaf.2017.09.002.

[17] F. Baillot and F. Lespinasse, 'Response of a laminar premixed V-flame to a high-frequency transverse acoustic field', *Combust Flame*, vol. 161, no. 5, pp. 1247–1267, May 2014, doi: 10.1016/J.COMBUSTFLAME.2013.11.009.

[18] E. Beisner *et al.*, 'Acoustic Flame Suppression Mechanics in a Microgravity Environment', *Microgravity Sci Technol*, vol. 27, no. 3, pp. 141–144, Jun. 2015, doi: 10.1007/S12217-015-9422-4/FIGURES/5.

[19] T. Y. T. K. M Tunabe, 'Numerical Simulation on the Flame Propagation in Acoustic Fields', *JASMA*, vol. 23, pp. 371–375, 2008.

[20] M. Z. Abbasi, P. S. Wilson, and O. A. Ezekoye, 'Modeling acoustic propagation in a compartment fire', *J Acoust Soc Am*, vol. 134, no. 5, pp. 4218–4218, Nov. 2013, doi: 10.1121/1.4831486.

[21] M. Z. Abbasi, O. A. Ezekoye, and P. S. Wilson, 'Measuring the acoustic response of a compartment fire', *Proceedings of Meetings on Acoustics*, vol. 19, 2013, doi: 10.1121/1.4799626.

[22] M. Z. Abbasi, P. S. Wilson, and O. A. Ezekoye, 'Change in acoustic impulse response of a room due to a fire', *J Acoust Soc Am*, vol. 147, no. 6, p. EL546, Jun. 2020, doi: 10.1121/10.0001415.

[23] M. J. Sousa, A. Moutinho, and M. Almeida, 'Classification of potential fire outbreaks', *Expert Syst Appl*, vol. 129, pp. 216–232, Sep. 2019, doi: 10.1016/J.ESWA.2019.03.030.

[24] Y. Ye, X. Luo, C. Dong, Y. Xu, and Z. Zhang, 'Numerical and experimental investigation of soot suppression by acoustic oscillated combustion', *ACS Omega*, vol. 5, no. 37, pp. 23866–23875, Sep. 2020, doi: 10.1021/ACSOMEGA.0C03107/SUPPL_FILE/AO0C03107_SI_006.AVI.

[25] J. Lloret, M. Garcia, D. Bri, and S. Sendra, 'A wireless sensor network deployment for rural and forest fire detection and verification', *Sensors*, vol. 9, no. 11, pp. 8722–8747, Nov. 2009, doi: 10.3390/S91108722.

[26] B. L. Wenning, D. Pesch, A. Timm-Giel, and C. Görg, 'Environmental monitoring aware routing: Making environmental sensor networks more robust', *Telecommun Syst*, vol. 43, no. 1–2, pp. 3–11, Feb. 2010, doi: 10.1007/S11235-009-9191-8.

[27] A. A. A. Alkhatib, 'A Review on Forest Fire Detection Techniques':, *http://dx.doi.org/10.1155/2014/597368*, vol. 2014, Mar. 2014, doi: 10.1155/2014/597368.

[28] P. Barmpoutis, P. Papaioannou, K. Dimitropoulos, and N. Grammalidis, 'A Review on Early Forest Fire Detection Systems Using Optical Remote Sensing', *Sensors 2020, Vol. 20, Page 6442*, vol. 20, no. 22, p. 6442, Nov. 2020, doi: 10.3390/S20226442.

[29] K. Grover, D. Kahali, S. Verma, and B. Subramanian, 'WSN-Based System for Forest Fire Detection and Mitigation', pp. 249–260, 2020, doi: 10.1007/978-981-13-7968-0_19.

[30] S. J. Chen, D. C. Hovde, K. A. Peterson, and A. W. Marshall, 'Fire detection using smoke and gas sensors', *Fire Saf J*, vol. 42, no. 8, pp. 507–515, Nov. 2007, doi: 10.1016/J.FIRESAF.2007.01.006.

[31] G. H. Mitri, I. Z. Gitas, G. H. Mitri, and I. Z. Gitas, 'Fire type mapping using object-based classification of Ikonos imagery', *Int J Wildland Fire*, vol. 15, no. 4, pp. 457–462, Dec. 2006, doi: 10.1071/WF05085.

[32] I. Z. Gitas, G. H. Mitri, and G. Ventura, 'Object-based image classification for burned area mapping of Creus Cape, Spain, using NOAA-AVHRR imagery', *Remote Sens Environ*, vol. 92, no. 3, pp. 409–413, Aug. 2004, doi: 10.1016/J.RSE.2004.06.006.

[33] M. G. Cruz, J. S. Gould, J. J. Hollis, and W. L. McCaw, 'A Hierarchical Classification of Wildland Fire Fuels for Australian Vegetation Types', *Fire 2018, Vol. 1, Page 13*, vol. 1, no. 1, p. 13, Apr. 2018, doi: 10.3390/FIRE1010013.

[34] 'Acoustic Extinguisher Fire Dataset | Kaggle'. https://www.kaggle.com/datasets/muratkokludataset/acoustic-extinguisher-fire-dataset (accessed May 27, 2022).

[35] Y. S. Taspinar, M. Koklu, and M. Altin, 'Classification of flame extinction based on acoustic oscillations using artificial intelligence methods', *Case Studies in Thermal Engineering*, vol. 28, Dec. 2021, doi: 10.1016/J.CSITE.2021.101561.

[36] Y. S. Taspinar, M. Koklu, and M. Altin, 'Acoustic-Driven Airflow Flame Extinguishing System Design and Analysis of Capabilities of Low Frequency in Different Fuels', *Fire Technol*, May 2022, doi: 10.1007/S10694-021-01208-9.

[37] M. Koklu and Y. S. Taspinar, 'Determining the Extinguishing Status of Fuel Flames with Sound Wave by Machine Learning Methods', *IEEE Access*, vol. 9, pp. 86207–86216, 2021, doi: 10.1109/ACCESS.2021.3088612.

[38] W. S. McCulloch and W. Pitts, 'A logical calculus of the ideas immanent in nervous activity', *Bull Math Biophys*, vol. 5, no. 4, pp. 115–133, Dec. 1943, doi: 10.1007/BF02478259/METRICS.

[39] 'Hebb, D. O. The organization of behavior: A neuropsychological theory. New York: John Wiley and Sons, Inc., 1949. 335 p. $4.00', *Sci Educ*, vol. 34, no. 5, pp. 336–337, Dec. 1950, doi: 10.1002/SCE.37303405110.

[40] F. Rosenblatt, 'The perceptron: A probabilistic model for information storage and organization in the brain', *Psychol Rev*, vol. 65, no. 6, pp. 386–408, Nov. 1958, doi: 10.1037/H0042519.

[41] C. A. Tudor, 'Analysis of the Rosenblatt process', *ESAIM: Probability and Statistics*, vol. 12, pp. 230–257, Oct. 2008, doi: 10.1051/PS:2007037.

[42] A. Shmilovici, 'Support Vector Machines', *Data Mining and Knowledge Discovery Handbook*, pp. 231–247, 2009, doi: 10.1007/978-0-387-09823-4_12.

[43] A. v. Joshi, 'Support Vector Machines', *Machine Learning and Artificial Intelligence*, pp. 89–99, 2023, doi: 10.1007/978-3-031-12282-8_8.

[44] Ingo. Steinwart and Andreas. Christmann, 'Support vector machines', p. 601, 2008.

[45] G. Teles, J. J. P. C. Rodrigues, R. A. L. Rabêlo, and S. A. Kozlov, 'Comparative study of support vector machines and random forests machine learning algorithms on credit operation', *Softw Pract Exp*, vol. 51, no. 12, pp. 2492–2500, Dec. 2021, doi: 10.1002/SPE.2842.

[46] S. Kim, Z. Yu, R. M. Kil, and M. Lee, 'Deep learning of support vector machines with class probability output networks', *Neural Networks*, vol. 64, pp. 19–28, Apr. 2015, doi: 10.1016/J.NEUNET.2014.09.007.

[47] B. Schölkopf, 'SVMs - A practical consequence of learning theory', *IEEE Intelligent Systems and Their Applications*, vol. 13, no. 4, pp. 18–21, Jul. 1998, doi: 10.1109/5254.708428.

[48] 'Support Vector Machines for Regression.', *Support Vector Machines*, pp. 330–351, Aug. 2008, doi: 10.1007/978-0-387-77242-4_9.

[49] S. Nusinovici *et al.*, 'Logistic regression was as good as machine learning for predicting major chronic diseases', *J Clin Epidemiol*, vol. 122, pp. 56–69, Jun. 2020, doi: 10.1016/J.JCLINEPI.2020.03.002.

[50] T. Rymarczyk, E. Kozłowski, G. Kłosowski, and K. Niderla, 'Logistic Regression for Machine Learning in Process Tomography', *Sensors 2019, Vol. 19, Page 3400*, vol. 19, no. 15, p. 3400, Aug. 2019, doi: 10.3390/S19153400.

[51] E. Bisong, 'Logistic Regression', *Building Machine Learning and Deep Learning Models on Google Cloud Platform*, pp. 243–250, 2019, doi: 10.1007/978-1-4842-4470-8_20.

[52] S. Orozco-Arias, J. S. Piña, R. Tabares-Soto, L. F. Castillo-Ossa, R. Guyot, and G. Isaza, 'Measuring Performance Metrics of Machine Learning Algorithms for Detecting and Classifying Transposable Elements', doi: 10.3390/pr8060638.

[53] S. Adhikari, S.-L. Normand, J. Bloom, D. Shahian, and S. Rose, 'Revisiting performance metrics for prediction with rare outcomes', doi: 10.1177/09622802211038754.

[54] M. Steurer, R. J. Hill, and N. Pfeifer, 'Metrics for evaluating the performance of machine learning based automated valuation models', *Journal of Property Research*, vol. 38, no. 2, pp. 99–129, 2021, doi: 10.1080/09599916.2020.1858937.

[55] A. Rácz, D. Bajusz, and K. Héberger, 'molecules Multi-Level Comparison of Machine Learning Classifiers and Their Performance Metrics', doi: 10.3390/molecules24152811.

[56] M. Dirik, 'Optimized Anfis Model with Hybrid Metaheuristic Algorithms for Facial Emotion Recognition', *International Journal of Fuzzy Systems*, pp. 1–12, Oct. 2022, doi: 10.1007/S40815-022-01402-Z/FIGURES/5.

[57] D. Chicco and G. Jurman, 'The advantages of the Matthews correlation coefficient (MCC) over F1 score and accuracy in binary classification evaluation', *BMC Genomics*, vol. 21, no. 1, pp. 6-1-6–13, Jan. 2020, doi: 10.1186/s12864-019-6413-7.

[58] D. M. W. Powers, 'Evaluation: From Precision, Recall and F-Measure to ROC, Informedness, Markedness & Correlation', *Journal of Machine Learning Technologies*, vol. 2, no. 1, pp. 37–63, 2011, Accessed: Oct. 28, 2021. [Online]. Available: https://api.semanticscholar.org/CorpusID:55767944#id-name=S2CID

[59] T. Fawcett, 'An Introduction to ROC Analysis', *Pattern Recognit Lett*, vol. 27, no. 8, pp. 861–874, Jun. 2006, doi: 10.1016/j.patrec.2005.10.010.

[60] D. Chicco and G. Jurman, 'The advantages of the Matthews correlation coefficient (MCC) over F1 score and accuracy in binary classification evaluation', *BMC Genomics*, vol. 21, no. 1, pp. 6-1-6–13, Jan. 2020, doi: 10.1186/s12864-019-6413-7.

[61] S. Josephine Isabella, S. Srinivasan, and G. Suseendran, 'An Efficient Study of Fraud Detection System Using Ml Techniques', in *Lecture Notes in Networks and Systems*, Springer, 2020, pp. 59–67. doi: 10.1007/978-981-15-3284-9_8.

[62] A. A. Taha and S. J. Malebary, 'An Intelligent Approach to Credit Card Fraud Detection Using an Optimized Light Gradient Boosting Machine', *IEEE Access*, vol. 8, pp. 25579–25587, 2020, doi: 10.1109/ACCESS.2020.2971354.

[63] 'Confusion Matrix - an overview | ScienceDirect Topics'. https://www.sciencedirect.com/topics/engineering/confusion-matrix (accessed Jan. 23, 2023).

[64] M. Makhtar, D. C. Neagu, and M. J. Ridley, 'Comparing multi-class classifiers: On the similarity of confusion matrices for predictive toxicology applications', *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, vol. 6936 LNCS, pp. 252–261, 2011, doi: 10.1007/978-3-642-23878-9_31.