



idap 2023

**International
Artificial
Intelligence
and Data Processing
Symposium**

September/07-08/2023

<http://idap.inonu.edu.tr> - <http://www.asoiu.edu.az>

1920
ANATOLIAN
SCIENCE

1987
İNÖNÜ ÜNİVERSİTESİ
MÜHENDİSLİK
FAKÜLTESİ

Signal processing hardware
Information security
Robotics
Signal processing
Data mining
Deep learning
Big data
Computer vision
Spatio-spectral analysis
Computational statistics
Brain-computer interfaces

Organization Committee

Honory Chair/s	Prof. Dr. Ahmet KIZILAY (İnönü University) Prof. Dr. Ali KARCI (İnönü University)
Chairman	Dr. Kenan İNCE Prof. Dr. Teymuraz ABBASOV
Technical Chair	Dr. Fatih Okumuş
Organization Committee	A.A. Aydın (İnönü University) A. Ateş (İnönü University) A. Başçı (Atatürk Üniversitesi) A. Karadoğan (İnönü University) B.B. Alagöz (İnönü University) Ç. Erçelik (İnönü University) C. Hark (İnönü University) C. İnce (İnönü University) D. Hanbay (İnönü University) E. A. Oymak (İnönü University) E. Seyyarer (Yüzüncü Yıl University) E. Seyyarer (Yüzüncü Yıl University) E. Yiğit (Yüzüncü Yıl University) F. Okumuş (İnönü University) F. Öztemiz (İnönü Üniversitesi) İ. Tuğal (Muş Alpaslan University) K. Hanbay (İnönü University) K. İnce (İnönü University) M. İnan (Yüzüncü Yıl University) Ö. F. Alçin (İnönü University) R. Daş (Firat University) R. Özdağ (Yüzüncü Yıl University) S.E. Hamamcı (İnönü University) S. Yakut (İnönü University) T. Uçkan (Yüzüncü Yıl University)

Scientific Committee

Dr. Ahmet Arif Aydın (İnönü University)
Dr. Ahmet Arslan (Konya Food & Agriculture University)
Dr. Ahmet Çınar (Fırat University)
Dr. Ali Erdoğan (İnönü University)
Dr. Ali Karci (İnönü University)
Dr. Ali Kaveh (Iran University Of Science & Technology)
Dr. Ali Safak Sekmen (Tennessee State University)
Dr. Asım Kaygusuz (İnönü University)
Dr. Aşkın Demirkol (Sakarya University)
Dr. Ayhan İstanbullu (Balıkesir University)
Dr. B.Baykant Alagöz (İnönü University)
Dr. Bedri Özer (Fırat University)
Dr. Berat Doğan (İnönü University)
Dr. Bilal Alataş (Fırat University)
Dr. Bilal Şenol (İnönü University)
Dr. B. Ramakrishnan (S.T. Hindu College)
Dr. Bugra Koku (Middle East Technical University)
Dr. Burhan Ergen (Fırat University)
Dr. Burhan SELÇUK (Karabük University)
Dr. Cafer Bal (Fırat University)
Dr. Celaledin Yeroğlu (İnönü University)
Dr. Cemil Çolak (İnönü University)
Dr. Çiğdem İnan ACI (Mersin University)
Dr. Davut Hanbay (İnönü University)
Dr. Deniz KILINÇ (Celal Bayar University)
Dr. Kazım Hanbay (İnönü University)
Dr. Erdem Erdemir (Tennessee State University)
Dr. Erdinç Avaroğlu (Mersin University)
Dr. Erhan Akbal (Fırat University)
Dr. Erkan Tanyıldızı (Fırat University)
Dr. Fatih Özkaynak (Fırat University)

Dr. Fatih Tüysüz (Harran University)
Dr. Fuad Mammadov (Azerbaijan State Oil and Industry University)
Dr. Furkan Öztemiz (İnönü Üniversitesi)
Dr. Galip Aydın (Fırat University)
Dr. İbrahim Berkan Aydılek (Harran University)
Dr. İbrahim Türkoğlu (Fırat University)
Dr. İlhan Aydın (Fırat University)
Dr. Jaroslav Koton (Brno University of Technology)
Dr. Kemal Balıkçı (Osmaniye University)
Dr. Latafat Gardashova (Azerbaijan State Oil and Industry University)
Dr. Mahmudur Rahman (Morgan State University)
Dr. Marouane EL MABROUK (Abdelmalek Essaadi University)
Dr. Mehmet ACI (Mersin University)
Dr. Mehmet Emin Tağluk (İnönü University)
Dr. Mehmet Karaköse (Fırat University)
Dr. Mehmet Kaya (Fırat University)
Dr. Morten Goodwin (University Of Agder)
Dr. M. Sıraç Özerdem (Dicle University)
Dr. Muhammed Fatih Talu (İnönü University)
Dr. Murat Canayaz (Yuzuncu Yil University)
Dr. Murat Demir (Mus Alpaslan University)
Dr. Murat Karabatak (Fırat University)
Dr. Musa Çıbuk (Bitlis University)
Dr. Mustafa Türk (Fırat University)
Dr. Ömer Faruk Ertuğrul (Batman University)
Dr. Omer Faruk Ozguven (İnönü University)
Dr. Prabir Bhattacharya (Morgan State University)
Dr. Recep Halicioğlu (Osmaniye Korkut Ata University)
Dr. Recep Ozdag (Yuzuncu Yil University)
Dr. Resul Daş (Fırat University)

Dr. Ruksar Fatima (KBN College of Engineering)
Dr. Sabri Koçer (Necmettin Erbakan University)
Dr. Seda Arslan Tuncer (Firat University)
Dr. Selman Yakut (İnönü University)
Dr. Serdar Ethem Hamamci (İnönü University)
Dr. Taner Tuncer (Firat University)
Dr. Tolga Aydın (Atatürk University)
Dr. Ulaş Baran Baloğlu (University of Bristol)
Dr. Walter Boles (Middle Tennessee State University)
Dr. Yanhui Guo (University Of Illinois Springfield)

İÇİNDEKİLER / CONTENTS

Gözde Demirsoy, Nihat Özsoy, Dilşat Berin Aytar, Berak Burak Kaya, Bülent Tuğrul, “ An analysis of Antalya’s wind energy potential utilizing machine learning algorithms ”	1
Ayşe Özavcı, Selman Yakut, “ Based on Malatya centrality algorithm development of suggestion system in social platforms and commercial applications ”	11
Enis Çetin, Zafer İşcan, “ Comparison of machine and human vision based on brightness and contrast using Yolov3 with fuzzy logic ”	19
Filiz Şenyüzlüler Özçelik, Adil Baykaşoğlu, “ Kiralık ev sektöründe veriye dayalı öneri sistemi geliştirilmesi ve uygulanması ”	30
Hayriye Tanyıldız, Canan Batur Şahin, Özlem Batur Dinler, “ Endüstriyel siber güvenlik sistemleri için Siemens S7-300 PLC honeypot tasarımı ”	36
Duran Bala, Ahmet Doğan, “ Güneş ısıtım değerlerinin uzun kısa süreli bellek yöntemi ile tahmin edilmesi ”	46
Özkan Arslan, “ Infant cry classification by using adaptive cepstral features and machine learning approach ”	54
Cemalettin Sonakalan, Furkan Öztemiz, “ Kablosuz sensor ağlarının Malatya minimum vertex cover yöntemi ile konumlandırılması ”	63
Ali Nihat Uzunalioglu, Behçet Mutlu, Deniz Kılınç, “ Optimizing shift scheduling with constraint programming: A practical approach using google OR-tools ”	70
Özgün Karadeniz, Erdem Yelken, “ Anomaly detection-based web application firewall using machine-learning techniques ”	79
Nida Tekedereli Özçelik, Barış Baykant Alagöz, “ Differential evolutionary optimal architecture deep neural network model for heating load estimation of buildings ”	88

An Analysis of Antalya's Wind Energy Potential Utilizing Machine Learning Algorithms

Gozde DEMIRSOY¹, Nihat OZSOY¹, Dilsat Berin AYTAZ¹,

Berat Burak KAYA¹, Bulent TUGRUL¹

¹Department of Computer Engineering, Ankara University, Ankara, Türkiye

(gzddemirsoy@ankara.edu.tr, nozsoy@ankara.edu.tr, dbaytar@ankara.edu.tr, bbkaya@ankara.edu.tr, btugrul@eng.ankara.edu.tr)

Abstract— Technology, population growth, economic prosperity, and industrialization have all contributed to an increase in energy demand. Society and industry are becoming more interested in renewable energy sources because fossil fuel reserves are dwindling and environmental and economic problems are increasing. Wind energy has emerged as one of the fastest-developing renewable energy technologies contributing to electricity production. As part of energy supply security, accurate forecasting of wind energy generation is crucial to reduce the negative impacts of wind energy on the electricity market, which has a variable profile, and thus allows for efficient and effective utilization of this energy. In this study, various machine learning algorithms are used to predict wind power in Antalya using daily wind speed data. According to the results, the SVR method performs better than other machine learning methods. This study shows that machine learning algorithms can select the most suitable locations for wind farms before installing costly wind energy equipment.

Keywords : Renewable energy, wind energy, wind power generation prediction, machine learning.

1. Introduction

Energy is the driving force behind countries' socioeconomic development and a strategic element of economic growth. Energy production and consumption are at the forefront of countries' development criteria (Greene et al. 2013). When energy production and consumption figures are analyzed, traditional energy sources have the largest share. However, as a result of the limited economic life of these resources, the pollution they cause to the environment, and the energy needs of future generations, renewable energy resources are becoming increasingly important.

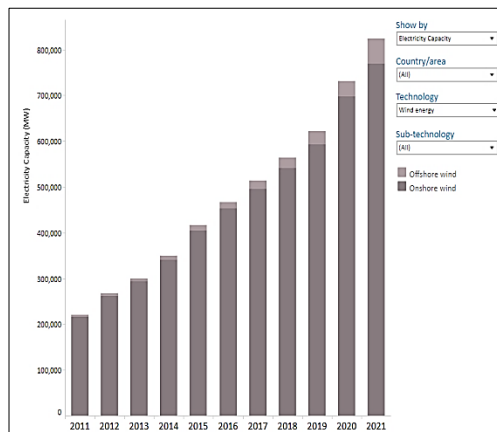
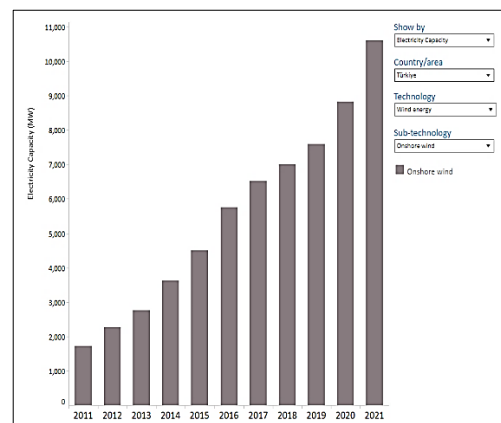
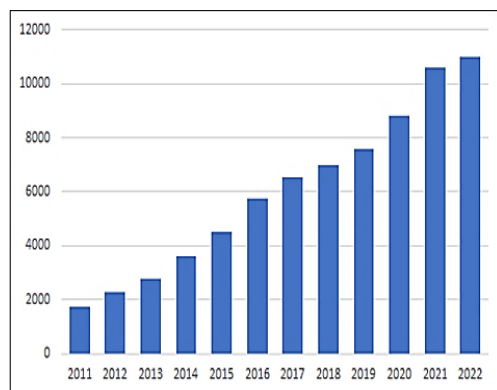
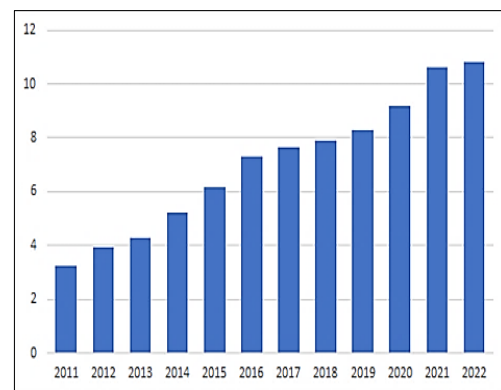
Wind energy is one of the alternative energy types that contributes to national energy supply security due to the decrease in its costs and increases in its efficiency every year, as well as being an environmentally friendly option in cases where the decrease in global reserves of fossil fuels threatens the long-term sustainability of the global economy. For these reasons, wind energy technologies have been focused on worldwide in recent years.

As in the rest of the world, Türkiye's trend towards sustainable energy sources has drawn attention recently because it has a wide range of renewable energy resources (Erdoğan 2009). With its strengths and significant potential, wind energy has become one of the most used and preferred clean energy sources. Türkiye's wind energy potential is approximately 48.000 MW based on the Wind Energy Potential Atlas, REPA (Republic of Türkiye The Ministry of Energy and Natural Resources (MENR), 2022). In this Atlas, wind resource information is produced using a medium-scale numerical weather forecast model and a micro-scale wind flow model. Table 1 explains it in more detail.

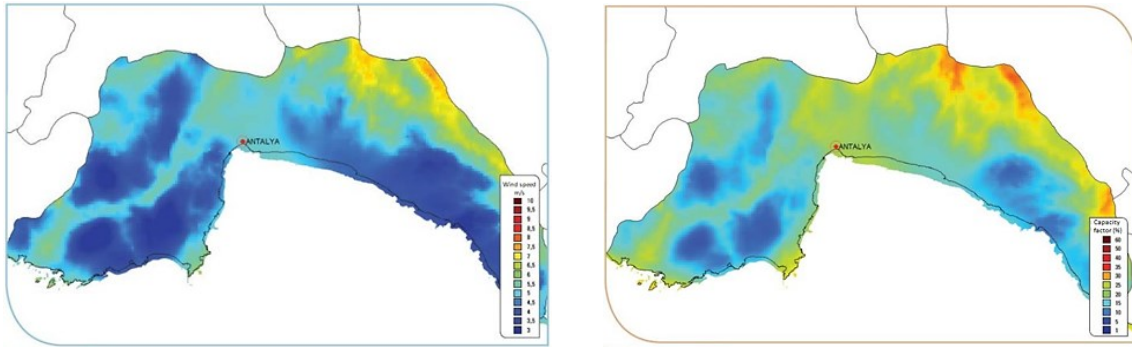
Table 1. Türkiye's annual wind energy potential (MENR, 2022)

Wind Class	Annual Wind Power Density (W / m ²)	Annual Average Wind Speed (m/s)	Total Capacity (MW)
4	400 – 500	7,0 – 7,5	29.259
5	500 – 600	7,5 – 8,0	12.994
6	600 – 800	8,0 – 9,0	5.400
7	>800	> 9,0	196
Total Capacity			47.849

In addition, as can be seen in Fig. 1 wind power capacity and electricity generation have increased steadily. Hence, while the world wind power installed capacity was 220.121 MW in 2011, it is 823.484 MW in 2021 (The International Renewable Energy Agency (IRENA), 2023). In addition, while wind energy installed power capacity in Türkiye was 1,729 MW in 2011, as of the end of June 2022, it reached 10.976 MW, corresponding to 10.81% of Türkiye's electricity produced power (Republic of Türkiye The Ministry of Energy and Natural Resources (MENR) 2023, IRENA 2023). Fig. 3 shows the distribution of provinces' installed wind power capacity.

**a. World****b. Türkiye****Fig 1.** Electricity capacity trends (2011-2021) (IRENA, 2023)**a. Installed power based on wind energy****b. The ratio of wind energy in total installed power**

The location is determined with respect to these parameters by benefiting from the wind energy potential atlas, REPA map prepared by the General Directorate of Renewable Energy (YEGM), earthquake map and bird migration map in Fig. 4-6 (Arca et al. 2020, Hacıoğlu et al. 2017, Solar Academy 2022, Birben 2019, İlkılıç 2012).



a. The wind speed distribution (m/s)

b. The capacity factor (%)

Fig 4. The wind speed distribution (m/s) and the capacity factor (%) in Antalya at 50 m (Solar Academy, 2022)

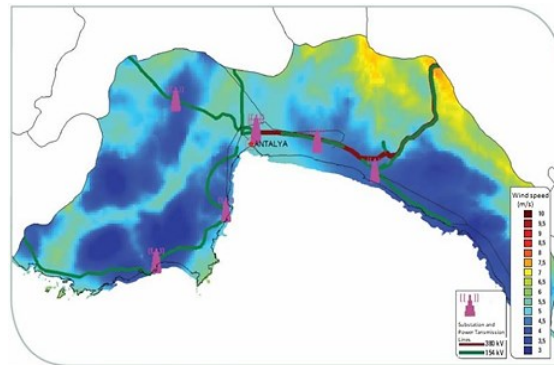


Fig 5. The substations and power transmission lines in Antalya (Solar Academy, 2022)

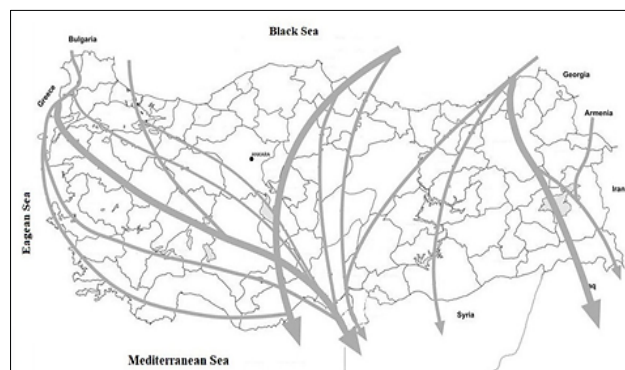


Fig 6. Main bird migration routes in Türkiye (Birben, 2019)

Northeast of Antalya was chosen as the location considering the figures. The data sets of Akseki Meteorological Station detailed in section 2.3 are utilized for estimation due to the data set availability.

The data set analyzed in this study includes meteorological observation data (wind speed, temperature, pressure, humidity values) between 2017 and 2021 obtained from the General Directorate of Meteorology (MGM). This study uses four years of hourly wind speed data in order to forecast long-term wind power generation for one year in advance using various machine learning algorithms.

2. Problem Definition and Data Set

The following section provides an overview of wind energy forecasting and presents calculations based on the data of wind speed measured hourly and converted into daily values. This section also provides a detailed description of the data set and how it was compiled.

2.1. Problem Definition:

The study's primary purpose is the prediction of wind power generation with regard to daily wind speed data. Four approaches were developed using machine learning algorithms to solve the problem. In the approaches, regression analysis was performed for prediction. In addition, Support Vector Regression (SVR), Random Forest (RF), Decision Tree (DT) and Gradient Boosting Regressor (GBR) were applied by data set, including daily wind speed.

In order to calculate wind power generation values for each day, we used the arithmetic mean. Machine learning models were trained with the data of daily wind speed and power. In order to determine whether they could provide sufficient wind power values based on the generated data set, they are tested using daily mean wind speed.

2.2. Wind Energy Output Calculations Based on Wind Speed:

The Eq. 1 calculates wind power generation in wind turbines stated below, where A is the area swept by the turbine blades, v is the wind speed, ρ is the air density and C_p is the wind turbine power coefficient [14].

$$P = \frac{1}{2} \rho A v^3 C_p \quad (1)$$

In this study, wind turbine LTW90 manufactured by Leitwind, and wind turbine E92 manufactured by Enercon are selected, taking into account the location of a power plant, manufacturability, and technical specifications of these turbines. Technical parameters of the considered turbines are stated below.

Table 3. Technical specifications of LTW90 and E92

Characteristics	LTW90	E92
Rated power (kW)	1.000	2.350
Hub height (m)	105	138
Rotor diameter (m)	90	92
Swept area (m ²)	6.404	6.648
Number of blades	3	3

2.3. Data Set:

Five-year data set of hourly wind speed observations obtained from Akseki, Antalya, Türkiye, is used in this study. The location of the meteorological station is indicated on the map in Fig. 7, and information about it is presented in Table 4.



Fig 7. Location of Akseki Meteorological Station

Table 4. Information of Akseki Meteorological Station

Name of the station	Latitude	Longitude	Altitude (m)
Akseki	37.0468	31.7971	1.063

Firstly, the values of hourly wind speed were converted into the values of daily average wind speed. In addition, daily wind speed data are observed at a height of 10 m. Since the wind speed values are used at the height of a wind turbine, the 10 m wind speed values need to be extrapolated to the hub heights of the turbines. Therefore, secondly, interpolation, the mathematical procedure, was applied to the values of daily average wind speed.

3. Results and Discussions

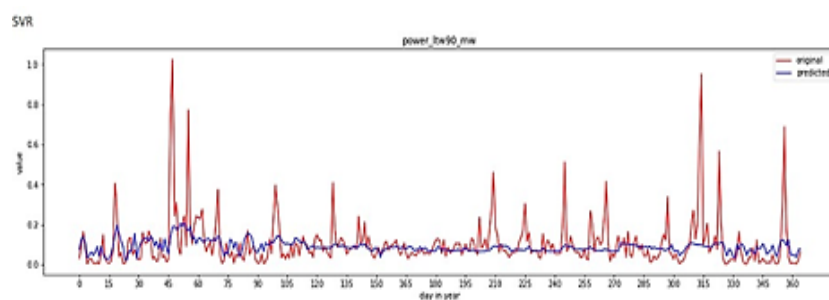
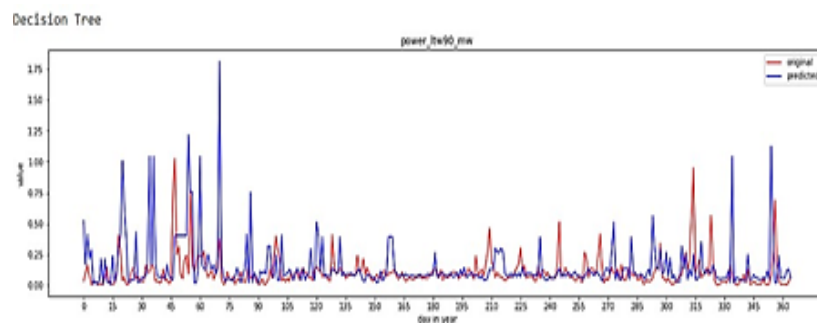
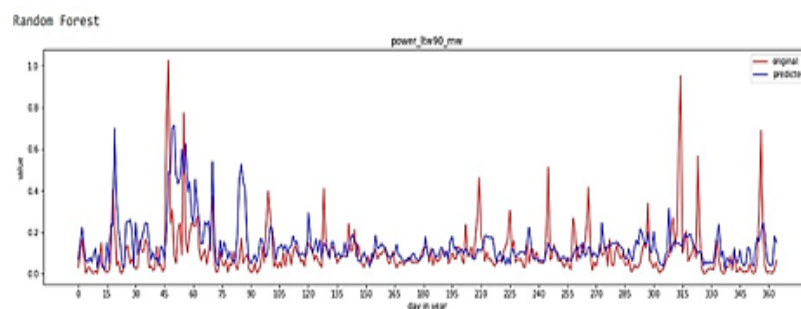
Several machine learning techniques are evaluated for wind power prediction. The performance of techniques is assessed by measuring mean absolute error (MAE), root mean square error (RMSE) and coefficient of determination (R^2).

Table 5. The metrics of the daily wind speed for LTW90 and E92

Algorithms	LTW90			E92		
	MAE	RMSE	R^2	MAE	RMSE	R^2
SVR	0,96	1,28	0,19	1,01	1,34	0,20
DT	1,14	1,57	-0,21	1,26	1,74	-0,33
RF	1,04	1,35	0,11	1,09	1,42	0,11
GBR	1,17	1,55	-0,18	1,26	1,67	-0,22

Table 6. The metrics of the daily wind power for LTW90 and E92

Algorithms	LTW90			E92		
	MAE	RMSE	R ²	MAE	RMSE	R ²
SVR	0,06	0,12	0,12	0,07	0,14	0,11
DT	0,11	0,22	-2,13	0,12	0,24	-1,71
RF	0,08	0,13	-0,20	0,10	0,16	-0,17
GBR	0,10	0,18	-1,25	0,12	0,21	-1,02

**Fig 8.** The original and predicted values of wind power of LTW90 for SVR**Fig 9.** The original and predicted values of wind power of LTW90 for DT**Fig 10.** The original and predicted values of wind power of LTW90 for RF

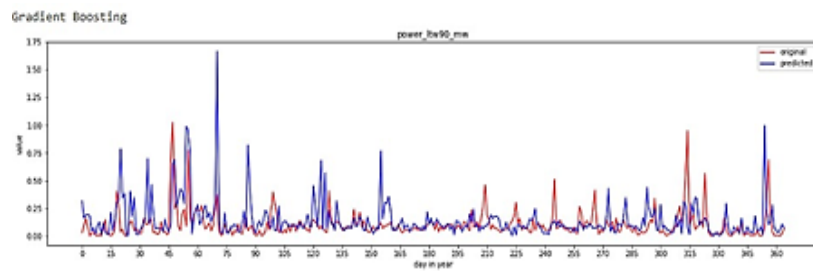


Fig 11. The original and predicted values of wind power of LTW90 for GBR

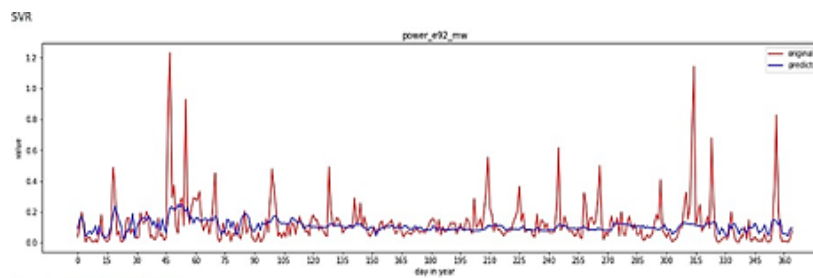


Fig 12. The original and predicted values of wind power of E92 for SVR

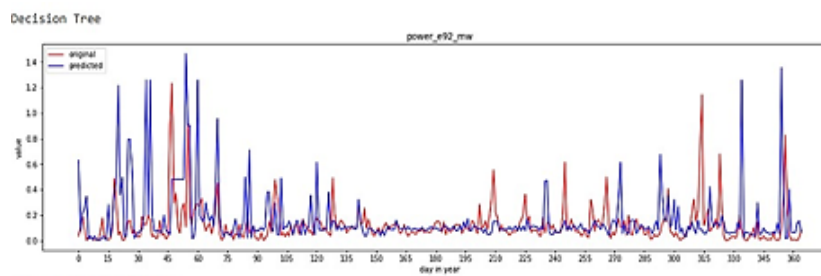


Fig 13. The original and predicted values of wind power of E92 for DT

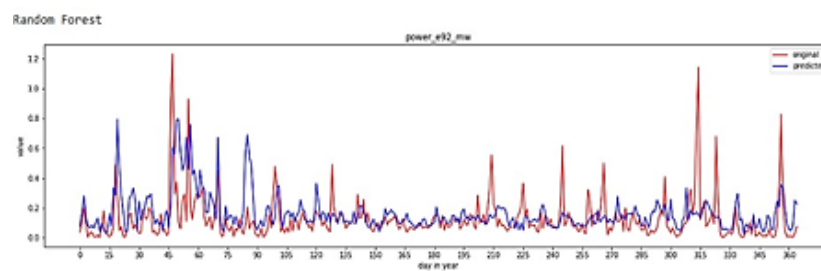


Fig 14. The original and predicted values of wind power of E92 for RF

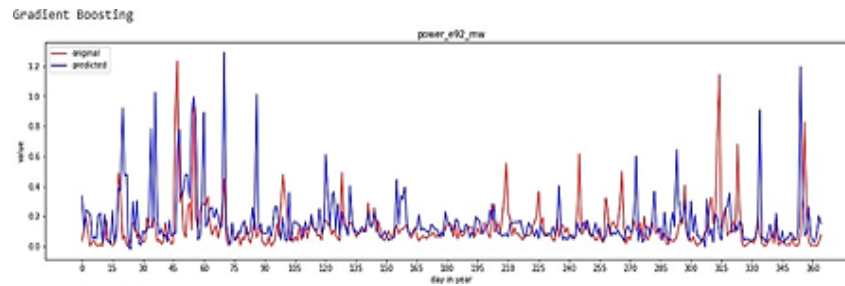


Fig 15. The original and predicted values of wind power of E92 for GBR

As can be observed from the tables and figures, the SVR algorithm appears to be more successful. This is because it has a higher R^2 value and lower error values among the algorithms we are analyzing.

4. Conclusion

Similar to other renewable energy sources, wind power is a domestic energy source that can help reduce our reliance on fossil fuels. As a result, wind energy forecasting is crucial to the successful operation of this resource in terms of reliability, economy, and quality. In this study, the daily wind speed parameter between 2017 and 2021 was utilized to predict daily wind energy generation. In addition, the different machine learning algorithms were used and evaluated for their performance by RMSE, MAE and R^2 . The proposed methods were applied to the Antalya in question to see if they could produce acceptable and reliable results based on trained data. According to the results, the SVR algorithm with an R^2 value of 0.12, RMSE of 0.12, and MAE of 0.06 for LTW90 is the best algorithm for forecasting wind power daily. For E92, the SVR with the R^2 value of 0.11, RMSE of 0.14, and MAE of 0.07 is the best algorithm. Among these algorithms, DT is the least effective algorithm for two turbines.

Acknowledgments

We would like to thank the Turkish State Meteorological Service (TSMS) for providing a data set including hourly wind speed, temperature, humidity, and pressure.

References

- Greene JS, Geisken M (2013), Socioeconomic impacts of wind farm development: a case study of Weatherford, Oklahoma. *Energy* 3(1): 1-9.
- Electricity capacity trends. <https://www.irena.org/>. Accessed 14 June 2022
- Erdoğan E (2009), On the wind energy in Turkey. *Renewable and Sustainable Energy Reviews* 13(6-7): 1361–1371.
- TÜRKİYE RÜZGAR ENERJİSİ POTANSİYELİ. <https://repa.enerji.gov.tr/REPA/>. Accessed 14 June 2022
- RÜZGAR. <https://enerji.gov.tr/eigm-yenilenebilir-enerji-kaynaklar-ruzgar#:~:text=T%C3%BCrkiye%20r%C3%BCzg%C3%A2r%20enerjisi%20potansiyeli%2048.000%20MW%20olarak%20belirlenmi%C5%9Ftir.,%2C30'una%20denk%20gelmektedir.> Accessed 14 June 2022
- Wind Energy. <https://www.irena.org/Energy-Transition/Technology/Wind-energy>. Accessed 20 July 2023
- RÜZGAR. <https://enerji.gov.tr/bilgi-merkezi-enerji-ruzgar>. Accessed 20 July 2023

- İllere göre rüzgar enerjisi kurulu gücü. <http://www.ntv.com.tr/ekonomi/illere-gore-ruzgar-enerjisi-kurulu-gucu,L76PfoQ-SE-f5R2IpiRDDw>. Accessed 14 June 2022
- Demolli H, Dokuz AS, Ecemiş A, Gökçek M (2019), Wind power forecasting based on daily wind speed data using machine learning algorithms. *Energy Conversion and Management* 198: 111823.
- Karık F, Sözen A, İzgeç MM (2017), Rüzgar gücü tahminlerinin önemi: Türkiye elektrik piyasasında bir uygulama. *Politeknik Dergisi* 20(4): 851–861.
- Türkiye Rüzgar Enerjisi Potansiyeli Haritası. <https://www.enerjiatlası.com/ruzgar-enerjisi-haritasi/turkiye>. Accessed 14 June 2022
- Elektrik Piyasası Yıllık Sektör Raporu Listesi. <https://www.epdk.gov.tr/Detay/Icerik/3-0-24/elektrikyillik-sektor-raporu>. Accessed 14 June 2022
- Van Haaren R, Fthenakis V (2011), GIS-based wind farm site selection using spatial multi-criteria analysis (smca): Evaluating the case for New York state. *Renewable and sustainable energy reviews* 15(7): 3332–3340.
- Aydın NY, Kentel E, Düzgün HS (2013), GIS-based site selection methodology for hybrid renewable energy systems: A case study from western Turkey. *Energy conversion and management* 70: 90–106.
- Arca D, Çıtıröğlü HK (2011), Rüzgar enerjisi santral (RES) yapım yerlerinin CBS dayalı çok kriterli karar analizi ile belirlenmesi: Yenice ilçesi (Karabük) örneği. *Karaelmas Science & Engineering Journal* 10(2).
- Hacıoğlu S, Dinçer E, İşler CT, Karapınar Z, Ataseven VS, Özkul A, Ergünay K (2017), A snapshot avian surveillance reveals west Nile virus and evidence of wild birds participating in Toscana virus circulation. *Vector-Borne and Zoonotic Diseases* 17(10): 698–708.
- Antalya-İli-Rüzgar-Kaynak-Bilgileri. <http://www.solar-academy.com/menus/Antalya-İli-Rüzgar-Kaynak-Bilgileri-REPA012822.pdf>. Accessed 14 June 2022
- Birben U (2019), The effectiveness of Protected Areas in Biodiversity Conservation: The Case of Turkey. *CERNE* 25(4): 424-438.
- İlkılıç C (2012), Wind energy and assessment of wind energy potential in Turkey. *Renewable and Sustainable Energy Reviews* 16(2): 1165–1173.

Based on Malatya Centrality Algorithm Development of Suggestion System in Social Platforms and Commercial Applications

Ayşe ÖZAVCI*¹ , Selman YAKUT*¹ 

¹Institute of Science, İnönü University, Department of MSc. Software Engineering, Malatya, TURKEY
(36223632003@ogr.inonu.edu.tr, selman.yakut@inonu.edu.tr)

Abstract—Recommendation systems are commonly utilized in social platforms and commercial applications. Recommendation systems constitute one of the domains of expert systems. The aim is to enhance user satisfaction through the provided recommendations. For this purpose, various parameters such as age, gender, profession are taken into account. Various algorithms have been employed for recommendation systems in previous studies, such as K-Means and KNN. The KNN algorithm is a supervised learning method that groups data with similar features. A portion of the data is used for training, typically around 90%. The remaining 10% constitutes the test data. There are various recommendation systems present in the literature. In this study, a new recommendation system was designed using the Malatya centrality algorithm. During the design of the recommendation system, work was conducted using movie data. A graph data structure was employed, where each node in the graph represents a movie. Edge weights were determined based on the number of common features between interconnected movies. By obtaining information about at least one movie the user has previously watched, the Malatya centrality algorithm was used to identify the most similar movie, which was then presented to the user as a recommendation.

Keywords : *Expert Systems, Graph Theory, Recommendation Systems, Malatya centrality algorithm*

1. Introduction

With the increasing use of social platforms and the internet in today's world, the need for recommendation systems has also grown. For instance, let's consider a movie platform; the goal here is to recommend movies that best match the user's preferences. This situation enables the user to use a personalized movie platform. Increased user satisfaction leads to more platform usage, ultimately allowing the provider to profit from this situation.

When examining the studies in the literature, it is observed that content-based filtering and collaborative filtering methods have been employed. In the content-based filtering method, recommendations are provided by suggesting movies similar to those the user has watched. In the collaborative filtering method, a technique of matching similar users has been utilized. This matching process takes into account user attributes such as occupation, gender, and age (Park, D. H., Kim, H. K., Choi, Y., and Kim, J. K., 2012). Recommendations can be given based on the user's similar attributes or through feedback received from the user. This recommendation method can be employed by having users rate their opinions. Consequently, the more positive the user's feedback, the better the quality of the recommendations can be asserted (Schelter, S., Boden, C., and Markl, V., 2012). Deep learning has also found its place in recommendation systems. In one of these studies, a Clothing Combination Recommendation System was designed using deep learning (Kara B., Kalkan H., 2020). In one of the previous studies on movie recommendation systems, using the MovieLens dataset, three distinct recommendation systems were developed and their performances were compared (Bozkurt M. B., Acı Ç. İ., 2021). Movie recommendation systems employing the K-NN and K-Means algorithms are also part of the conducted studies. The K-NN algorithm has been used to compute various parameters of movies using Euclidean distance, and movies with the closest distances have been recommended to the

user. Through the K-Means algorithm, recommendations have been presented to the user using clustering methodology (Ahuja, R., Solanki, A., & Nayyar, A., 2019). News recommendation systems, another application area of recommendation systems, have also been the subject of detailed scientific articles (Özgöbek Ö., Erdur R. C.). Another popular use of recommendation systems has been scientific studies. One of the challenges researchers face today is being able to access accurate and effective publications related to their research topics. In a study, a recommendation system was developed to suggest publications on research topics. The system was built upon the similarities and differences among scientific studies (Deniz E., Öz V. K., Keser S., Okyay S., Kartal Y., 2021). Furthermore, in today's context, recommendation systems have gained significant importance within e-commerce platforms containing vast datasets of big data. Over time, these platforms have developed the need to offer personalized services to users. Studies have also been conducted in this field utilizing recommendation systems (Utku A., Akcayol M. A., 2018). In the academic realm, it's crucial to avoid using sentences and methods from previous works, both in thesis research and academic article writing. This requirement necessitates conducting extensive literature searches to determine if similar studies exist, which can be time-consuming. To address this issue, in one of the studies conducted in the literature, a recommendation system was developed using text mining to suggest journals (Kartal E., Emre İ. E., Özen Z., 2018). In another field, namely the finance sector, where big data is extensively used, providing accurate recommendations to customers is of great significance. In one of the studies, hyperparameter optimization was performed with the aim of delivering effective outcomes to customers (Pınar M., Okumuş O., Turgut U. O., Kalıpsız O., Aktaş M. S.). Recommendation systems have provided solutions not only in online platforms like the internet but also in the real world for various problems. For instance, in today's context, the negative impacts of global warming and climate change on agriculture are evident. In one of the studies, a recommendation system was developed using sensors to measure and analyze soil conditions, incorporating smart farming techniques to address these challenges (Özer B., Kuş S., Yıldız O., 2022). In enhancing the success of recommendation systems, thesis studies utilizing artificial intelligence have also been conducted. In one such study, the performance of two algorithms (Naive Bayes - Complementary Naive Bayes) was compared, followed by normalization. Subsequently, user-based recommendation systems were developed, and the performance of predictions based on users' similar attributes was measured. Artificial neural networks were then employed to further improve performance and enhance prediction accuracy (Şeref B., 2021). In one of the AI-based recommendation systems, a personalized platform was developed in tourism marketing by analyzing users' characteristics and providing a customized experience accordingly (Ercan F., 2020). In another study on movie recommendation systems, the IMDb dataset was used. By using BiLSTM and GRU deep neural networks together, the ratings of the movies by the users were estimated. Some of the features used in the data set included actors, director and movie genre (Türk V., Aydılek İ. B., 2021). In one of the proposal studies in e-commerce applications, suggestions were made by considering parameters such as similar products, customer comments and ratings (Şahiner G., 2019). At the IDAP congress, the similarity of authors' papers was determined using a social network similarity method. The similarity ratios of authors working on similar topics were calculated using Euclidean, Jaccard, and Cosine similarity methods, with Jaccard proving to be the most successful method among them (Öztemiz, F. & Karıcı, A., 2020). In another study, for academic research purposes, articles closely related to the research text were recommended based on keywords to find the most relevant articles in the field. The user's past research was also taken into consideration before making recommendations (E. Deniz, V. K. Öz, S. Bozkurt Keser, S. Okyay, and Y. Kartal, 2021).

In our mentioned study, the aim was to recommend movies that are most relevant to the user's interests, focusing on movie platforms. Firstly, a graph data structure was designed, where nodes represented movies and edge weights indicated the number of common features between connected movies. Then, by considering at least one movie the user had previously watched, an examination was conducted on movies that were most relevant to the user's interests. During the comparison, parameters such as category, release year, duration, language, and city were taken into account. The movie with the highest similarity among movies connected to the one the user watched was calculated using the Malatya

centrality algorithm on the graph data structure, and a recommendation was made (Karci, A., Yakut, S. & Öztemiz, F., 2022).

The continuation of the article consists of four sections: Method, Analysis and Conclusion. In the "Proposed Method" section, the proposed method is discussed, presenting a flowchart and pseudo-codes. The "Analysis" section exemplifies the proposed method and provides experimental results. In the "Conclusion" section, the performance of the applied method and its comparison with other algorithms are presented based on the analysis.

2. Proposed Method

A graph data structure was employed in the design of the recommendation system, with movie information stored within this structure. Each node represented an individual movie, and the edge weights indicated the level of similarity between movies. For example, in the following graph with two nodes, if the edge weight is five, it signifies the count of shared attributes between the movies.

In this particular instance, the two movies share common attributes such as category, language, city, point, and director. Therefore, the total count of shared attributes is five, resulting in an edge weight of five.

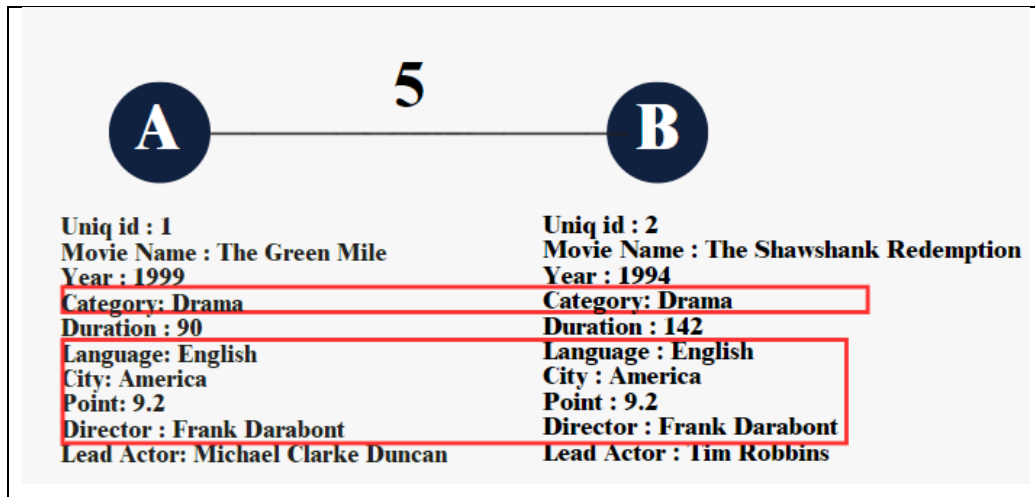


Figure 1: Node and Edge Structure on the Graph

After constructing the graph data structure, we obtained information from the viewer about at least one movie they have previously watched. We are aware that this movie exists on the same platform, thus it exists within the graph data structure. By examining the neighboring edge weights of the movie the viewer watched, the Malatya centrality algorithm was applied. The main goal here is to achieve centrality in order to recommend the most similar movie. When assessing similarity, various parameters of the movie were taken into account: year, category, duration, language, and city – a total of five parameters. Centrality was calculated using the Malatya centrality algorithm with the following method.

$$\Sigma \left(\frac{\text{The edge weight of the movie the viewer watched}}{\text{The edge weight of neighboring movies}} \right)$$

The pseudo code for the Malatya centrality algorithm is provided in Figure 2. (Karci, A. , Yakut, S. & Öztemiz, F., 2022)

```

1. G:(V,E) // G graph
2. MalatyaCentralityMethod <- function (g){ //Malatya algorithm is defined
3. VertexList <- c(V(g)) // Throw vertex from graph to array
4. for (i in VertexList ) // Work as many vertexes in the array
5. Vdegree <- degree ( g,v = V(g)[i]) // Calculate the node degree of the
corresponding vertex
6. AdjacentDegree <- degree ( g,v = neighbors ( g,v = V(g)[i])) // Calculate the
node degree of the neighbors of the relevant node
7. Value <- Vdegree / AdjacentDegree // Degree of related node / degree of the
adjacent node
8. MalatyaCentralityValue <- print ( paste (V(g)[i], sum (Value)), digits = 3) // New
centrality value results
9. return ( MalatyaCentralityValue ) // Returns the maximum values of the graphs
10. }

```

Figure 2: Pseudo code of the Malatya centrality algorithm

The flow diagram related to the created recommendation system is provided in Figure 3.

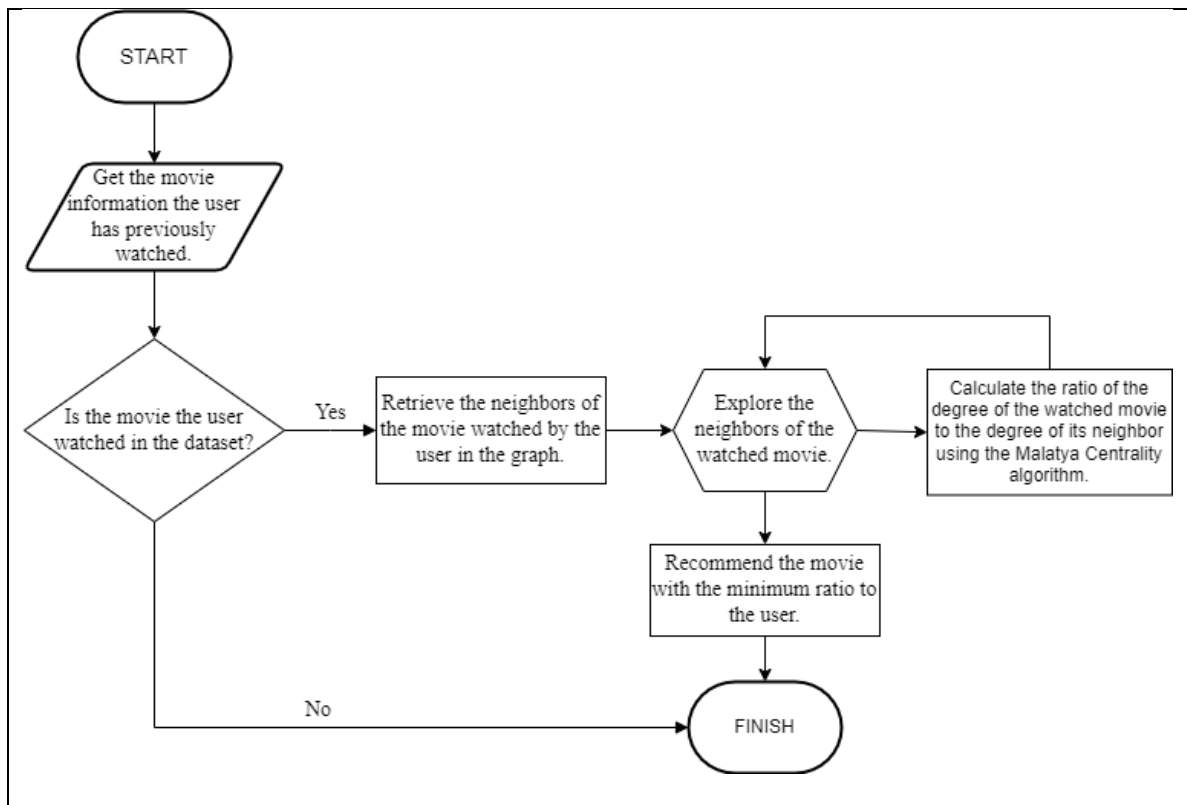


Figure 3: Flow diagram of the proposed method.

Firstly, by taking a certain number of movies, a graph data structure was designed where each node represents a movie, and the edge weight represents the number of common features with each connected movie. The attributes of the movies such as title, release year, category, duration, language,

and city were stored. Subsequently, to ensure high accuracy of recommendations, the user was prompted to provide information about at least one movie they have watched in the past. Based on the attributes of this movie, movies with the most similar features were suggested from the movies stored in the graph data structure. Using the Malatya centrality algorithm, the most similar movie was recommended considering the common attributes among the movies stored in the graph data structure.

In the developed recommendation system, the pseudocode shown in Figure 4 was utilized for the implementation of the Malatya centrality algorithm.

1. START
2. Get movie information from user previously watched.
3. If the movie the user watched is in the dataset, go to step 4, otherwise go to step 8.
4. Find the movie the user watched in the graph and retrieve its neighbors in the graph.
5. Iterate through the neighbors.
6. Calculate the ratio of the degree of the watched movie to the degree of its neighbor using the Malatya Centrality algorithm.
7. Recommend the movie with the minimum ratio to the user.
8. FINISH.

Figure 4: Pseudo code for the application of the Malatya centrality algorithm in the Recommendation System.

3. Analysis

The proposed algorithm suggests a novel approach to recommendation systems. By utilizing the Malatya Algorithm, a graph was constructed for the sample data given in Figure 5. Various parameters are employed to characterize and assess these movies. These parameters are used in defining the connections and degrees within the created graph. Using the Malatya algorithm, centrality values for the nodes in this graph were calculated. In this calculation, data for 10 example movies were used to create a sample graph. Each of these movies corresponds to a node, and the number of common features determines the edge weights. The data and parameters utilized in the experimental application are provided in Figure 5.

```
#Movie Dataset
movies = [
    {'name': 'Movie1', 'year': 2020, 'category': 'Action', 'duration': 120, 'language': 'Turkish', 'city': 'Turkey'},
    {'name': 'Movie2', 'year': 2018, 'category': 'Drama', 'duration': 105, 'language': 'English', 'city': 'America'},
    {'name': 'Movie3', 'year': 2019, 'category': 'Comedy', 'duration': 95, 'language': 'Turkish', 'city': 'Turkey'},
    {'name': 'Movie4', 'year': 2007, 'category': 'Drama', 'duration': 110, 'language': 'English', 'city': 'America'},
    {'name': 'Movie5', 'year': 2020, 'category': 'Action', 'duration': 115, 'language': 'Turkish', 'city': 'Turkey'},
    {'name': 'Movie6', 'year': 2019, 'category': 'Comedy', 'duration': 100, 'language': 'Turkish', 'city': 'Turkey'},
    {'name': 'Movie7', 'year': 1995, 'category': 'Comedy', 'duration': 135, 'language': 'Russian', 'city': 'Russia'},
    {'name': 'Movie8', 'year': 2007, 'category': 'Drama', 'duration': 140, 'language': 'French', 'city': 'Fransa'},
    {'name': 'Movie9', 'year': 1995, 'category': 'Comic Book', 'duration': 135, 'language': 'Turkish', 'city': 'Turkey'},
    {'name': 'Movie10', 'year': 2007, 'category': 'Science Fiction', 'duration': 135, 'language': 'Turkish', 'city': 'Turkey'}
]
```

Figure 5: Parameters in which movies are stored in the recommendation system.

Let's take a look at the 10 example movies. Each of these movies corresponds to a node, and the number of common features determines the edge weights. For instance, since Movie1 and Movie5 share common attributes such as release year, category, language, and city, the edge weight will be 4. Based on this, the connections and edge weights between these nodes in the graph are provided in Table 1.

Table 1: Edge weights based on the number of common attributes among movies on the graph.

Related Movie 1	Related Movie 2	Edge weight
Movie1	Movie3	2
Movie1	Movie5	4
Movie1	Movie6	2
Movie1	Movie9	2
Movie1	Movie10	2
Movie2	Movie4	3
Movie2	Movie8	1
Movie3	Movie5	2
Movie3	Movie6	4
Movie3	Movie7	1
Movie3	Movie9	2
Movie3	Movie10	2
Movie4	Movie8	1
Movie4	Movie10	1
Movie5	Movie6	2
Movie5	Movie9	2
Movie5	Movie10	2
Movie6	Movie7	1
Movie6	Movie9	2
Movie6	Movie10	2
Movie7	Movie9	2
Movie7	Movie10	1
Movie8	Movie10	1
Movie9	Movie10	3

When the user enters the movie they have previously watched as Movie1 into the system, the centrality values were calculated using the Malatya centrality algorithm as follows:

Movie3 : 0.17647058823529413

Movie5 : 0.16129032258064516

Movie6 : 0.17647058823529413

Movie9 : 0.17647058823529413

Movie10 : 0.21052631578947367

While making the proposal, the movie with the minimum centrality rate is recommended. The reason for this is that as the number of common features increases, the centrality ratio decreases. Movie1's

neighbor with the most common features is Movie5. And in the above example, it determined the most similar movie as Movie5 and recommended it to the user.

In Figure 6, the relationships of the movies with each other can be seen in the diagram. Here, each node represents a movie. The connections with each other show that they have at least 1 common feature. For example, since Movie7 and Movie8 have no common features, they do not have any connection with each other.

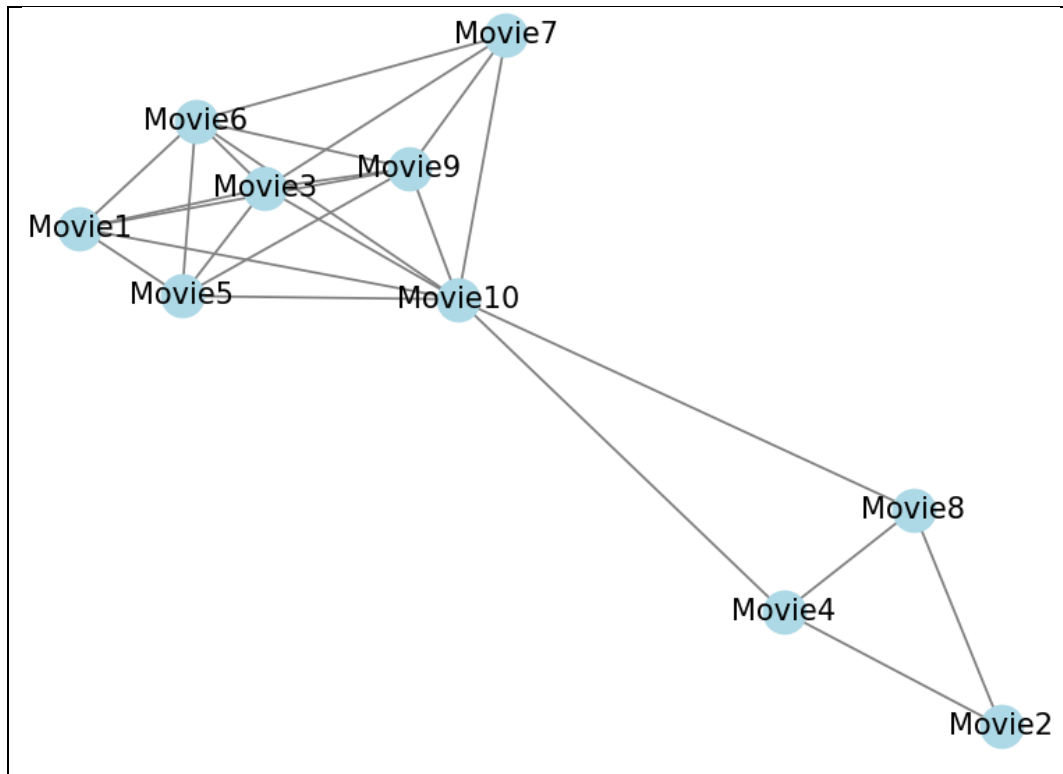


Figure 6: Formed graph structures

4. Conclusion

In this study, a recommendation system was developed using a Graph data structure that offers personalized service to users and recommends the most similar movie based on the movie information obtained from the user, utilizing the Malatya centrality algorithm. In the graph, each node represents a movie, and edge weights denote the number of common features between similar movies. Analysis was conducted based on attributes like release year, category, duration, language, and city of the movies, and this set of parameters can be expanded.

With the use of Malatya centrality algorithm, the most similar movie to the movie that the user has watched in the past was determined. While detecting the similarity, the ratio of the edge weight of the node where the movie watched by the user is kept to the edge weight of the neighboring nodes was calculated. Thus, like the K-NN algorithms, the similarity ratio was not calculated for all nodes, but only the similarity ratio of the most central node to its neighbors. In this way, a more effective and performance recommendation system was created.

Acknowledgement

This study was supported by Malatya İnönü University Scientific Research Projects Management Department (BAP) fund under project ID, 3136 project code, FBG-2023-3136.

Resources

- Park, D. H., Kim, H. K., Choi, Y. and Kim, J. K. 2012. A literature review and classification of recommender systems research. *Expert systems with applications*, 39(11), 10059–10072.
- Schelter, S., Boden, C. ve Markl, V. (2012). Scalable similarity-based neighborhood methods with MapReduce. In *Proceedings of the sixth ACM*.
- Kara B., Kalkan H. (2020). Cloth Combine Estimation System Using Deep Learning.
- Bozkurt M. B., Acı Ç. İ., 2021, Öneri Algoritmalarının Film Önerme Problemi Üzerinde Karşılaştırılması: MovieLens Örneği.
- Ahuja, R., Solanki, A., & Nayyar, A. (2019, January). Movie recommender system using K-Means clustering and K-Nearest Neighbor. In *2019 9th International Conference on Cloud Computing, Data Science & Engineering (Confluence)* (pp. 263-268). IEEE.
- Özgöbek Ö., Erdur R. C., Öneri Sistemleri ve Bir Uygulama Alanı Olarak Haber Öneri Sistemleri.
- Deniz E., Öz V. K., Keser S., Okyay S., Kartal Y., 2021, İçerik tabanlı bilimsel yayın öneri sisteminde benzerlik ölçümlerinin incelenmesi.
- Utku A., Akcayol M. A., 2018, Tavsiye Sistemlerinde Büyük Verinin Kullanımı Üzerine Kapsamlı Bir İnceleme.
- Kartal E., Emre İ. E., Özen Z., 2018, Metin Madenciliği ile Türkçe Bir Dergi Öneri Sisteminin Geliştirilmesi.
- Pınar M., Okumuş O., Turgut U. O., Kalıpsız O., Aktaş M. S., Büyük Veri İçeren Öneri Sistemleri İçin Hiperparametre Optimizasyonu.
- Özer B., Kuş S., Yıldız O., 2022, VERİ MADENCİLİĞİ YÖNTEMLERİ İLE TARIMSAL VERİ ANALİZİ: BİR AKILLI TARIM SİSTEMİ ÖNERİSİ.
- Şeref B., 2021, ÖNERİ SİSTEMLERİNDE BAŞARIMI ARTIRMAK İÇİN YAPAY ZEKA TABANLI YAKLAŞIMLAR.
- Ercan F., 2020, Turizm Pazarlamasında Yapay Zekâ Teknolojilerinin Kullanımı ve Uygulama Örnekleri.
- Türk V., Aydılek İ. B., 2021 IN-DNET İLE FİLM TAVSİYE SİSTEMİ.
- Şahiner G., 2019, Öneri sistemleri ve E- ticarete öneri sistemlerinin kullanımı.
- Öztemiz, F. & Karci, A. (2020). AKADEMİK YAZARLARIN YAYINLARI ARASINDAKİ İLİŞKİNİN SOSYAL AĞ BENZERLİK YÖNTEMLERİ İLE TESPİT EDİLMESİ. *Uludağ Üniversitesi Mühendislik Fakültesi Dergisi* , 25 (1) , 591-608 . DOI: 10.17482/uumfd.533476.
- E. Deniz , V. K. Öz , S. Bozkurt Keser , S. Okyay ve Y. Kartal , 2021 "İçerik tabanlı bilimsel yayın öneri sisteminde benzerlik ölçümlerinin incelenmesi", *Dicle Üniversitesi Mühendislik Fakültesi Mühendislik Dergisi*, c. 12, sayı. 2, ss. 221-228, doi:10.24012/dumf.838084.
- Karci, A. , Yakut, S. & Öztemiz, F. (2022). A New Approach Based on Centrality Value in Solving the Minimum Vertex Cover Problem: Malatya Centrality Algorithm . *Computer Science* , Vol:7 (Issue:2) , 81-88 . DOI: 10.53070/bbd.1195501.

Comparison of Machine and Human Vision based on Brightness and Contrast using YOLOv3 with Fuzzy Logic

Enis Çetin , Zafer İscan 

Department of Electrical and Electronics Engineering, Bahcesehir University, İstanbul, Turkey

(enis.cetin@bahcesehir.edu.tr, zafer.iscan@eng.bau.edu.tr)

Abstract — This study aims to compare human and machine vision using fuzzy logic with YOLOV3 algorithm based on brightness and contrast values in the images. In order to determine detection thresholds for the contrast and the brightness parameters, we conducted a survey including 50 participants. We also examined how these two parameters affect each other depending on the thresholds. The thresholds are normalized, before being given as input to fuzzy logic. We also analyzed the effect of the weight parameter, that represents the size of the object found in the image. In fact, weight affects the sensitivity of the object in the image to the brightness and contrast parameters. We obtain a fuzzy result by processing the brightness, contrast, and weight parameters with fuzzy logic. This result is compared with the output from YOLOV3. In this way, human and machine vision are compared on the same platform. Depending on this comparison, we observed that the human eye performs better in capturing objects in extreme brightness and contrast conditions than the machine. We think that this might be due to some extra information (e.g. experience) involved in human recognition.

Keywords: YOLOV3, Contrast, Brightness, Fuzzy Logic, Image Processing

1. Introduction

In this thesis, we investigated the visual differences between human (Peli, 1990) and machine vision using the parameters of brightness (Firdausy et al., 2007) (Nilizadeh et al., 2022) and contrast (Park & Kim, 2019) (Simone et al., 2012). Humans use an optical perception, while machines use a perception based entirely on numbers. In order to bring these different types of perception into the same environment, we show some digital images to humans and machines and ask them to interpret these images, thus bringing these two different types of vision into the same environment. On this basis, by considering contrast and brightness as parameters, we digitally set visual range boundaries and create the same environment conditions for comparing human and machine vision. These measurements are made and evaluated for high and low brightness or contrast (Tolat et al., 1991) levels of the image. We use fuzzy logic (Zadeh, 1965) for human vision and YOLOV3 (Redmon & Farhadi, 2018) for machine vision.

An object in an image has an area (number of pixels) relative to the image and we call this ratio the weight (Kish, 1990). A higher weight means that the object features in the image are more robust to changing brightness and contrast parameters. While humans have a range of vision based on brightness and contrast, the camera performs image interpretation within a certain range of digital values held in the pixel (Lyon, 2006) (He et al., 2001). Since we are working with multiple factors, when we process human vision with values from a fuzzy logic questionnaire, we use human and machine vision in the same environment. In the comparison, we think that the human has a better visual range than the machine and the reason for this is that the human has more experience.

There are two important points where this work differs from other work in this field. First, we address brightness and contrast with the machine learning algorithm: YOLOV3. With these two parameters, we measure the impact of the machine on object recognition and discuss the relationship between the weight of the object in the image and the brightness and contrast. Second, we combine fuzzy logic with image processing in the field of object recognition for human vision. With a survey of fifty people, we establish

certain visual thresholds in a digital image that are perceived by the human eye. At this point, we use fuzzy logic to establish meaning between thresholds of different variables (brightness and contrast). We determine threshold with YOLO (Redmon et al., 2016) algorithms. YOLO is trained based on CNN (Yang & Guan, 2021), R-CNN (Chen et al., 2017) and FAST R-CNN (Cao et al., 2019) (Girshick, 2015). The COCO dataset (Puri, 2019) was also used. So far, many people using YOLO have worked on the object recognition problem. One of the most important parts of the solution we use in this problem is fuzzy logic (Zadeh, 1999).

2. Methodology

2.1. Adjustment of Brightness and Contrast

1) We create trackbars (slide buttons) to control brightness and contrast. We need to set the range for these trackbars. For brightness we create a range of 511 levels from -255 to 255. Again, for contrast we create a range of 255 levels between -127 and 127. Minus values are the minimum levels, zero is the original image and plus values are the maximum value for brightness and contrast in the image. Figure 1 shows the block diagram of our approach, while Table 1 shows the equation (1)-(2) for brightness and contrast adjustment (<https://www.geeksforgeeks.org/>).

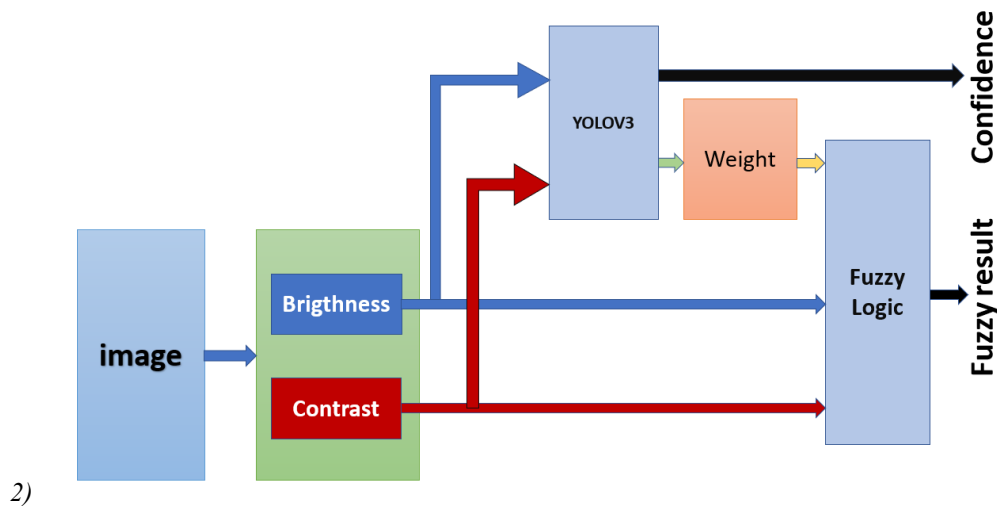


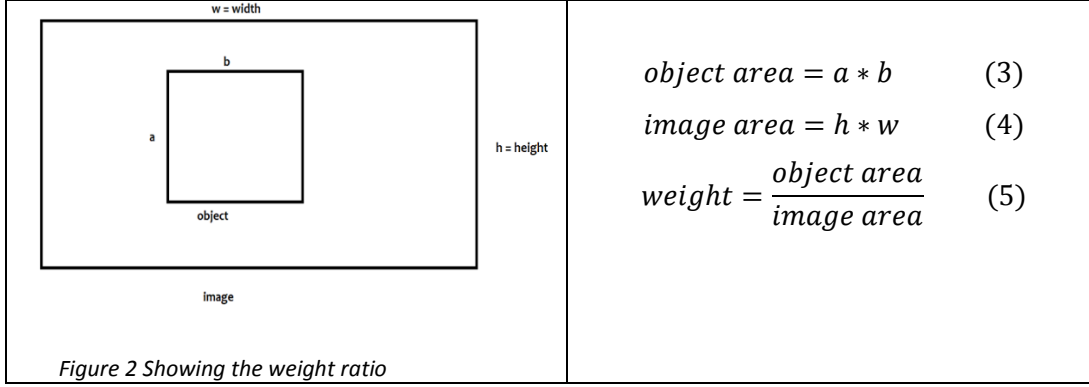
Figure 1 Fuzzy logic and YOLOV3 diagram with brightness and contrast adjustment

Table 1 Equation to adjust brightness and contrast.

$B_{range} = 510$ $B_{min} = -255$ $B_{max} = 255$ Brighthness $= \frac{(B_{adj} - 0) * (B_{max} - (B_{min}))}{(B_{range} - 0) + (-B_{min})} \quad (1)$	$C_{range} = 254$ $C_{min} = -127$ $C_{max} = 127$ Contrast $= \frac{(C_{adj} - 0) * (C_{max} - (C_{min}))}{(C_{range} - 0) + (-C_{min})} \quad (2)$
B_{adj} brightness value from trackbar	C_{adj} Contrast value from trackbar

2.2. Weighting Coefficient

Through equations (1) & (2), we obtain new contrast and brightness values and apply them to the image. We send these values to the YOLOV3 algorithm. The algorithm recognizes an object in the image and draws a square around it. The ratio of this square to the whole image gives us the weight and this is shown in Figure 2 with equations (3)(4)(5).



2.3. Brightness and Contrast Calculation

Brightness is obtained by summing and averaging the values of each of the three channels (RGB) and shown with equation (6). To find the standard deviation we first need the mean. Then we find the contrast using the standard deviation method. In the paper (Zarie et al., 2019), the mean in equations (7) and the standard deviation in equation (8) are calculated.

$$\text{brightness} = \frac{R + G + B}{3} \quad (6)$$

$$\mu = \frac{1}{MN} \sum_{i=0}^{M-1} \sum_{j=0}^{N-1} x_{(i,j)} \quad i = 1,2,3, \dots, M \quad j = 1,2,3, \dots, N \quad (7)$$

$$\sigma = \sqrt{\frac{1}{MN} \sum_{i=0}^{M-1} \sum_{j=0}^{N-1} [x_{(i,j)} - \mu]^2} \quad i = 1,2,3, \dots, M \quad j = 1,2,3, \dots, N \quad (8)$$

2.4. Brightness and contrast normalization

We normalize equations (9) and (10) to remove the mismatch between contrast and brightness.

$$B_{\text{new}} = \frac{\text{brightness}}{B_{\text{range}}} * 100 \quad (9)$$

$$C_{\text{new}} = \frac{\text{Contrast}}{C_{\text{range}}} * 100 \quad (10)$$

2.5. Determine the range of fuzzy input values.

The brightness and contrast thresholds of the images shown to humans are shown in Table 2.

Table 2 Survey results for brightness and contrast thresholds

	Where the object is first distinguishable in the image	The level at which the object in the image is seen with all its features	Where the object in the image begins to lose all its features	Where the object completely disappears in the image
1. image average	1.55	2.45	95.72	97.63
2. image average	1.58	2.48	95.60	97.50
3. image average	1.96	3.07	94.97	96.86
4. image average	1.64	2.70	94.96	97.40
5. image average	1.45	2.44	95.21	97.14
6. image average	2.01	3.15	95.43	97.41
7. image average	1.76	2.90	95.10	97.35
8. image average	2.86	4.04	93.50	96.00
9. image average	1.60	2.62	93.90	97.31
10. image average	2.06	3.03	94.77	96.96
General average	1.85	2.89	94.92	97.16

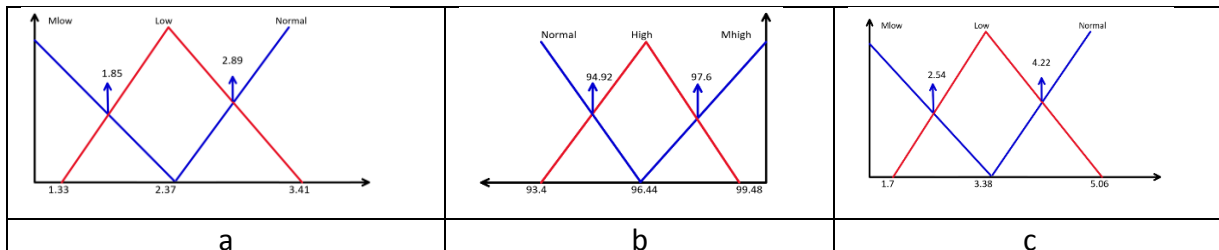
Brightness Average for each step

	Where the object is first distinguishable in the image	The level at which the object in the image is seen with all its features	Where the object in the image begins to lose all its features	Where the object completely disappears in the image
1. image average	2.39	4.10		
2. image average	2.30	4.01		
3. image average	2.76	4.45		
4. image average	2.32	4.09		
5. image average	2.43	3.94		
6. image average	2.32	4.12		
7. image average	2.34	4.05		
8. image average	3.36	5.17		
9. image average	2.48	3.86		
10. image average	2.74	4.20		
General average	2.54	4.20		

Contrast Average for each step

The meaningful equivalents of the lower limits for brightness and contrast are shown in Table 3.

Table 3 Visualization of verbal expressions fuzzy logic for lower (a) or upper (b) limits of brightness and lower bounds (c) of contrast.



The adjusted range for each parameter for fuzzy logic is shown in Table 4 with start and end values and the amount of increase.

Table 4 Adjusted Range For Each Parameter.

	Start value	End value	Step
Brightness	0	101	1
Contrast	0	101	1
Weight	0	1,01	0,01
Fuzzy rate	0	1,01	0,01

In this step, verbal expressions for the brightness, contrast, weight, and fuzzy ratio parameters need to be created and the limits of these verbal expressions for determined. As can be seen in Table 5, expressions and their limits were determined for each parameter. Expressions with 3 thresholds represent triangular and expressions with 4 thresholds represent trapezoidal formation membership. According to the rule list for these formations, it then generates the fuzzy result of each rule. Each fuzzy result is obtained from the membership of the input parameters of the rule to which it belongs. These membership values correspond to a rule in our rule list. This is done by calculating an overall result using the min and max functions. The fuzzy logic rule list is also written in Table 6.

Table 5 Expressive Boundaries For Fuzzy Logic.

parameter	expression	boundaries			
		1. Threshold	2. Threshold	3. Threshold	4. Threshold
Brightness	Mlow	0	0	2.37	
	low	1.33	2.37	3.41	
	normal	2.37	3.41	93.4	96.44
	high	93.44	96.44	98.48	
	Mhigh	96.44	100	100	
Contrast	Mlow	0	0	3.38	
	low	1.7	2.38	5.06	
	normal	3.38	5.06	50	75
	high	50	75	90	100
	Mhigh	75	90	100	100
Weight	small	0	0	1.1	
	medium	0	0.1	0.2	
	large	0.1	0.2	1	1
Fuzzy rate	poor	0	0	0.25	
	bad	0	0.25	0.5	
	average	0.25	0.5	0.75	
	good	0.5	0.75	1.01	
	perfect	0.75	1.01	1.01	

Table 6 Rule List

rule 1	IF	brightness	IS	Mlow	THAN	fuzzy rate	IS	poor								
rule 2	IF	brightness	IS	Mhigh	THAN	fuzzy rate	IS	poor								
rule 3	IF	contrast	IS	Mlow	THAN	fuzzy rate	IS	poor								
rule 4	IF	contrast	IS	Mhigh	THAN	fuzzy rate	IS	poor								
rule 5	IF	brightness	IS	low	AND	contrast	IS	low	AND	weighted	IS	medium	THAN	fuzzy rate	IS	bad
rule 6	IF	brightness	IS	low	AND	contrast	IS	normal	AND	weighted	IS	medium	THAN	fuzzy rate	IS	average
rule 7	IF	brightness	IS	normal	AND	contrast	IS	low	AND	weighted	IS	medium	THAN	fuzzy rate	IS	average
rule 8	IF	brightness	IS	normal	AND	contrast	IS	high	AND	weighted	IS	large	THAN	fuzzy rate	IS	good
rule 9	IF	brightness	IS	normal	AND	contrast	IS	normal	AND	weighted	IS	small	THAN	fuzzy rate	IS	perfect
rule 10	IF	brightness	IS	normal	AND	contrast	IS	normal	AND	weighted	IS	medium	THAN	fuzzy rate	IS	perfect
rule 11	IF	brightness	IS	normal	AND	contrast	IS	normal	AND	weighted	IS	large	THAN	fuzzy rate	IS	perfect
rule 11	IF	brightness	IS	normal	AND	contrast	IS	Mhigh	AND	weighted	IS	small	THAN	fuzzy rate	IS	bad
rule 13	IF	brightness	IS	high	AND	contrast	IS	low	AND	weighted	IS	medium	THAN	fuzzy rate	IS	bad
rule 14	IF	brightness	IS	high	AND	contrast	IS	normal	AND	weighted	IS	medium	THAN	fuzzy rate	IS	average
rule 15	IF	brightness	IS	high	AND	contrast	IS	high	AND	weighted	IS	medium	THAN	fuzzy rate	IS	average

2.6. Limitation

We have two important limitations: First, the eye health of the participants and second, the ambient brightness and screen type. Despite these important limitations, we were able to collect a significant amount of data, both in terms of the number of participants and the number of images, and this amount of data is sufficient to compensate for these limitations.

3. Results

We showed the ten images in the Figure 3 to the participants.

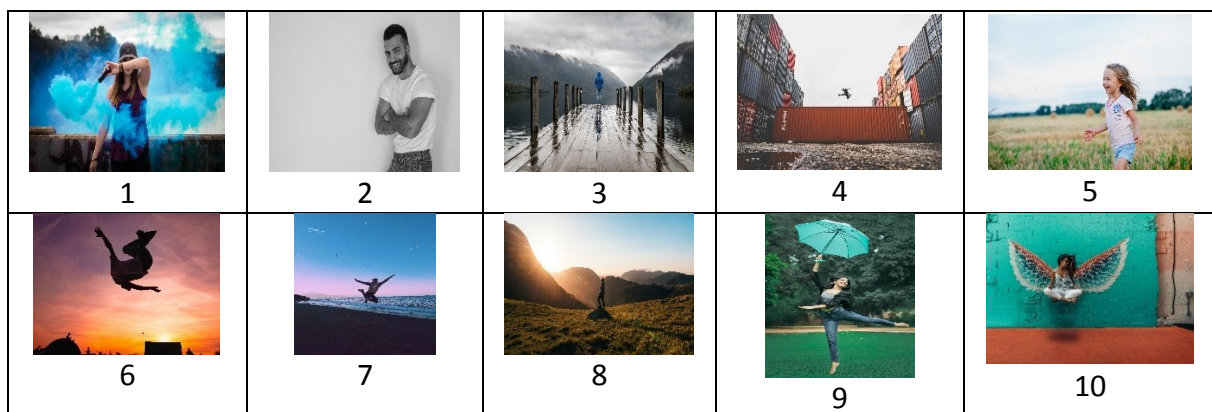


Figure 3 Images shown to participants and their order.

When Table 2 is examined, it can be seen that the average of each image is taken. According to the average values, while the object is recognized as a human in the first image, the average is 1.55 at low brightness. Again, with a brightness value of 97.63, people saw the object at the highest brightness in the first image. Again, according to Table 2, when the visual range between the low mean luminance

value of 2.86 and the high mean value of 96 is analyzed, the narrowest visual range belongs to the image number 8.

Table 7 (a) the lowest brightness level at which an object can be selected in the image at the lowest brightness level, (b) the brightness level at which an object at low brightness can be selected well with its feature, (c) the brightness level at which an object at high brightness can be selected well with its feature, (d) the brightness level at which an object at high brightness can be selected.

Table 7 Distribution of images shown to participants depending on brightness variables.

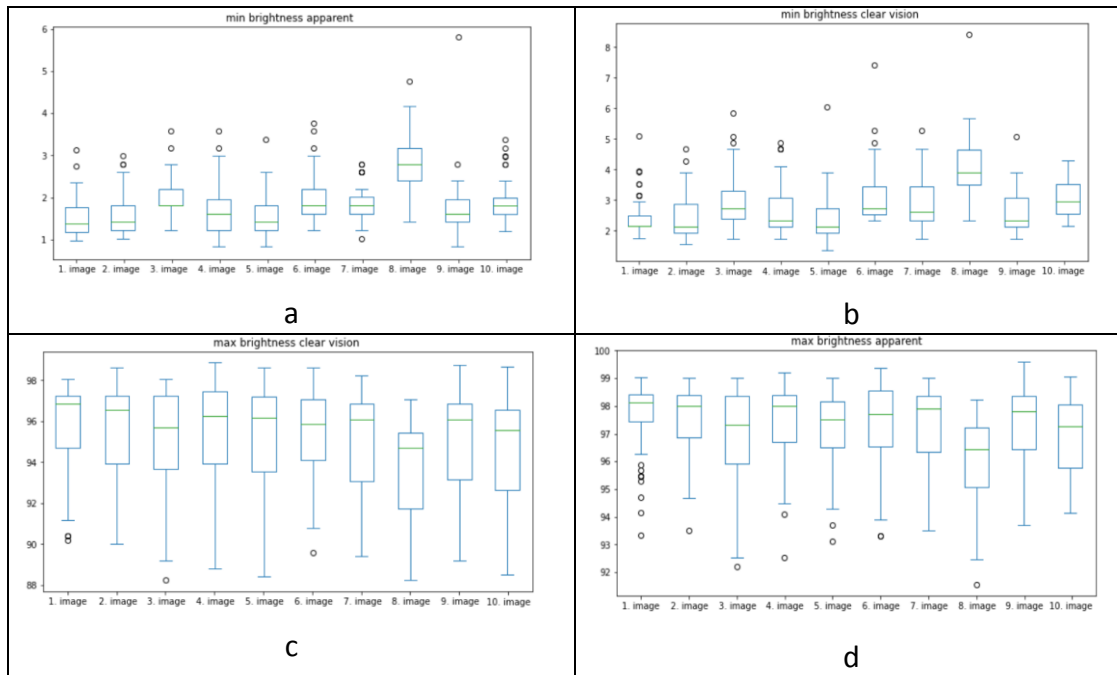
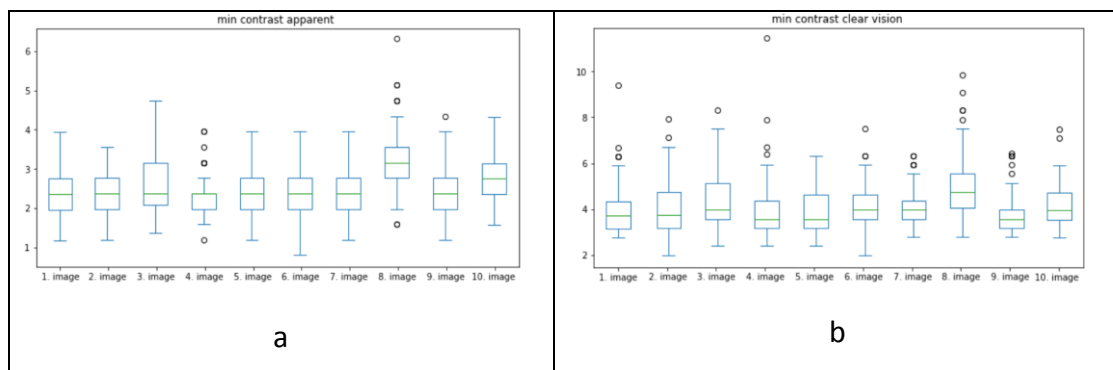


Table 8 Distribution of images shown to participants depending on contrast variables. (a) the lowest contrast level at which an object can be selected in the image at the lowest contrast level, (b) the contrast level at which a low contrast object can be well selected with its feature.



In Figure 4 we have the image back from the YOLOV3 algorithm. As can be seen in the figure, YOLOV3 correctly predicted the person in the image 99.95% of the time.

Figure 5 shows the fuzzy logic result. As can be seen from the figure, the fuzzy logic vision captured the objects with very high accuracy. According to this figure, the input of brightness and

contrast values sends us back a result, which also has a fuzzy verbal meaning. Fuzzy logic uses thresholds that are derived from the results of our survey. Therefore, since this value is taken from people, it expresses the visual limits of human beings. In Figure 6, the fuzzy logic shows the best result, 'excellent'.

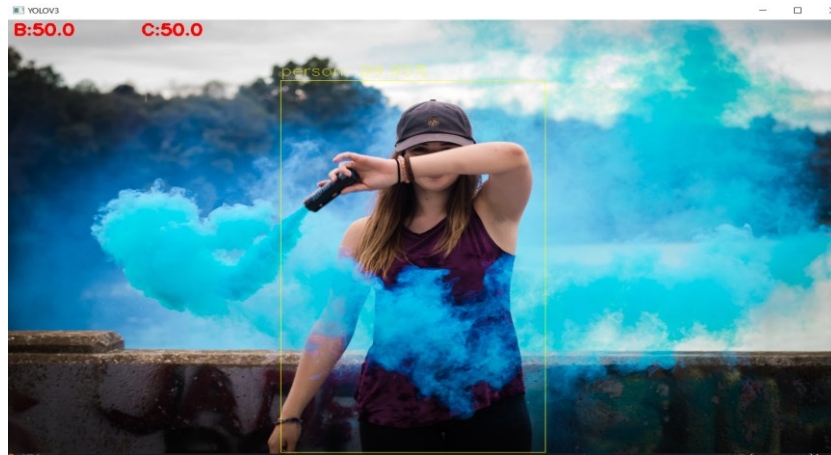


Figure 4 YOLOV3 result window for image.

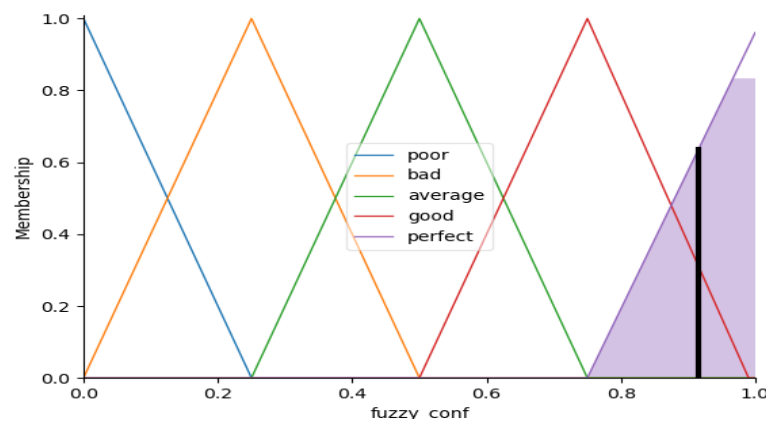


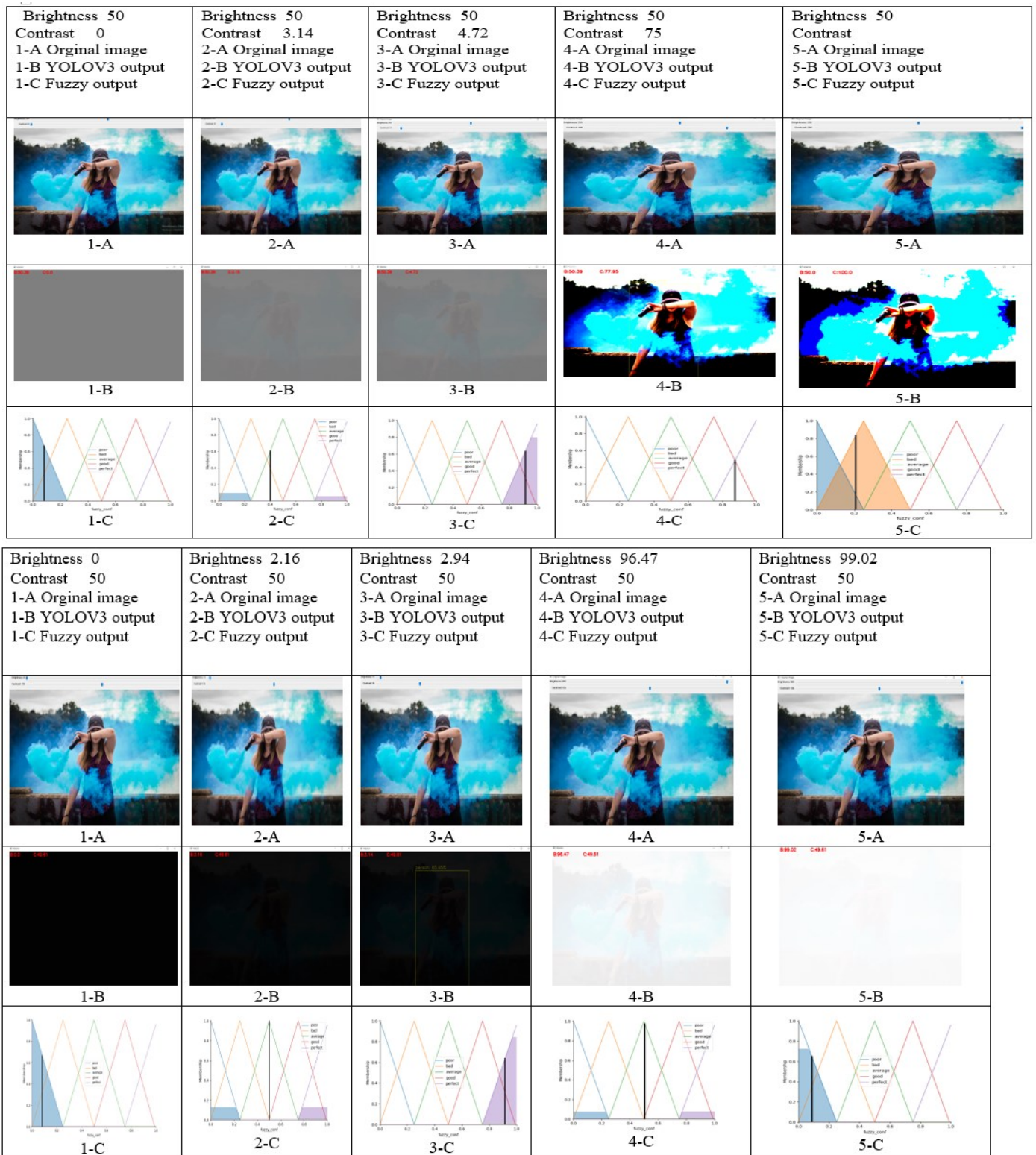
Figure 5 Fuzzy result window of the unit image.

The above values are the results obtained without changing the brightness and contrast of an image. Table 9 below shows the different brightness values under normal contrast and both YOLOV3 and fuzzy logic. The inputs are recorded in the first row of results. The second row holds the original images. The third row contains the YOLOV3 results. The last row contains the fuzzy logic results. There are five columns in the same table. Each column has a different brightness level.

In Table 10, the contrast levels are varied while keeping the brightness constant. When we increase the contrast, object recognition behaves normally for values between 0-50. After fifty, increasing the contrast does not degrade the image much. In an image with maximum contrast, even if the image is bad, if there is an object in the image, that object is recognized by the human. But we cannot say the same for machine vision. Depending on the weight of the object and the appearance of its features, it is slightly recognized by the machine.

Table 9 The effect of brightness under a specific contrast

Table 10 The effect of contrast under a specific brightness



4. Conclusions

In a digital image, the amount of light emitted by each pixel directly affects the brightness of the image, while contrast focuses on contrasting colors in the image. In our tests, low and high brightness and

contrast have a direct impact on both human and machine vision. However, we found that the human has better vision than the machine when considering low or high brightness and contrast values.

Although the weight parameter did not have a significant effect on human vision, it directly affected both the recognition and classification ability of the machine. Especially in the fourth image, the machine recognized an object with a lower weight than the others as a human only at a contrast of around 70 and classified it as a bird at other contrast levels. This is the most obvious effect of image weight on machine recognition.

One of the important points of this study is that it gave us the opportunity to compare human and machine vision in the same environment. The data from the questionnaires is a digitized version of human vision. This data was fed into fuzzy logic to combine two different types of vision in the same environment.

For this reason, it was desired to calculate human vision with the help of a machine and make an inference. Fuzzy logic was preferred for this. Fuzzy expressions are developed by means of fuzzy thresholds and fuzzy results are produced based on them. In this method, we have seen that the fact that the fuzzy set represents human opinion and that the comparison based on it has made an important contribution to this problem. When each image is analyzed individually, it will be seen that the human recognizes the object in the image better than the machine and also makes a much better classification.

If a constant environment is created in the surveys and screen and ambient light variables that directly affect the participants can be controlled, much more stable data will be obtained for fuzzy logic. The YOLOV3 algorithm makes object recognition prediction threshold when it exceeds 50%. As these algorithms improve and the machine's visual threshold is improved, there will be a significant improvement in the machine's visual range based on brightness and contrast. Although we are comparing machine and human, Fuzzy logic can build a perfect bridge between human and machine. Therefore, researchers interested in image processing can benefit from this field.

		brightness level under normal contrast				
		Min Low	Low	Normal	High	Max High
1 Image	Human	NaN	Person	Person	Person	NaN
	Machine	NaN	Person	Person	Person	NaN
2 Image	Human	NaN	Person	Person	Person	NaN
	Machine	NaN	Person	Person	Person	NaN
3 Image	Human	NaN	Person	Person	Person	NaN
	Machine	NaN	Person	Person	Person	NaN
4 Image	Human	NaN	Person	Person	Person	NaN
	Machine	NaN	Bird (F)	Bird (F)	Bird (F)	NaN
5 Image	Human	NaN	Person	Person	Person	NaN
	Machine	NaN	Person	Person	Person	NaN
6 Image	Human	NaN	Person	Person	Person	NaN
	Machine	NaN	Bird (F)	Person	NAN	NaN
7 Image	Human	NaN	Person	Person	Person	NaN
	Machine	NaN	NAN	Person	NAN	NaN
8 Image	Human	NaN	Person	Person	Person	NaN
	Machine	NaN	NAN	Person	NAN	NaN
9 Image	Human	NaN	Person	Person	Person	NaN
	Machine	NaN	Person	Person	Person	NaN
10 Image	Human	NaN	Person	Person	Person	NaN
	Machine	NaN	Person	Person	Person	NaN

a

		contrast level under normal brightness				
		Min Low	Low	Normal	High	Max High
1 Image	Human	NaN	Person	Person	Person	Person
	Machine	NaN	NaN	Person	Person	Person
2 Image	Human	NaN	Person	Person	Person	Person
	Machine	NaN	NaN	Person	Person	NaN
3 Image	Human	NaN	Person	Person	Person	Person
	Machine	NaN	NaN	Person	Person (T) backpack (F)	Person
4 Image	Human	NaN	Person	Person	Person	Person
	Machine	NaN	NaN	Bird	Person	Person
5 Image	Human	NaN	Person	Person	Person	Person
	Machine	NaN	Person	Person	Person	Person
6 Image	Human	NaN	Person	Person	Person	Person
	Machine	NaN	NaN	Person (T) Cat(F)	Cat	Cat
7 Image	Human	NaN	Person	Person	Person	Person
	Machine	NaN	NaN	Airplane(F) Person(T) Sunboard(F)	Bird (F) Person(T) Sunboard(F)	Bird (F) Person(T) Sunboard(F)
8 Image	Human	NaN	Person	Person	Person	Person
	Machine	NaN	NaN	Person(T) NaN Bed (F)	Person (F) Person (F)	Person(F) Person(F)
9 Image	Human	NaN	Person	Person	Person	Person
	Machine	NaN	Umbrella (T) Person (T)	Umbrella (T) Person (T) Person (F) Frisbee (F)	Umbrella (T) Person (T) Person (T)	Umbrella (T) Person (T)
10 Image	Human	NaN	Person	Person	Person	Person
	Machine	NaN	Umbrella (F) Person (T)	Umbrella (F) Person (T)	Umbrella (F) Person (T)	NaN

b

Figure 6 Vision and Classification of Human and Machine at Different Brightness(a)/Contrast(b) Levels

References

- Cao, C., Wang, B., Zhang, W., Zeng, X., Yan, X., Feng, Z., Liu, Y., & Wu, Z. (2019). An improved faster R-CNN for small object detection. *Ieee Access*, 7, 106838–106846.
- Chen, C., Liu, M.-Y., Tuzel, O., & Xiao, J. (2017). R-CNN for small object detection. *Asian Conference on Computer Vision*, 214–230.
- Firdausy, K., Sutikno, T., & Prasetyo, E. (2007). Image enhancement using contrast stretching on RGB and IHS digital image. *TELKOMNIKA Indonesian Journal of Electrical Engineering*, 5(1), 45–50.
- Girshick, R. (2015). Fast r-cnn. *Proceedings of the IEEE International Conference on Computer Vision*, 1440–1448.
- He, Y., Hattori, R., & Kanicki, J. (2001). Improved a-Si: H TFT pixel electrode circuits for active-matrix organic light emitting displays. *IEEE Transactions on Electron Devices*, 48(7), 1322–1325.
- Kish, L. (1990). Weighting: Why, when, and how. *Proceedings of the Survey Research Methods Section*, 121–130.
- Lyon, R. F. (2006). A brief history of “pixel.” In N. Sampat, J. M. DiCarlo, & R. A. Martin (Eds.), *Digital Photography II* (Vol. 6069, p. 606901). SPIE. <https://doi.org/10.1117/12.644941>
- Nilizadeh, A., Nilizadeh, S., Mazurczyk, W., Zou, C., & Leavens, G. T. (2022). Adaptive Matrix Pattern Steganography on RGB Images. *Journal of Cyber Security and Mobility*, 1–28.
- Park, H. J., & Kim, K. B. (2019). Estimation of object location probability for object detection using brightness feature only. *International Journal of Electrical & Computer Engineering (2088-8708)*, 9(6).
- Peli, E. (1990). Contrast in complex images. *J. Opt. Soc. Am. A*, 7(10), 2032–2040. <https://doi.org/10.1364/JOSAA.7.002032>
- Puri, D. (2019). COCO dataset stuff segmentation challenge. *2019 5th International Conference on Computing, Communication, Control and Automation (ICCUBEA)*, 1–5.
- Redmon, J., Divvala, S., Girshick, R., & Farhadi, A. (2016). You only look once: Unified, real-time object detection. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 779–788.
- Redmon, J., & Farhadi, A. (2018). YOLOv3: An Incremental Improvement. <https://pjreddie.com/yolo/>.
- Simone, G., Pedersen, M., & Hardeberg, J. Y. (2012). Measuring perceptual contrast in digital images. *Journal of Visual Communication and Image Representation*, 23(3), 491–506.
- Tolat, A. R., Mcneill, S. R., & Sutton, M. A. (1991). Effects of contrast and brightness on subpixel image correlation. *The Twenty-Third Southeastern Symposium on System Theory*, 604–605.
- Yang, Y., & Guan, C. (2021). A Novel Convolutional Neural Network for Image Classification. *2021 IEEE 3rd International Conference on Frontiers Technology of Information and Computer (ICFTIC)*, 330–333.
- Zadeh, L. A. (1965). Fuzzy sets. *Information and Control*, 8(3), 338–353.
- Zadeh, L. A. (1999). Fuzzy sets as a basis for a theory of possibility. *Fuzzy Sets and Systems*, 100, 9–34. [https://doi.org/https://doi.org/10.1016/S0165-0114\(99\)80004-9](https://doi.org/https://doi.org/10.1016/S0165-0114(99)80004-9)

Kiralık Ev Sektöründe Veriye Dayalı Ürün Öneri Sistemi Geliştirilmesi ve Uygulanması

Development and Application of Data-Driven Product Recommendation System for House Rental Market

Filiz Şenyüzlüler Özçelik^{*1} , Adil Baykasoglu² 

¹Endüstri Mühendisliği Bölümü, Dokuz Eylül Üniversitesi, İzmir, Türkiye

²Endüstri Mühendisliği Bölümü, Dokuz Eylül Üniversitesi, İzmir, Türkiye

(filiz.senyuzlulerozcelik@ogr.deu.edu.tr, adil.baykasoglu@deu.edu.tr)

Özetçe— Çok sayıda kriterin bulunduğu gayrimenkul sektöründe karar vericiye en uygun mülkü önermek karmaşık bir karar verme sürecidir. Karar vericilerin tercihlerini çeşitli şekillerde ifade edebilecekleri, önem ağırlığı verebilecekleri veri güdümlü sistemlere uygulamada rastlanılmamaktadır. Mevcut sistemler genellikle basit sorgulama olanağı sağlayan sistemlerdir. Emlak piyasasında son kullanıcıya yönelik akıllı öneri sistemleri geliştirilmesine ihtiyaç duyulmaktadır. Kullanıcı için en uygun kiralık evi arama sürecinde, kullanıcının her bir kritere verdiği önem ve yapılandırılmış/yapılandırılmamış bilgilerin analizi, sonuçların kalitesini önemli ölçüde etkiler. Mevcut çalışmada, Ağırlıklandırılmış Hiyerarşik Bulanık Aksiyomatik Tasarım (WFAD) yöntemine dayanan internet ortamında çalışabilen kiralık ev öneri sistemi geliştirilmiştir. Geliştirilen sistemde yapılandırılmamış kiralık ev tanımlama verilerinden anlamlı bilgiler çıkarmak için doğal dil işleme (NLP) yöntemleri kullanılmıştır. Ayrıca etkileşimli harita içeren bir web uygulaması aracılığıyla kullanıcıdan tercih ettiği lokasyonları tanımlamasına olanak sağlanmıştır. Alternatif kiralık evler için veriler web-scraping yöntemi kullanılarak elde edilmiş ve modelin etkinliğini göstermek için vaka çalışmaları sunulmuştur.

Anahtar Kelimeler: Bulanık Aksiyomatik Tasarım (WFAD), Öneri Sistemleri, Bilgi Çıkarma, Web Kazıma

Abstract— Finding the most suitable property for the user is a complex problem in the real estate sector, where there are many criteria to be considered in the decision-making process. Another problem is that users are not always able to find systems where they can fully express their preferences. Recommendation systems in the Real Estate Market are smart systems designed to solve this problem. In the process of searching for the most suitable rental house for the user, the importance given to each criterion by the user and the analysis of unstructured information, significantly affect the quality of the results. In the present study, a house rental recommendation system has been developed by using the Weighted Hierarchical Fuzzy Axiomatic Design (WFAD). Natural language processing (NLP) methods are used to extract meaningful information from unstructured house description data. The requirements are taken from the user through the developed web application, including an interactive map for determining the important locations for the user. Data for alternative rental houses is obtained by using the web-scraping method and several case studies are presented in order to exhibit the working of the proposed rental house recommendation system.

Keywords: Fuzzy Axiomatic Design (WFAD), Recommendation Systems, Information Extraction, Web scraping

1. Introduction

Recommendation systems make the online product-finding process simpler and easier for the user. A good product recommendation system should enable end users to express their needs in the best possible way. While developing recommendation systems, various methods have been used by researchers in order to best appeal to the user and make recommendations with high performance. Gharahighehi et al. (2021) categorized a total of 26 papers in the real estate sector into six general methodological approaches, namely collaborative filtering (CF), content-based filtering (CB), Knowledge-Based recommendation systems (KB), reinforcement learning (RL), multi-criteria decision making (MCDM) and hybrid approach (HB). In addition to these methods, Viappiani et al. (2006) presented a preference-based search to reveal user needs and an example-criticism approach to improve these preferences.

The data received from the user during the recommendation process, may not always be structured. In such cases, although the information is available, some alternatives may be overlooked in the product recommendation process. The number of studies, in which such unstructured data are analyzed, is scarcely few in the literature. In the proposed study, the big data of rental house alternatives is obtained by web scraping method. By applying data analysis and information extraction methods on the big data, meaningful information is obtained from the unstructured product description and product title. In the determination of the rental house suggestion list, the weighted hierarchical fuzzy axiomatic design (WFAD) approach is used, which is a good fit for complex multi-criteria decision-making (MCDM) problems. Functional requirements such as the net/brut size, room number, monthly rental price, number of floors, location, and heating are taken from the user by the developed UI and the user is expected to give weights to each of these features.

The rest of this paper is arranged as follows: Section 2, covers an overview of the axiomatic design methodology. Section 3, describes the web scraping technique that is used to obtain the data. Section 4, presents a real-life application of a house rental recommendation system followed by the conclusions in Section 5.

2. Overview of Axiomatic Design (AD) Methodology

AD methodology deals with the transformation of complex customer needs into functional requirements (FRs) and design parameters (DPs) using a design matrix to represent the relationship between the FRs and DPs. Axioms are defined as general principles which are obvious facts that cannot be proven to be correct but do not have counter-examples (Suh, 1990). According to the main axioms, each FR should be independent and the information content of the system should be minimized.

The AD method is assumed as the most suitable method for this study due to the characteristic of the data (mixed data types with fuzzy and interval characteristics) that need to be handled in the multiple criteria decision making process in recommending/searching for rental houses. According to the AD methodology, a feasible design emerges when the design range and the system range overlaps and produces a common range. Figure 1 and Figure 2 illustrate the design and system ranges under crisp and fuzzy decision-making environments respectively (Subulan and Baykasoğlu, 2021).

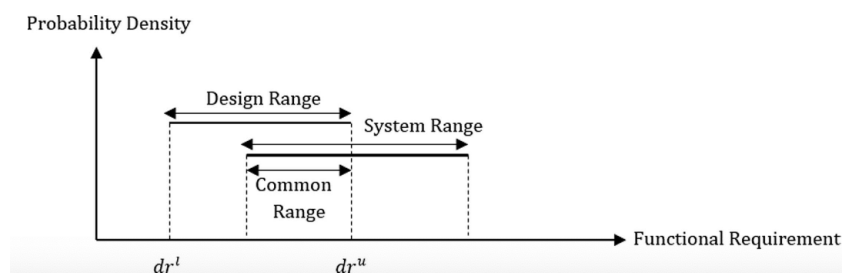


Figure 1. The design and system ranges under crisp decision-making environment

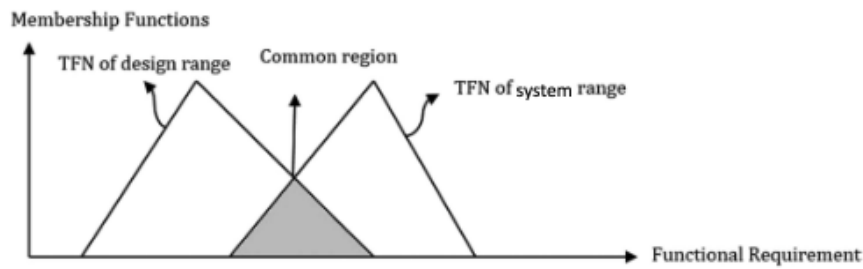


Figure 2. The design and system ranges under fuzzy decision-making environment

When functional requirement (FR_j) has a uniform probability density function, the probability of achieving FR_j can be calculated as follows:

$$p_j = \text{Common Range} / \text{System Range} \quad (1.1)$$

In accordance with the formula (1.1), the information content for any FR_j can be expressed as:

$$I_j = \log_2 (1 / p_j) \text{ or } \log_2 (\text{System Range} / \text{Common Range}) \quad (1.2)$$

The developed system in this study, let the users to use both crisp, interval and fuzzy values while determining their requirements and also allows giving weights to each of them. The weighted information content values are calculated by using the equations presented in the study of Subulan and Baykasoğlu (2021), in which, the shortcomings of the classic AD methodology are modified in the way that benefit type FRs and cost type FRs are also calculated. They mentioned that there may be alternatives with a better performance which is out of the desired range and thus be overlooked in the design. After total information content value is calculated for each alternative rental house in the dataset, the information content values are ranked in an increasing order and presented to the user.

3. Web Scraping

Web scraping (WS) that is also called as web data extraction is a data engineering technique for extracting information from websites with or without consent from the website owner. Specific data is gathered from the web and stored in a database for later retrieval or analysis. The scraping process can be done manually by a software user but is mostly done automatically by an implemented bot or web crawler because of efficiency concerns. Khder (2021) discussed web scraping techniques and approaches in detail in his study.

The most important thing to consider when web scraping is that the scraped data should be public and should not be used for malicious purposes. Kratov and Silva (2018) have created a list of questions that need to be addressed to make a Web Scraping project legal and ethical. Kratov et al. (2020), presented a tutorial on the legality and Ethics of Web Scraping to help researchers to decrease the likelihood of ethical and legal controversies in their work.

In the current study, in order to obtain the rental house data, a web scraper program is developed in Python language by using the BeautifulSoup library. Using this library, firstly the HTML codes of the pages are downloaded, then these codes are broken down and the necessary data is scraped.

4. Application on the House Rental Recommendation System

In order to apply the WFAD methodology to a real-life example, a web application for rental house recommendation, has been developed in Python language by using Streamlit and Folium libraries. The data for rental house alternatives are collected via the web-scraping method and stored in Mongo Atlas DB. The collected data has undergone data cleansing operation in order to make it ready for the information extraction process (see Figure 3).

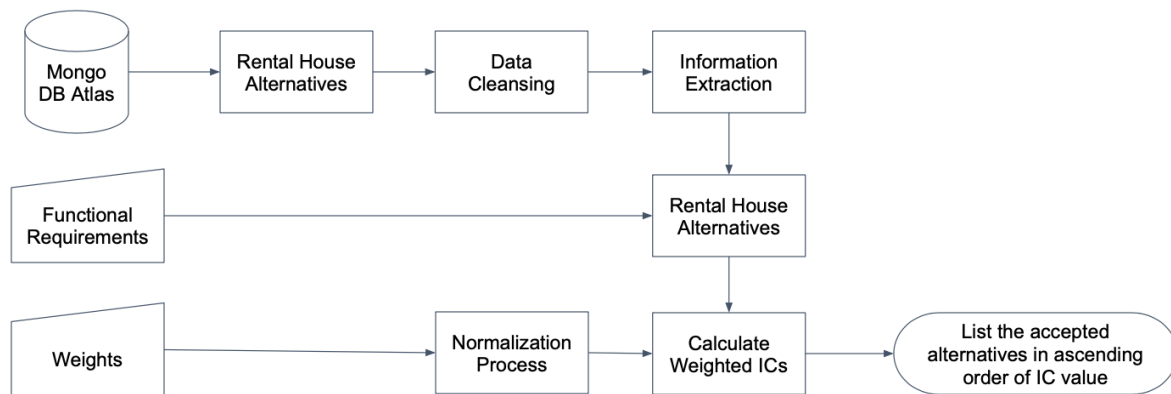


Figure 3. The application procedure of the WFAD Methodology

The functional requirements in the case study have a certain hierarchy within themselves, and this hierarchy is depicted in Figure 4. The system is designed as the user can only assign weights to the features in Criterion Level 1, and assign values and weights for the lower-level criteria.

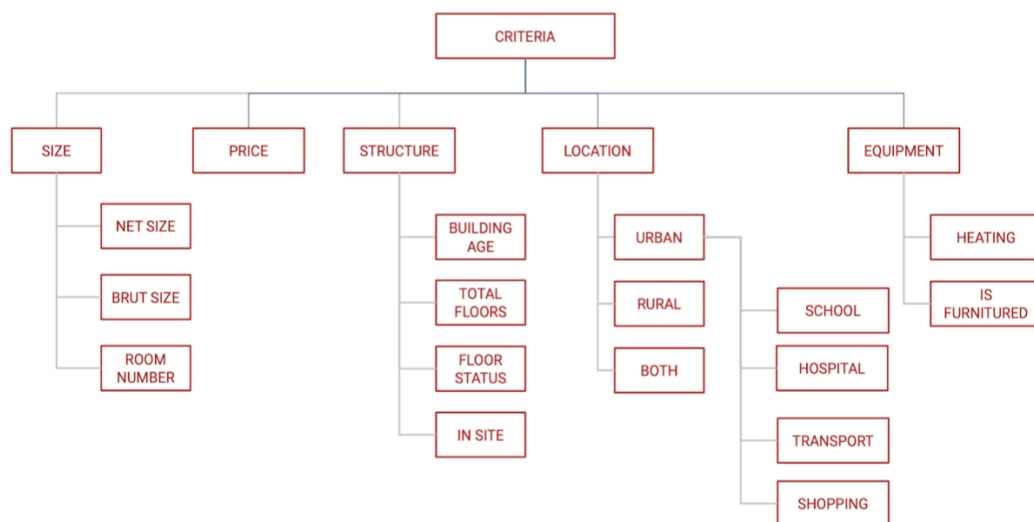


Figure 4. The criteria hierarchy of the case study

Functional requirements are taken from the user via the developed UI as shown in Figure 5. The system also allows the user to assign weights for each search criteria. In this way, the importance given to each feature by the user is determined in a better way, and more meaningful results are aimed to be obtained in the results. After the weights are obtained from the user, the normalized weights are calculated by the system.

The application gives the user the ability to select important locations on the map and assign weights to each of these locations. Then the acceptable distance range is obtained from the user. The rental house alternatives that are going to be recommended to the user, will be in this distance range to the selected locations.

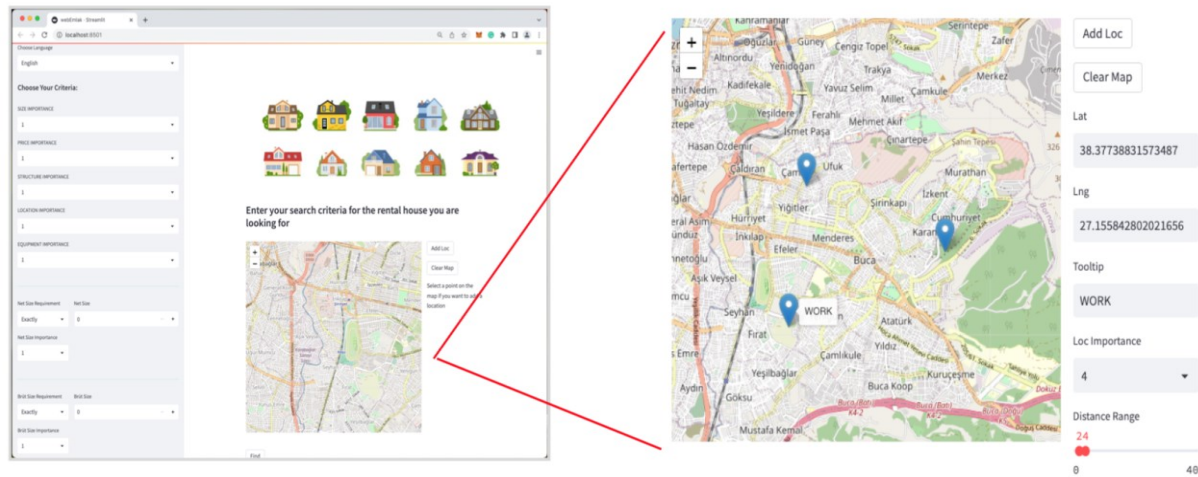


Figure 5. UI of the developed web application for house rental recommendation

Three case studies are performed in order to examine the system behavior when different location data is entered by the user. In the first case study the user mentioned no choice for location. In the second case study the user selects 'Urban' area in which the houses are near to the schools and hospitals. In the third case, to be more specific, the user selects the important locations on a geographic map and enters the acceptable distance ranges. Then the developed recommendation system calculates the overall information content value of all features for any alternative rental house in the system. Finally, the rental houses which are suitable for all the given requirements, are recommended to the user according to the convenience rate, in other words the information content value.

5. Conclusions

In this study, a web based rental house recommendation system that makes use of the improved extension of the weighted hierarchical FAD methodology of Subulan and Baykasoğlu (2021) is developed. Case studies have shown that the adopted FAD methodology fits well with the proposed rental house recommendation system. The study is enriched by the information extraction methods so that the rental house data is analyzed in a better way. As a result, the hidden an unstructured information given for a candidate rental house is converted into valuable data. Additionally, the integrated geographic map makes it available for the user to add important locations to the system to be used in the filtering processes. Thus the system recommends to the user more accurate results that many other recommendation system may ignore.

B) References

- Gharahighehi, A., Pliakos, K., Vens, C. (2021) Recommender Systems in the Real Estate Market—A Survey. *Applied Sciences*. 11: 7502 - 7521
- Khder, M. (2021) Web Scraping or Web Crawling: State of Art, Techniques, Approaches, and Application. *International Journal of Advances in Soft Computing and its Applications*. 13: 145-168

- Krotov, V., Silva, L., (2018) Legality and ethics of web scraping. *Twenty-fourth Americas Conference on Information Systems, New Orleans*.
- Krotov, V., Johnson, L., Silva, L. (2020). Tutorial: Legality and Ethics of Web Scraping. *Communications of the Association for Information Systems*, 47: 555 - 581
- Subulan, K., Baykasoğlu A. (2021) An Improved Extension of Weighted Hierarchical Fuzzy Axiomatic Design. *Sustainable Production and Logistics*, CRC Press, Boca Raton, 321-357.
- Suh, N. P. (1990). Design of thinking design machine. *Annals of the CIRP*, 39(1): 145–149.
- Viappiani, P., Faltings, B., Pu, P. (2006). Preference-based Search using Example-Critiquing with Suggestions. *The Journal of Artificial Intelligence Research*. 27: 465-503

Endüstriyel Siber Güvenlik Sistemleri için Siemens S7-300 PLC Honeypot Tasarımı

Siemens S7-300 PLC Honeypot Design for Industrial Cyber Security Systems

Hayriye TANYILDIZ^{*1} , Canan BATUR ŞAHİN² , Özlem BATUR DİNLER³ 

^{1,2}Yazılım Mühendisliği Bölümü, Malatya Turgut Özal Üniversitesi, Malatya, Türkiye

³ Bilgisayar Mühendisliği Bölümü, Siirt Üniversitesi, Siirt, Türkiye

(hayriye.tanyildiz@tedas.gov.tr, canan.batur@ozal.edu.tr, o.b.dinler@siirt.edu.tr)

Özetçe— Endüstriyel operasyonlarda, Endüstriyel kontrol Sistemleri (ICS) ve nesnelerin İnterneti (IoT) teknolojilerinin benimsenmesi artmaya devam ettikçe, bu sistemleri kötü niyetli siber saldırılardan korumak zorunlu hale gelmiştir.

Bu makale, birçok endüstriyel kontrol sisteminin temel bileşenleri olan S7-300 Programlanabilir Mantık Denetleyicilerine (PLC) izinsiz girişleri tespit etmek için bal küplerinin kullanımına ilişkin bir vaka çalışması sunmaktadır. Honeypot tabanlı izinsiz giriş tespit mekanizmalarının etkinliğini araştırıyor, avantajlarını ve dezavantajlarını tartışıyor ve endüstriyel siber güvenlik bağlamında potansiyellerinden yararlanmak için etkili stratejiler öneriyoruz.

Anahtar Kelimeler : Balküpe, endüstriyel siber güvenlik, S7-300 PLC , Siemens, IoT.

Abstract— As the adoption of Industrial Control Systems (ICS) and Internet of Things (IoT) technologies in industrial operations continues to increase, it has become imperative to protect these systems from malicious cyber attacks.

This article presents a case study of using honeypots to detect intrusions into S7-300 Programmable Logic Controllers (PLCs), which are key components of many industrial control systems. We explore the effectiveness of Honeypot-based intrusion detection mechanisms, discuss their advantages and disadvantages, and propose effective strategies to exploit their potential in the context of industrial cybersecurity.

Keywords : Honeypot, Industiral siber security, S7-300 PLC, Siemens, IOT.

1.Giriş

Endüstriyel Kontrol sistemleri, enerji üretiminden imalata, lojistikten altyapı yönetimine kadar birçok endüstriyel operasyonların desteklenmesinde ve otomatikleştirmesinde çok önemli bir rol oynamaktadır. Bu sistemlerde özellikle programlanabilir Mantık Denetleyiciler (PLC) makinelerin bileşenidir.

Siemens S7-300 sağlamlığı, esnekliği ve endüstri sektörlerinde geniş uygulanabilirliği ile tanınan bir PLC'dir. Diğer birçok PLC çeşitlerinde olduğu gibi S7-300 PLC'ler de endüstriyel süreçleri dijitalleştirmeyi ve ağ oluşturmayı amaçlayan ve Endüstri 4.0 girişimlerinin bir parçası olarak diğer sistemlerle daha fazla birbiri ile iletişim kurabilir bir mekanizmayı içermeye potansiyeline sahip olmaktadır. Bu birbirine bağlanabilirlik, sistemlerde operasyonel verimliliği artırırken, veri erişilebilirliği ve süreç otomasyonu açısından kapsamlı faydalar sağlarken, siber tehditler için potansiyel saldırı yüzeyini de genişletiyor. Sonuç olarak, daha önce yalıtılmış olan bu sistemler siber saldırılara karşı daha savunmasız hale gelmiş ve bu da güvenlikleriyle ilgili endişeleri artırmıştır.

Endüstriyel sistemlere yönelik siber saldırılar ciddi zararlar verme potansiyeline sahip olup, diğer siber tehditlerin aksine gerçek dünyada oldukça fazla zararlara sebebiyet vermektedir. Başarılı izinsiz

girişler, makinelerin yetkisiz kontrolüne, süreçlerin manipüle edilmesine ve hatta kritik altyapının kapanmasına yol açabilmektedir. Bu olaylar, bu sistemlerin kontrol ettiği gerçek dünyadaki fiziksel süreçler göz önüne alındığında, yalnızca önemli ekonomik kayıplara değil, aynı zamanda önemli güvenlik risklerine de neden olabilmektedir.

Bu risklerin farkında olarak, bu kritik sistemleri korumak için siber güvenlik önlemlerinin geliştirilmesine ve uygulanmasına artan bir odaklanma olmuştur. Böyle bir önlem, potansiyel siber izinsiz girişleri tespit etmek ve analiz etmek için bal küpü sistemlerinin kullanılmasıdır. Temelde hedef sistemleri taklit etmek, saldırganları çekmek ve gerçek sistemlermiş gibi onlarla etkileşime geçmek için tasarlanan tuzak sistemler olan bal küpleri, stratejileri ve yöntemleri hakkında değerli bilgilerin toplanmasına olanak tanımaktadır.

S7 -300 PLC için Geliştirilmiş HoneyPot'lar:

Conpot , IEC 60870-5-104, Bina Otomasyonu ve Kontrol Ağı (BACnet), Modbus, s7comm ve HTTP, SNMP ve TFTP gibi diğer protokoller dahil olmak üzere çeşitli endüstriyel protokolleri destekleyen, açık kaynaklı, düşük etkileşimli bir bal küpüdür. Conpot ve Conpot tabanlı bal küpü, araştırmacılar tarafından kullanılan en popüler ICS aldatma uygulamaları arasındadır (Buza ve ark., 2014).

XPOT, programları çalıştırabilen, yazılım tabanlı, yüksek etkileşimli bir PLC bal küpüdür. Siemens S7-300 serisi PLC'leri simüle eder ve saldırganın PLC programlarını derlemesine, yorumlamasına ve XPOT'a yüklemesine olanak tanır. XPOT, S7comm ve SNMP protokollerini destekler ve ilk yüksek etkileşimli PLC bal küpüdür. Yazılım tabanlı olduğu için çok ölçeklenebilir ve büyük tuzak veya sensör ağlarına olanak sağlar.

CryPLH: Siemens Simatic S7-300 PLC'yi simüle eden yüksek etkileşimli ve sanal bir Smart-Grid ICS bal küpüdür. Linux tabanlı bir ana bilgisayarda çalışır ve HTTP(S)'yi simüle etmek için MiniWeb HTTP sunucularını simüle etmek için bir Python betiğini kullanır. CryPLH'nin etkileşim yeteneği, ICS protokollerinin simülasyonundan ICS ortamlarına doğru giderek artmaktadır (Zhang ve ark., 2022).

Bu çalışma, S7-300 PLC'leri korumak için bal küplerinin kullanımına odaklanmaktadır. Bu denetleyiciler için bal küpü özellikli izinsiz giriş tespitinin etkinliğini araştırmayı, bu yaklaşımın faydalarını ve sınırlamalarını değerlendirmeyi ve endüstriyel siber güvenliği geliştirmek için bu aracın kullanılmasına yönelik içgörüler sunmayı hedefliyoruz. Bu keşif, endüstriyel siber güvenlik alanında artan bilgi birikimine katkıda bulunacak ve S7-300 PLC'ler gibi kritik endüstriyel kontrol sistemlerini siber tehditlere karşı korumak için pratik öneriler sağlayacaktır.

(Serbanescu ve ark., 2015) genel ağda açığa çıkan endüstriyel ekipmanın saldırgan için çekiciliğini ve saldırganın davranışını büyük ölçekte düşük etkileşimli ICS bal küpü ayarlayarak analiz etti.

(Buza ve ark., 2014), CryPLH adında saldırganlar tarafından Siemens s7-300 gibi görülecek yüksek etkileşimli bir bal küpü oluşturarak PLC'den yararlanmaya çalışan bir saldırganın gerçekleştirdiği tüm eylemleri günlüğe kaydetmişlerdir. Gerçek cihazda bulunan tüm servisleri inceleyerek Linux tabanlı bir sanal makineye entegre etmişlerdir. Böylece daha sonra bu günlük dosyaları analiz edilerek yeni hedefli saldırılar ortaya çıkarmışlardır.

(Dodson ve ark., 2020), 120 kişilik bir ağda 22 ülkede programlanabilir mantık denetleyicilerini taklit eden yüksek etkileşimli bal küpleri tasarlayarak 80.000 etkileşimin ayrıntılı bir analizini yapmış ve 13 aydan fazla süre yaptıkları çalışmada 9 tanesinin bir endüstriyel protokolü kötü niyetli olarak kullandığını tespit etmişlerdir.

(Yuo ve ark., 2021) çalışmalarında, yüksek maliyet performansı sağlamak için yarı sanal ve yarı fiziksel bir bal küpü tasarımı ve uygulamasını desteklemek için HoneyVP adlı yeni bir bal küpü mimarisi önermiş. ICS cihazlarına yönelik siber saldırıları farklı etkileşim seviyeleri açısından analiz etmiş, daha sonra, bu saldırılarla başa çıkmak için HoneyVP mimarisini sanal bileşen, fiziksel bileşen, ve koordinatör olacak şekilde üç temel bağımsız ve işbirliğine dayalı bileşeni açıkça tanımlamışlardır. Önerdikleri mimariyi kullanmanın önceki bal küpü çözümleriyle karşılaştırıldığında avantajlarını

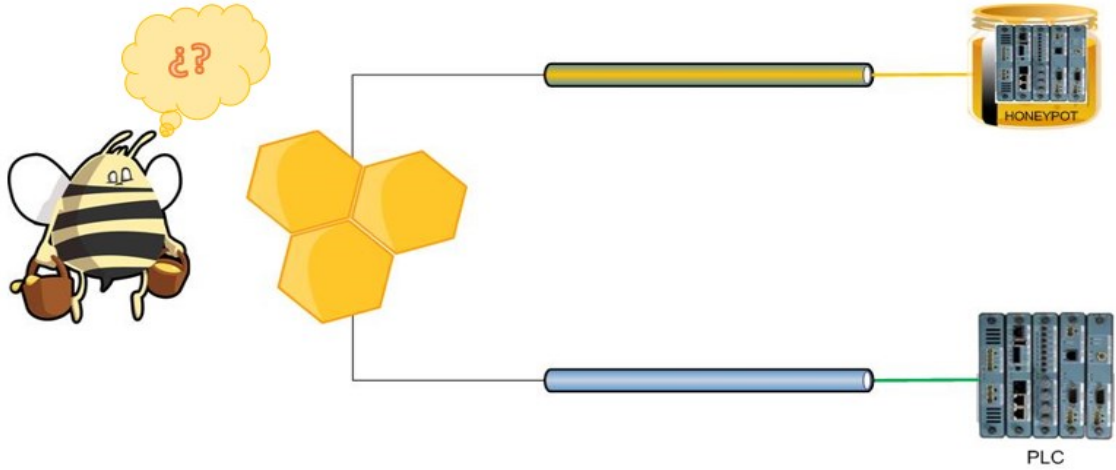
görmüş, HoneyVP'nin, uygun maliyetli bir çözüm sunarak ICS bal küplerini daha çekici hale getirerek ve fiziksel etkileşimleri yakalamayı mümkün kıldığını göstermişlerdir.

2. Genel Bilgiler

2.1. Honeypot Etkin Saldırı Tespiti:

Endüstriyel kontrol sistemine yönelik yoğun olarak gerçekleşen siber saldırılar, kritik ulusal altyapı için ciddi bir tehdit oluşturmaktadır. Bal küpü teknolojisi ile endüstriyel kontrol sistemi için tespit ve saldırı verilerinin yakalanması, ölümcül bir saldırı gerçekleşmeden önce potansiyel saldırganları ve motivasyonlarını ortaya çıkaracak durum farkındalığı yeteneği sağladığı için önemlidir (Xia0 ve ark., 2017).

Bal küpleri, saldırganların kullandığı birçok farklı yöntem ve stratejiyi daha iyi anlamak için kötü amaçlı yazılım yükü gibi gerçek verileri toplamak için oldukça değerli araçlar olduğundan siber güvenlik alanında değerli savunma araçları olarak hizmet eder. Özünde bal küpleri, gerçek sistemlerin özelliklerini ve davranışlarını simüle etmek için titizlikle tasarlanmış ve saldırganlar için gerçek hedeflerle etkileşime girdikleri yanılsamasını yaratan tuzak sistemlerdir. Bir bal küpünün birincil amacı, bir gerçek ağı mümkün olduğunca gerçekçi bir şekilde simüle etmektir, bu da üretim sistemleri, sunucular, hizmetler vb. dahil olmak üzere çeşitli unsurları içerir. Honeynetler, saldırganlar tarafından ele geçirilebilir şekilde tasarlanmıştır ve en iyi şekilde, siber saldırıların tarafından kullanılan teknikler ve araçlar hakkında bilgi toplamayı hedeflemektedir.



Şekil 1. Örnek bir HoneyPot taslağı

Honeypotlar, işlevsellikleri ve etkileşim düzeyleri açısından farklı kategorilere ayrılmaktadırlar. İşlevsellik, honeypotun bir üretim sisteminin parçası olup olmadığını, belirli bir ağ ya da cihaz için güvenlik çözümü olarak mı yoksa saldırganları çekip davranışlarını analiz etme amacıyla mı tasarlandığını belirlemektedir. Etkileşim düzeyi, honeypotun hedef cihazı ne kadar gerçekçi bir şekilde simüle ettiği ve saldırganın karşısındaki sistemin bir honeypot olduğunu fark etme olasılığını ifade etmektedir. Etkileşim seviyeleri genellikle düşük, orta ve yüksek olmak üzere üçe ayrılmaktadır, fakat bu kategorilerin net tanımlamaları bulunmamaktadır. Düşük etkileşimli bir honeypot, giriş ekranını simüle eden ancak daha karmaşık cihaz davranışlarını taklit etmeyen basit bir script olabilir. Öte yandan yüksek etkileşimli bir honeypot, gerçek bir cihaz veya sistem olabilir ve saldırganın sistemdeki

faaliyetlerinin ayrıntılarını kaydetmek için donatılmış olabilmektedir. Bal küpü kavramı, saldırganları geciktirmeyi veya dikkatini dağıtmayı amaçlayan basit tuzaklardan, saldırganları gerçek sistemlerin davranışlarını taklit ederek karmaşık etkileşimlere sokabilen yüksek düzeyde etkileşimli sistemlere kadar çeşitli biçimlerde gerçekleştirilebilmektedir. Siemens S7-300 gibi Programlanabilir Mantık Denetleyicilerini (PLC) koruma bağlamında bal küpü, potansiyel saldırganlara gerçek ve potansiyel olarak savunmasız bir PLC sistemi gibi görünen simüle edilmiş bir PLC ortamı sağlamaktadır. Saldırı tespitinde bal küplerinin değeri, bal küpü ile herhangi bir etkileşimin doğası gereği kötü niyetli olma ihtimalinden kaynaklanmaktadır. Normal şartlar altında, herhangi bir varlığın tuzak sistemle etkileşime girmesi için meşru bir sebep olmamalıdır. Bu nedenle, gözlemlenen tüm etkileşimler güvenli bir şekilde şüpheli veya kötü niyetli etkinlik olarak kategorize edilebilir, bu da meşru ve kötü niyetli eylemleri ayırt etme sorununu basitleştirmektedir- diğer birçok izinsiz giriş tespit yaklaşımında önemli bir zorluktur. Bir saldırgan bal küpü ile etkileşime geçtiğinde, bal küpü sistemi saldırganın IP adresi, kullanılan teknikler ve gerçekleştirdiği eylem dizileri gibi verileri yakalayarak etkinliklerini kaydeder. Bu bilgi, potansiyel olarak yeni saldırı modellerini ve stratejilerini açığa çıkararak saldırganın yöntemleri hakkında değerli bilgiler sağlayabilir. Zamanla, toplanan veriler izinsiz giriş tespit algoritmalarını iyileştirmek, daha sağlam savunma stratejileri geliştirmek ve potansiyel olarak gelecekteki saldırıları tahmin etmek ve önlemek için kullanılabilecek bir tehdit istihbaratı havuzuna katkıda bulunabilir.

Ancak, bal küpü sisteminin korumak için tasarlandığı orijinal sistemlerden izole kalmasını sağlamak çok önemlidir. Uygun izolasyon ve güvenli tasarım olmadan, bir bal küpü, bir saldırganın onu bir tuzak olarak algılaması ve gerçek sistemlerle olan bağlantılarından yararlanmaya çalışması durumunda, yanlışlıkla gerçek sistemlere yönelik saldırılar için bir fırlatma rampası işlevi görebilir. Genel olarak, bal küpü özellikli izinsiz giriş tespiti, bazı zorluklar ve riskler sunsa da, doğru bir şekilde uygulandığında, siber güvenlik cephaneliğinde güçlü bir araç olarak hizmet edebilir, potansiyel tehditlerle etkileşimde bulunmak ve faaliyetlerine ilişkin değerli içgörüler elde etmek için proaktif bir

yol sağlayabilmektedir. Ayrıca, potansiyel saldırıları gerçek sistemlerden uzaklaştırarak ve tehditlerin erken tespitine izin vererek S7-300 gibi PLC'ler için ek bir güvenlik katmanı sunmaktadır.

3. Metot ve Yöntemler

3.1. Reconnaissance Saldırıları:

Sistemin ilk taramasını ve haritalanmasını içermektedir. Saldırganlar genellikle kullanımda olan PLC'lerin türünü, ürün yazılımı sürümlerini ve çalıştırıyor olabilecekleri açık bağlantı noktalarını veya hizmetleri belirlemeye çalışır. Toplanan bu bilgiler daha sonra daha hedefli ve potansiyel olarak yıkıcı saldırıları planlamak ve başlatmak için kullanılır. Keşif saldırılarının, özellikle Siemens S7 300 gibi Programlanabilir Mantık Denetleyicileri (PLC) bağlamında nasıl çalıştığı açıklanmıştır:

PLC'leri Tanımlama: Bir saldırgan için ilk adım, kullanılan PLC türlerini belirlemek olabilir. Bu, ağa bağlı cihazları belirlemek için IP adresi taraması ve nasıl yanıt verdiklerini görmek için cihazlara paketler göndermek dahil olmak üzere çeşitli tekniklerle yapılabilir.

Üretici Yazılımı Sürümlerini Belirleme: Saldırgan, ne tür PLC'lerin kullanımda olduğunu öğrendiğinde, bu PLC'lerde çalışan belirli üretici yazılımı sürümlerini belirlemeye çalışabilir. Bu, cihazların daha ayrıntılı bir şekilde incelenmesi veya belirli ürün yazılımı sürümlerine özgü belirli yanıtların aranması yoluyla yapılabilir. Üretici yazılımı sürümünü bilmek, saldırganların yararlanabilecekleri güvenlik açıklarını belirlemelerine yardımcı olur.

Açık Bağlantı Noktaları ve Hizmetler: Saldırganlar ayrıca PLC'lerde açık bağlantı noktaları ve çalışan hizmetler arar. Açık bağlantı noktaları ağıta bir ağ geçidi sağlayabilirken, çalışan hizmetler kötüye kullanılabilecek ek işlevler sağlayabilir. Bu, bağlantı noktası taraması ve hizmet tespiti için Nmap gibi araçlar kullanılarak gerçekleştirilebilir.

Ağ Haritalama: Saldırganlar genellikle farklı cihazlar arasındaki bağlantıları detaylandıran bir ağ haritası oluşturur. Bu, ağ mimarisini anlamalarına ve potansiyel saldırı noktalarını belirlemelerine yardımcı olabilir.

Güvenlik Açığı Taraması: Saldırganlar, toplanan bilgilere dayanarak, kullanılan sistemlerde, hizmetlerde veya uygulamalarda bilinen güvenlik açıklarını taramak için otomatik araçlar kullanabilir.

Sosyal Mühendislik : Bazı durumlarda, saldırganlar bilgi toplamak için sosyal mühendislik tekniklerini de kullanabilir. Bu, kimlik avı girişimlerini veya hassas bilgileri ifşa etmeleri için kandırmak üzere güvenilir bir kişi veya kuruluş kisvesi altında çalışanlarla doğrudan iletişim kurmayı içerebilir.

Tüm bu adımlar, saldırganların daha sonra hedefli saldırıları planlamak ve başlatmak için kullanabilecekleri sistemi ayrıntılı bir şekilde anlamalarına yardımcı olur. Bu etkinliklerin birçoğunun iyi bir ağ izleme ve izinsiz giriş tespit sistemleriyle tespit edilebileceğini ve önlenilebileceğini not etmek önemlidir. Ayrıca, düzenli güncellemeler ve yamalar, gereksiz portların kapatılması, erişimin sınırlandırılması ve personel eğitimi, başarılı keşif ve müteakip saldırıların önlenmesine yardımcı olabilir.

3.2. Hizmet Reddi (Denial of Service- DOS) Saldırıları:

Siemens S7-300 Programlanabilir Mantık Denetleyicileri (PLC), endüstriyel sistemlerde makine veya prosesleri kontrol etmek için yaygın olarak kullanılmaktadır. Ağa bağlı tüm cihazlar gibi, Hizmet Reddi (DoS) saldırıları da dahil olmak üzere siber saldırılara karşı potansiyel olarak savunmasız olabilirler. Bir DoS saldırısı, bir makineyi, ağı veya hizmeti, bir internet trafiği seline boğarak hedeflenen kullanıcılar için kullanılamaz hale getirmeyi amaçlamaktadır.

Bir S7-300 PLC'ye yapılan bir DoS saldırısı, potansiyel gerçek dünya etkileri nedeniyle özellikle sorunlu olabilir. Bir PLC, bir DoS saldırısı nedeniyle çökerse, kontrol ettiği herhangi bir endüstriyel süreç kesintiye uğrayarak potansiyel olarak önemli finansal kayıplara, fiziksel hasarlara ve hatta güvenlik risklerine yol açabilir. Bir DoS saldırısı gerçekleştirmek için bir saldırganın öncelikle PLC'nin belirli IP adresini bilmesi gerekir. Ardından, PLC'nin gelen ağ trafiğini işleme yeteneğini aşırı yüklemek amacıyla, bu IP adresine çok büyük miktarda veri göndermek için bir tür yazılım aracı kullanılmaktadır.

Saldırı birkaç şekilde gerçekleştirilebilir:

TCP/IP tabanlı saldırılar : Bunlar, mevcut tüm bant genişliğini tüketmek için hedef makineyi rasgele, işe yaramaz verilerle doyurmayı ve meşru paketlerin hedeflerine ulaşmasını engellemeyi içerir.

Uygulama katmanı saldırıları : Bunlar, bir uygulamanın veya hizmetin belirli bir yönünün aşırı yüklenmesini içerir ve bant genişliğinden çok cihazın kaynaklarını tüketmekle ilgilidir.

Protokol saldırıları : Bunlar, PLC'lerin sıklıkla kullandığı Profibus, Profinet, Modbus vb. endüstriyel iletişim protokolleri dahil olmak üzere protokolün kendisindeki güvenlik açıklarından yararlanır.

Büyütme saldırıları : Saldırgan, PLC tarafından işlendiğinde sistemi aşırı yükleyerek çok daha büyük miktarda trafik oluşturan nispeten küçük bir trafik oluşturur.

3.3. Zararlı Komut Saldırıları (Malicious command Attacks):

Zararlı komut saldırıları bir bilgisayar sistemine veya ağa zararlı komutlar göndererek hedeflenen sistemi etkilemeye çalışan bir siber saldırı türüdür. Bu tür saldırılar, sistemdeki güvenlik açıklarından yararlanarak gerçekleştirilebilir. Zararlı komut saldırıları, genellikle kötü niyetli kişiler tarafından yapılır ve amaçları şunlar olabilir:

Yetkisiz erişim sağlamak: Sistemdeki güvenlik açıklarını kullanarak saldırganlar, hedeflenen sistem veya ağa yetkisiz erişim elde etmeye çalışır. Böylece, sistemi kontrol edebilir ve verilere, kullanıcı hesaplarına veya diğer kaynaklara erişebilmektedirler.

Veri çalmak: Saldırganlar, sisteme sızdıktan sonra önemli bilgileri çalmak için komutları kullanabilirler. Bu veriler, kişisel bilgiler, finansal bilgiler, şirket sırları veya diğer hassas bilgiler olabilmektedir.

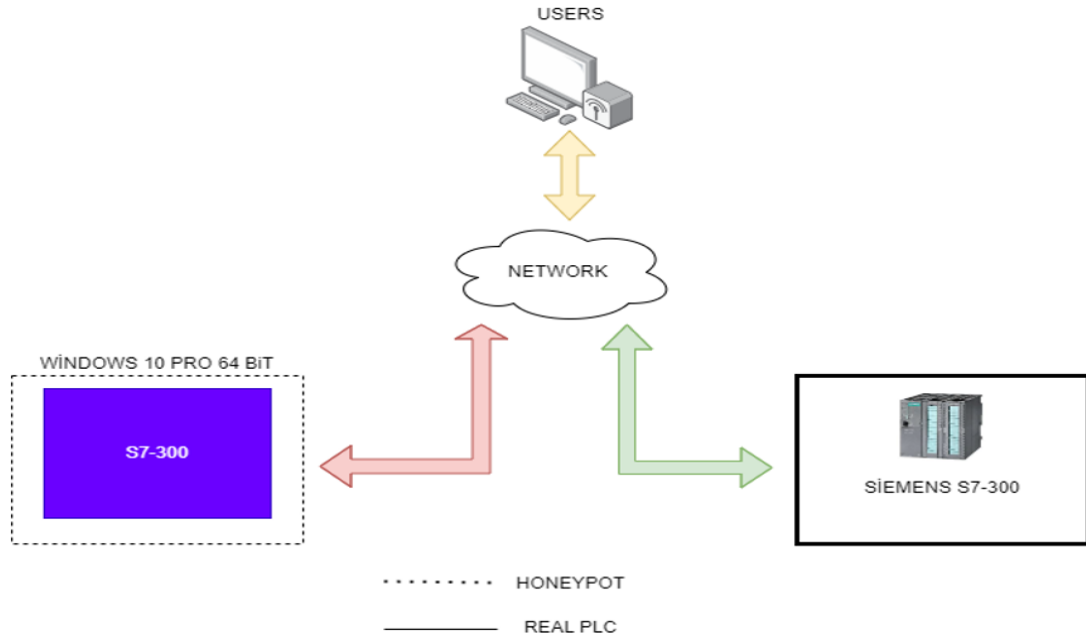
Zarar vermek: Saldırganlar, sisteme zarar vermek için zararlı komutları kullanabilirler. Bu, veri kaybına neden olmak, sistemleri çökertmek veya hizmet kesintilerine yol açmak gibi sonuçlara yol açabilmektedir.

Sistemi istismar etmek: Saldırganlar, hedef sistemi kötü amaçlar için kullanmak için kontrolü ele geçirebilir. Bu tür saldırılar genellikle botnet adı verilen kötü amaçlı bir ağ oluşturmak için kullanılabilmektedir.

4. Metodoloji ve Vaka Çalışmaları

4.1. Bal Küpünün Yüklenmesi ve Saldırıları:

Bu çalışma metodolojik olarak, bir bal küpü görevi görecektir S7-300 PLC cihazlarının simüle edilmiş bir ağının tasarlanmasını ve geliştirilmesini içermektedir. Bu süreçte balküpü gerçek bir S7-300 PLC'lerin işlevselliklerini ve davranışlarını taklit edecek şekilde olması planlanmıştır ve saldırganlara gerçek S7-300 cihazlarından oluşan bir ağ gibi görünen bir ortam oluşturulmuştur. Python kullanılarak sanal bir sunucu oluşturularak S7-300 davranışlarının taklit etmesi sağlanmıştır. Önerilen Balküpü modeli Şekil 2'de gösterilmektedir.



Şekil 2. Önerilen Honeypot Modeli

Tablo 1. Honeypot geliştirme Özellikleri

Taklit Edilen ICS Bileşeni	Geliştirme Dili	İşletim Sistemi
SIEMENS S7-300	Python	Windows 10 Pro

4.2. Vaka Çalışmaları:

Bu bölümde, saldırı seneryoları için istek betikleri oluşturularak honeypot performansı izlenmiştir.

4.1.1. Keşif Saldırıları Vaka Çalışması:

Python Scapy kütüphanesi kullanılarak ARP istekleri ile ağ analizi yapılmıştır. Ağ'da tanımlı cihazların IP taraması ve Mac adresi taraması yapılmıştır.

```

import scapy.all

from scapy.all import ARP, Ether, srp

target_ip = "192.168.1.15/24"
arp = ARP(pdst=target_ip)
ether = Ether(dst="ff:ff:ff:ff:ff:ff")
packet = ether/arp
result = srp(packet, timeout=3)[0]
clients = []
for sent, received in result:
    clients.append({'ip': received.psrc, 'mac': received.hwsrc})
# print clients
print("Available Devices in the Network:")
print("IP" + " " * 18 + "MAC")
for client in clients:
    print("{:16} {}".format(client['ip'], client['mac']))

```

Begin emission:
Finished sending 256 packets.

Received 260 packets, got 3 answers, remaining 253 packets
Available Devices in the Network:

IP	MAC
192.168.1.1	c0:fd:84:ed:9e:00
192.168.1.15	b0:7d:64:de:ee:a5
192.168.1.2	d0:03:df:65:4a:0c

Şekil 3. Keşif Saldırısı Ağ tarama Betiği

```

import socket

target_host = "192.168.1.15"
target_port = 102
recon_request_threshold = 1

s = socket.socket(socket.AF_INET, socket.SOCK_STREAM)

for i in range(recon_request_threshold + 1):
    try:
        s.connect((target_host, target_port))
        request = 'GET / HTTP/1.1\r\nHost: {}\r\n\r\n'.format(target_host)
        s.send(request.encode())
        s.close()

```

Şekil 4. Keşif Saldırısı TCP/Get istek Betiği

4.1.2. Dos Saldırıları Vaka Çalışması :

Bulunan IP adreslerine TCP/Get isteği gönderilerek cihazların varlığının tespit edilmesi amaçlanmıştır.

```

import socket
import struct

def generate_read_request(db, address):
    request = b'\x03\x00\x00\x1f\x02\xf0\x80\x32\x01\x00\x00' # Header
    request += b'\x01' # Number of items
    request += b'\x12' # Variable specification
    request += b'\x0a\x10' # Length of item
    request += struct.pack(">H", db) # DB number
    request += struct.pack(">I", address * 8)[1:] # DB address
    print(f"Received response: {response.hex()}")
    return request

def main():
    client = socket.socket(socket.AF_INET, socket.SOCK_STREAM)
    client.connect(("192.168.1.15", 102))

    request = generate_read_request(10, 20) # DB10, DBD20
    client.send(request)

    response = client.recv(1024)
    print(f"Received response: {response.hex()}")

if __name__ == "__main__":
    main()

```

Şekil 5. DOS Saldırı istek Betiği

4.1.3. Zararlı Komut Saldırıları Vaka Çalışması:

PLC üzerinde kullanıcı tarafında okunmaması gereken bir değer için okuma talebi gönderilerek PLC'de yoğun bir ağ trafiği oluşturulmuştur. Zararlı komutların listesi mevcut endüstriyel ağın özel özelliklerine ve ihtiyaçlarına bağlıdır. S7 protokolünde komutların özel bir PDU türüne karşılık gelen kodlar Tablo 2'deki gibidir.

Tablo 2. S7 PDU kodları

PDU Numarası	Kodu
1	Setup communication
2	Read var
3	Write var
4	Read multiple vars
5	Write multiple vars
6	Download block, Download request
7	Upload block, Upload request
8	Download ended, Upload ended

Sistemdeki belirli bir süreç asla "Download block" veya "Upload block" komutlarını kullanmaz, bu komutların kullanılması potansiyel olarak zararlı bir işlemi göstermektedir. Bu komutlarla süreç müdahale edilmesi de honeypot a yapılan bir saldırıyı temsil etmektedir.

```

2023-07-15 03:24:18,196 - INFO - S7-300 honeypot started on 192.168.1.15:102
2023-07-15 03:24:25,845 - INFO - Connection from 192.168.1.15:51757
2023-07-15 03:24:25,846 - INFO - b'GET / HTTP/1.1\r\nHost: 192.168.1.15\r\n\r\n'
2023-07-15 03:24:25,846 - INFO - Received data from 192.168.1.15:51757
2023-07-15 03:24:25,847 - INFO - 1
2023-07-15 03:24:34,871 - INFO - Connection from 192.168.1.15:51758
2023-07-15 03:24:34,872 - INFO - b'GET / HTTP/1.1\r\nHost: 192.168.1.15\r\n\r\n'
2023-07-15 03:24:34,873 - INFO - Received data from 192.168.1.15:51758
2023-07-15 03:24:34,873 - INFO - 2
2023-07-15 03:24:42,919 - INFO - Connection from 192.168.1.15:51760
2023-07-15 03:24:42,920 - INFO - b'GET / HTTP/1.1\r\nHost: 192.168.1.15\r\n\r\n'
2023-07-15 03:24:42,921 - INFO - Received data from 192.168.1.15:51760
2023-07-15 03:24:42,922 - INFO - 3
2023-07-15 03:24:42,923 - WARNING - Reconnaissance attack detected from 192.168.1.15:51760
2023-07-15 03:24:42,923 - WARNING - Received reconnaissance request: GET / HTTP/1.1
Host: 192.168.1.15

2023-07-15 17:13:09,353 - INFO - S7-300 honeypot started on 192.168.1.15:102
2023-07-15 17:13:13,875 - INFO - Connection from 192.168.1.15:54552
2023-07-15 17:13:13,876 - INFO - b'\x03\x00\x00\x1f\x02\xf0\x802\x01\x00\x00\x01\x12\n\x10\x00\n\x00\x00\xa0'
2023-07-15 17:13:13,876 - INFO - 1
2023-07-15 17:13:13,877 - WARNING - DDos attack detected from 192.168.1.15:54552
2023-07-15 17:13:13,878 - WARNING - Anomaly detected: Multiple read requests from 192.168.1.15:54552

2023-07-25 15:59:11,659 - INFO - 2
2023-07-25 15:59:11,659 - WARNING - Download block request Received Malicious command attack detected.

```

Şekil 6. Saldırı Logları

Şekil 6'da bal küpü sistemlerinin, S7-300'ün belirli güvenlik açıklarını kullanan daha karmaşık saldırıları, çeşitli izinsiz girişleri etkili bir şekilde tespit edebildiğini göstermektedir. Ek olarak bal küpü, saldırganlar tarafından kullanılan ve gerçek PLC sistemlerinin güvenliğini artırmak için kullanılabilecek yöntemler ve teknikler hakkında önemli veriler toplayabilmektedir.

5. Sonuç

Bu vaka çalışmasından elde edilen bulgular, bal küplerinin S7-300 PLC'ler gibi endüstriyel kontrol sistemlerinin güvenliğinde değerli bir rol oynayabileceğini göstermektedir. Umut verici sonuçlara rağmen, saldırı tespiti için bal küplerinin kullanımı sınırsız değildir. Bu tür stratejilerin farkında olan saldırganlar, kasıtlı olarak bal küplerinden kaçınabilir veya bunları yanıltıcı bilgiler göndermek için kullanabilmektedir. Ayrıca, bal küplerini gerçek dünya sistemlerini yansıtacak şekilde sürdürmek ve güncellemek, kaynak açısından yoğun olabilmektedir.

Tek başına bir çözüm olmamakla birlikte, diğer siber güvenlik önlemleriyle birlikte kullanıldığında, endüstriyel operasyonların genel güvenlik duruşunu önemli ölçüde artırabilirler. HoneyPotların yapay zekâ ile desteklenmesi ileride çalışmayı hedeflediğimiz konular arasındadır.

Kaynaklar

- Serbanescu, A.V., et al. (2015) ICS threat analysis using a large-scale honeynet. In: Proceedings of the 3rd International Symposium for ICS & SCADA Cyber Security Research. British Computer Society
- Xiao, F., Chen, E., & Xu, Q. (2017). S7commTrace: A High Interactive Honeypot for Industrial Control System Based on S7 Protocol. International Conference on Information, Communications and Signal Processing.

Buza, D. I., Juhász, F., Miru, G., Félegyházi, M., & Holczer, T. (2014). CryPLH: Protecting Smart Energy Systems from Targeted Attacks with a PLC Honeypot. *Smart Grid Security*, 181–192. doi:10.1007/978-3-319-10329-7_12

Dodson, M., Beresford, A. R., & Vingaard, M. (2020). Using Global Honeypot Networks to Detect Targeted ICS Attacks. 2020 12th International Conference on Cyber Conflict (CyCon). doi:10.23919/cycon49761.2020.9131734

J. You, S. Lv, Y. Sun, H. Wen and L. Sun, (2021) HoneyVP: A Cost-Effective Hybrid Honeypot Architecture for Industrial Control Systems," ICC 2021 - IEEE International Conference on Communications, Montreal, QC, Canada, 2021, pp. 1-6, doi: 10.1109/ICC42927.2021.9500567.

<http://conpot.org/> Accessed 25 July 2023

Buza, D.I., Juhász, F., Miru, G., Félegyházi, M., Holczer, T. (2014). CryPLH: Protecting Smart Energy Systems from Targeted Attacks with a PLC Honeypot. In: Cuellar, J. (eds) *Smart Grid Security. SmartGridSec 2014. Lecture Notes in Computer Science()*, vol 8448. Springer, Cham. https://doi.org/10.1007/978-3-319-10329-7_12

Zhang, Yipeng, Min Li, Xiaoming Zhang, Yueying He, and Zhoujun Li. (2022) Defeat Magic with Magic: A Novel Ransomware Attack Method to Dynamically Generate Malicious Payloads Based on PLC Control Logic *Applied Sciences* 12, no. 17: 8408. <https://doi.org/10.3390/app12178408>.

Mesbah, Mohamed, Mahmoud Said Elsayed, Anca Delia Jurcut, and Marianne Azer. (2023) Analysis of ICS and SCADA Systems Attacks Using Honeypots *Future Internet* 15, no. 7: 241. <https://doi.org/10.3390/fi15070241>.

Hadžiosmanović, D., Sommer, R., Zambon, E., Hartel, P.H. (2014) Through the eye of the PLC: Semantic security monitoring for industrial processes. In: *Proceedings of the 30th Annual Computer Security Applications Conference (ACSAC)*



Serhane, A., Raad, R., Susilo, W. et al. (2022) Applied methods to detect and prevent vulnerabilities within PLC alarms code. *SN Appl. Sci.* 4, 127. <https://doi.org/10.1007/s42452-022-05019-7>

Humayun M, Niazi M, Jhanjhi NZ, Alshayeb M, Mahmood S. (2020) Cyber security threats and vulnerabilities: a systematic mapping study. *Arab J Sci Eng* 45(4):3171–3189.

Noorizadeh M, Shakerpour M, Meskin N, Unal D, Khorasani K. (2021) A cyber-security methodology for a cyber-physical industrial control system testbed. *IEEE Access* 9:16239–16253. <https://doi.org/10.1109/ACCESS.2021.3053135>

Güneş Işınım Değerlerinin Uzun Kısa Süreli Bellek Yöntemi İle Tahmin Edilmesi

Solar Irradiance Forecasting Using Long Short Term Memory Method

Duran Bala¹ , Ahmet Dogan^{*2} 

¹ Fen Bilimleri Enstitüsü, Nuh Naci Yazgan Üniversitesi, Kayseri, Türkiye

² Elektrik - Elektronik Mühendisliği Bölümü, Nuh Naci Yazgan Üniversitesi, Kayseri, Türkiye

(duran.bala@ozkoyuncu.com, adogan@nny.edu.tr)

Özetçe— Fotovoltaik (FV) enerji en önemli yenilenebilir enerji kaynaklarından biridir. Fakat FV gücün değeri iklim ve mevsimsel faktörlerle birlikte değişmektedir. Bu nedenle, iklimsel değişikliklerin öngörülemez davranışı, güç çıkışını etkiler ve şebekenin kararlılığı, güvenilirliği ve işleyişi üzerinde olumsuz bir etkiye neden olur. Bu nedenle, FV güç çıkışının doğru bir şekilde tahmin edilmesi önemli bir gerekliliktir. Son yıllarda güneş ışınım ve güç değerlerinin tahmin edilebilmesi için yapay zeka yöntemlerinden sıklıkla yararlanılmaktadır. Yapay zekâ yöntemleri güneş ışınımı ve çevresel parametreleri girdi değişkenleri olarak girdilerden öğrenme yoluyla girdi ve çıktı arasındaki matematiksel ilişkiyi kurmaya çalışır. Bu çalışmada, FV gücün hesaplanmasında ana değişken olan güneş ışınım değerleri, derin makine öğrenimi yöntemi olan uzun kısa süreli bellek (LSTM) ve regresyon algoritmaları olan Gradyan Arttırma (Gradien Boost-GB), Aşırı Ağaçlar (Extra Trees-ET), Karar Ağacı (Decision Tree-DT), Rastgele Orman (Random Forest-RF), K-Komşu (kNeighbour-kNR) yöntemleri ile gerçekleştirilmiştir. Yöntemlerin hata değerleri korelasyon katsayısı (R^2), mutlak hataların ortalaması (MAE) ve hataların karelerinin ortalamaları (MSE) parametreleri açısından karşılaştırılmıştır. LSTM yöntemi GB, ET, DT, RF, kNR regresyon algoritmalarına göre hata hassasiyet parametreleri açısından en düşük hata ile tahmin gerçekleştiren yöntem olmuştur.

Anahtar Kelimeler : Fotovoltaik Güç, Tahmin, Makine Öğrenmesi Algoritmaları

Abstract— Photovoltaic (PV) energy is one of the most important renewable energy sources. However, the value of PV power changes with climate and seasonal factors. Therefore, the unpredictable behavior of climatic changes has an adverse effect on the grid's stability, dependability, and efficiency. Therefore, an accurate forecasting of PV output is an important requirement. In recent years, artificial intelligence methods have been frequently used to predict solar radiation and power values. Artificial intelligence methods try to establish the mathematical relationship between input and output by learning from inputs and improving results by taking solar radiation and environmental parameters as input variables. In this study, solar radiation values, which are the main variables in the calculation of solar energy, is forecasted using long short-term memory (LSTM), which is a deep machine learning method, and Gradient Boost (GB), Extra Trees (ET), Decision Tree (DT), Random Forest (RF), K-Neighbour (kNR) methods which are regression algorithms. The forecasting accuracy of the methods were compared in terms of correlation coefficient (R^2), mean absolute error (MAE) and mean squares errors (MSE) parameters. The LSTM method has been a method that performs with minimum error in terms of accuracy parameters compare to GB, DT, ET, RF, kNR regression algorithms.

Keywords : Photovoltaic Power, Forecasting, Machine Learning Algorithms

1.Giriş

Teknolojinin hızla ilerlemesi ve nüfus yoğunluğunun artması, enerjiye olan ihtiyacı ve talebi artırmaktadır. Bunun yanında, fosil yakıtlar gibi yenilenemeyen enerji kaynaklarının kullanımı, yalnızca

önemli miktarda kirliliğe neden olmakla kalmaz, aynı zamanda sınırlı kaynakların tükenmesine de neden olur. Bu nedenle, yenilenebilir enerji kaynaklarının kullanımı dünya genelinde teşvik edilmektedir. Akıllı güç sistemlerine yenilenebilir enerji entegrasyonunu amaçlayan çalışmalar literatürde yoğun olarak yer almaktadır (Elkazaz et al., 2020; Piazza et al., 2021; Ullah et al., 2019). Yenilenebilir enerji sistemleri, şebekenin her kademesine entegre edilebilir. Yenilenebilir enerji kaynakları arasında, FV sistemler, gelecekte enerji talebini karşılaması beklenen en önemli teknolojilerden biri olarak kabul edilmektedir (Vandeventer et al., 2019). Diğer taraftan, teknolojiye gelişmeler FV sistemlerin maliyetlerinin sürekli olarak düşmesini sağlamaktadır (Mahmoud, 2019). Uluslararası Enerji Ajansı'nın tahminlerine göre FV sistemlerin kurulu kapasitesi, 2027 yılına kadar kömürünü geçerek en büyüğü enerji kaynağı olması öngörülmektedir. Kümülatif güneş enerjisi kapasitesi, neredeyse üç kat artarak dönem boyunca yaklaşık 1500 GW büyümesi beklenmektedir (IEA - International Energy Agency, 2022).

FV sistemlerin elektrik şebekesine entegrasyonu, güneş enerjisinin kesintili ve değişken doğası nedeniyle şebeke yönetimini ve üretim/tüketim dengesinin sağlanmasını zorlaştırır (Akhter et al., 2019). Güneş enerjisi üretiminin kontrol edilemeyen çıkış gücü nedeniyle, gerilim dalgalanmaları, güç kalitesi ve kararlılık sorunları gibi başka sorunları da neden olur (Voyant et al., 2017). Bu nedenle, güneş enerjisi sistemlerinin çıkış gücünün hassas bir şekilde tahmini, elektrik şebekesinin etkili bir şekilde çalışması FV gücün optimal yönetimi ve maliyetlerin düşürülmesi açısından çok önemlidir. Literatürde, FV sistemlerle ilgili doğru tahminler elde etmeyi amaçlayan geniş bir çalışma yelpazesi bulunmaktadır (Ahmed et al., 2020; Yu et al., 2020; Moreno et al., 2020). FV gücünü tahmin etmek için çeşitli tahmin yöntemleri geliştirilmiştir. FV güç üretimini tahmin etmek için kullanılan yöntemler fiziksel, istatistiksel, yapay zeka ve hibrit yöntemler olmak üzere dört ana grupta toplanabilir (Mahmoud, 2019). Bazı kaynaklarda Makine öğrenimi modelleri de istatistiksel yöntemler altında da değerlendirilmektedir (Dogan & Cidem Dogan, 2022). Hibrit yöntemler ise istatistik ve fiziksel ve yapay zeka yöntemlerinin birlikte kullanılması ile elde edilen yöntemlerdir.

Fiziksel yöntemler, sayısal hava tahmin modeli, gökyüzü görüntü modeli ve uydu görüntüleme veya uzaktan algılama modeli olmak üzere üç alt modelden oluşur. Fiziksel yöntemler, atmosferdeki güneş radyasyonunun fiziksel durumu ile dinamik hareketi arasındaki etkileşime dayanmaktadır (Akhter et al., 2019). İstatistiksel yöntemler, tarihsel zaman serisi verileri ile güneş enerjisi çıkışı arasında bir haritalama ilişkisi kurmayı amaçlar. Yaygın olarak kullanılan istatistiksel yöntemler, otoregresif (AR) modeli, hareketli ortalama (MA) modelleri, otoregresif hareketli ortalama (ARMA) modeli ve durağan olmayan zaman serisidir (Kumar et al., 2018). Genel olarak istatistiksel yöntemler, fiziksel yöntemlere göre daha basit bir modelleme sürecine sahiptir (Wang et al., 2020). Yapay zeka yöntemleri, FV panellerin çalışma durumunu ve çevresel parametreleri girdi değişkenleri olarak alarak girdi ve çıktı arasındaki matematiksel ilişkiyi kurmaya çalışır. Yaygın yapay zeka yöntemleri arasında yapay sinir ağları (ANN), destek vektör makinesi (SVM), aşırı öğrenme makinesi (ELM) ve derin öğrenme (DL) yer almaktadır (Wang et al., 2020). Pek çok makale, sinir ağları veya destek vektör regresyonu gibi metodolojileri FV güç tahmininde kullanılsa da, regresyon yöntemleri de FV güç tahmininde kullanılmaktadır (Voyant et al., 2017). (Persson et al., 2017) saatlik FV güç üretiminin tahmini için GB, RF ve hibrid tahmin yöntemleri kullanılmıştır. (Bae et al., 2017), bulut yapısı dahil olmak üzere çeşitli meteorolojik faktörlere dayanan güneş ışıının bir saatlik ileri tahmini için SVM'ye dayalı güneş tahmin şeması önermektedir. (Benali et al., 2019) YSA kullanarak mevsimsel bir çalışma gerçekleştirmiştir ve ilkbahar ve sonbaharda güneş ışıını tahmininin kış ve yaza göre daha az güvenilir olduğunu, çünkü bu dönemlerde meteorolojik değişkenliğin daha önemli olduğu belirtmektedir. (Zhen et al., 2020) ayrı ayrı, evrişim sinir ağları (CNN) ve LSTM'ye dayalı uçtan uca ışıını haritalama modelleri geliştirmiştir.

Bu çalışmada, İzmir'e ait 15 dakikalık aralıklarla bir yıllık meteorolojik veriler güneş ışıma tahmini için kullanılmıştır. Veriler, ilk aşamada analiz edilerek ışıma ile korelasyon oranı 0.5'in üstünde olan verileri girdi verisi olarak kullanılmıştır. Daha sonra makine öğrenmesi algoritmaları olan uzun kısa süreli bellek (LSTM), Gradyan Arttırma (Gradien Boost-GB), Aşırı Ağaçlar (Extra Trees-ET), Karar Ağacı (Decision Tree-DT), Rastgele Orman (Random Forest-RF), k-Komşu (kNR) yöntemleri ile gerçekleştirilmiştir. Elde edilen tahmin sonuçlarının başarısı farklı hassasiyet ölçüm parametreleri kullanılarak karşılaştırmalı olarak değerlendirilmiştir.

2. Fotovoltaik (FV) Sistemler

Güneş enerjisi gibi yenilenebilir enerji kaynakları çevre dostudur ve kolay erişilebilirdir. FV sistemlerinin çıkış gücü büyük ölçüde güneş ışınımı gibi hava koşullarına bağlıdır. FV çıkış gücü (Bhamidi & Sivasubramani, 2020);

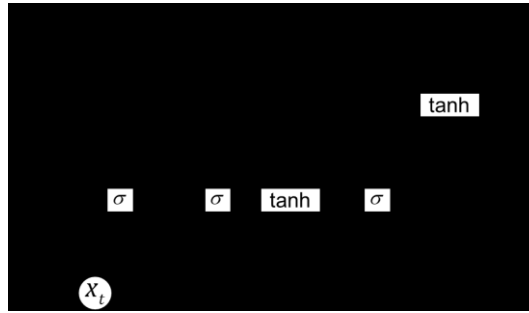
$$P_{FV} = \begin{cases} P_{FVn} * \frac{I^2}{I_{cp} I_{sv}}; & 0 \leq I \leq I_{cp} \\ P_{FVn} * \frac{I^2}{I_{sv}}; & I_{cp} \leq I \leq I_{sv} \\ P_{FVn} & ; I_{cp} \leq I \leq I_{sv} \end{cases} \quad (1)$$

Burada P_{FVn} FV ünitesinin nominal güç çıkışıdır. I_{cp} , genellikle 0,15 kW/m²'ye ayarlanan belirli bir güneş ışınımı noktasıdır ve I_{sv} , genellikle 1 kW/m²'ye ayarlanan standart güneş ışınımı değeridir.

3. Uzun Kısa Süreli Bellek (LSTM)

Makine öğrenimi modelleri, eğitim veri kümesi tarafından iyi eğitildikten sonra, karar vericiler, tahmin girdi verilerini iyi eğitilmiş modellere besleyerek tatmin edici tahmin çıktı değerleri elde edebilir. Veri ön işleme prosedürü, makine öğreniminde önemli bir rol oynar ve makine öğreniminin performansını verimli bir şekilde artırabilir. Bilgi teknolojisinin donanım ve yazılım alanındaki hızlı gelişimi nedeniyle, makine öğreniminin bir alt alanı olan derin öğrenmeye son zamanlarda ilgi artmaktadır (Moreno et al., 2020). LSTM de bir derin öğrenme yöntemi olarak yenilemeli sinir ağlarının bir alt dalıdır.

LSTM, tarihsel bilgi belleği özelliğine sahiptir ve modelin uzun vadeli bağımlılığını ortadan kaldırmaktadır. Bu yöntemde, etkili bilgi güncellenirken önemsiz bilginin unutulması ile gradyan dağılımı sorunu LSTM ile çözülmektedir. LSTM'nin özyinelemeli bir gizleme katmanı Şekil 7'deki gibi giriş kapısı, unutma kapısı ve çıkış kapısı olmak üzere üç kapı tarafından kontrol edilen bellek modülünden oluşur (Dogan & Cidem Dogan, 2022).



Şekil-1. LSTM yönteminin genel yapısı

Giriş kapısı, hangi bilginin hafızaya eklenip eklenmeyeceğine karar verir. Yeni bilgi için bir ağırlık oluşturur ve bu ağırlığı sigmoid aktivasyon fonksiyonu (σ) ile 0 ile 1 arasına çeker. Denklem (2) değeri önceki birim $t - 1$ 'de h_{t-1} çıkışının belirtir ve t zamanında mevcut x_t girişinin eklenmesiyle üretilir. (Fan et al., 2019; Rahman et al., 2018).

$$i_t = \sigma(w_i \times [h_{t-1}, x_t] + b_i) \quad (2)$$

Unutma kapısı, mevcut hafızadan hangi bilginin silineceğine karar verir. Geçmiş hafıza için bir ağırlık oluşturur ve bu ağırlığı sigmoid fonksiyonu ile ayarlar. Denklem (3) kullanılarak unutulacak bilgiye karar verilir.

$$f_t = \sigma(w_f \times [h_{t-1}, x_t] + b_f) \quad (3)$$

Çıkış kapısı Denklem (4) ile alınan bilgiyi nasıl kullanılacağına karar verir.

$$o_t = \sigma(w_o \times [h_{t-1}, x_t] + b_o) \quad (4)$$

Aynı zamanda *tanh* katmanı tarafından aşağıdakine benzer bir prosedürle \bar{c}_t seçim mesajı alınır;

$$\bar{c}_t = \tanh(w_c \times [h_{t-1}, x_t] + b_c) \quad (5)$$

Hafıza Hücresi, LSTM'in "hatıra" olarak adlandırılan bölümüdür ve uzun vadeli bağımlılıkları korur. Elde edilen hücre durumu \bar{c}_t değerlerinin evrimi Denklem (6)'da verilmiştir.

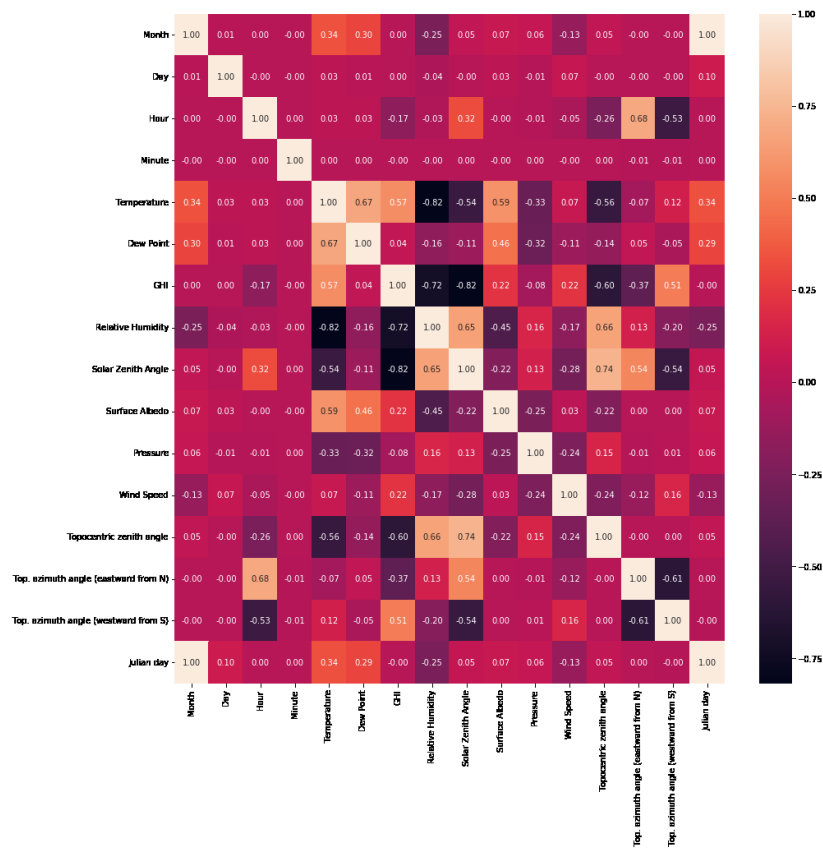
$$c_t = f_t \times c_{t-1} + i_t \times \bar{c}_t \quad (6)$$

c_t , hücre durumu için yeni bir adaydır. Önceki hücre durumu c_{t-1} , LSTM aracılığıyla durum bilgisi olarak çalışan yeni hücre durumuna c_t güncellenir. w_i , w_f ve w_o sırasıyla girdi, unutma, çıktı ve hücre durumunun ağırlıklarıdır. b_i , b_f ve b_o sırasıyla giriş, unutma ve çıkış geçerli hücre durumunun sapmalarıdır.

4. Benzetim Çalışması

Bu çalışmada, FV güç çıkışının hesaplamada kullanılan ışınlam değerlerinin LSTM ile birlikte GB, ET, DT, RF ve kNR regresyon algoritmaları kullanılarak tahmin işlemi gerçekleştirilmiştir. LSTM ve regresyon algoritmaları ile elde edilen sonuçlar birbirleri ile karşılaştırmalı olarak sunulmuştur. Çalışma için İzmir'e ait meteorolojik veriler gerçek veri seti olarak kullanılmıştır (*Solar-Radiation-Dataset*). Veri seti; sıcaklık, direkt küresel ışınlam, çiğ noktaları, direkt normal ışınlam, küresel yatay ışınlam, bağıl nem, solar zenit açısı, basınç, rüzgâr hızı vs. gibi 19 farklı meteorolojik veriyi kapsamaktadır. Veri 1 Ocak Haziran 2019'dan 31 Aralık 2019'a kadar 15 dakikalık periyotlar halindedir. Öncelikle bu veriler içerisinde ışıma ile ilgili olan altı adet veri seti ışıma değeri ile çok yüksek korelasyon sağladığı için simülasyonlarda kullanılmamıştır. Bunlar dışındaki veriler kullanılarak Şekil-2'deki gibi bir ısı haritası oluşturulmuştur.

Korelasyon ısı haritası, veri kümesindeki değişkenler arasındaki korelasyonları görselleştiren bir matrisdir. Bu sayede, veri kümesini daha iyi anlaşılabilir ve gelecekteki analizler için doğru değişkenler seçilebilir. Korelasyon, -1 ile +1 arasında değerler alır. Pozitif korelasyon, bir değişken artarken diğerinin de arttığı durumda oluşurken, negatif korelasyon, bir değişken artarken diğerinin azaldığı durumu ifade eder. 0 korelasyon ise iki değişken arasında bir ilişki olmadığını gösterir. Bu değerler ısı haritasında renk skalası kullanılarak görselleştirilmiştir. Bu çalışmada GHI; yani küresel dikey ışıma ile korelasyon oranı 0.5'in üstünde olan veriler algoritmaya giriş verisi olarak kullanılmıştır. Bu veriler ve korelasyon değerleri Tablo-1'de verilmiştir.

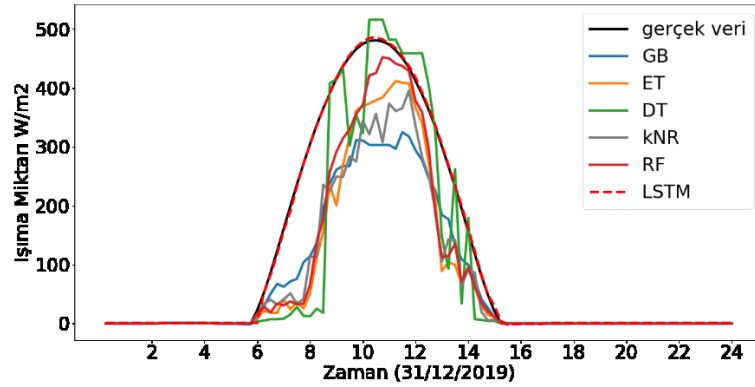


Şekil-2. Veri setlerinin korelasyon ısı haritası

Tablo-1. Işınlam ile korelasyon oranı 0.5'in üstünde olan veriler setleri

Veri adı	Korelasyon değeri
Sıcaklık	0.571595
Bağıl nesi	0.719513
Solar zenit açısı	0.815941
Toposentrik zenit açısı	0.602830
Tepe azimut açısı	0.506268

Kullanılan makine öğrenmesi algoritmalarının tahmin performansını gözlemlemek için burada ortalama korelasyon katsayısı (R^2), ortalama mutlak hatası (MAE) ve ortalama karekök hatası (MSE) gibi üç kriter dikkate alınmıştır. R^2 değeri, bağımlı değişkenin varyantının ne kadarının bağımsız değişkenler tarafından açıklandığını ifade eder. Elde edilen değer 1'e ne kadar yakında modelin o kadar iyi olduğu söylenir. MAE, regresyon analizinde ve tahmin modellerinin performansını değerlendirmek için kullanılan bir hata ölçüsüdür. Bu ölçü, gerçek değerler ile tahmin edilen değerler arasındaki mutlak farkların ortalamasını hesaplar. MSE gerçek değerler ile tahmin edilen değerler arasındaki farkların karelerinin ortalamasını hesaplar. MAE ve MSE'nin düşük olması, modelin gerçek değerlerle tahminleri arasında düşük hata olduğunu gösterir. MAE ve MSE'nin yüksek olması, modelin gerçek değerlerle tahminleri arasında daha büyük hatalar olduğunu ve modelin daha kötü bir performans sergilediğini gösterir.



Şekil-3. GB, ET, DT, kNR RF ve LSTM ile ısıma tahmini sonuçları

1 Ocaktan 31 Aralığa olan veriler eğitim verisi olarak kullanılmış ve 31 Aralık için ısıtım tahmini gerçekleştirilmiştir. Şekil-3’de 2019 yılının son günü olan 31 Aralık tarihi için LSTM ve regresyon algoritmaları ile yapılan ısıtım tahminleri gerçek veri ve birbirleri ile karşılaştırmalı olarak sunulmuştur. LSTM algoritması gerçek veriye çok yakın bir şekilde günlük ısıtım değerlerini elde ederken regresyon algoritmaları aynı başarıyı gösterememiştir.

Algoritmaların tahmin hassasiyetleri R^2 , MAE ve MSE hassasiyet değerleri olarak karşılaştırmalı olarak Tablo-2’de verilmiştir. Buna göre LSTM’in R^2 değeri 0.999 olarak gerçekleşmiş ve gerçek veriye çok yakın bir değer elde edilmiştir. GB, ET, DT, kNR, RF regresyon algoritmalarının R^2 değerleri ise 0.76-0.78 arasında değişmektedir. LSTM yönteminde MAE hassasiyet değeri 2.014 iken, RF, 36.536 ile ikinci en iyi MAE değerine sahiptir. MSE açısından LSTM yine 3.440 değeri ile açık ara en hassas değeri elde ederken, MSE için en büyük hata değeri 83.360 ile DT algoritmasına aittir.

Tablo-2. Algoritmaların tahmin hata değerleri

Algoritmalar	R^2	MAE	MSE
GB	0.779	45.855	81.057
ET	0.773	43.638	82.147
DT	0.766	37.260	83.360
kNR	0.770	45.812	82.733
RF	0.824	36.536	72.353
LSTM	0.999	2.014	3.440

5. Sonuç

Bu çalışmada meteorolojik veriler kullanılarak FV güç hesaplanması için kullanılan güneş ısıtım değerleri tahmini gerçekleştirilmiştir. Giriş verisi olarak ısıtım ile alakalı veriler göz önünde bulundurulmamış diğer meteorolojik veriler dikkate alınmıştır. İlk önce girdi verilerinin çıkış verisi ısıtım ile kolerasyon değerleri ısı haritası ile belirlenmiş ve kolerasyon değeri 0.5’in üzerinde olan değerler girdi olarak kullanılmıştır. Faydalı verileri tutup faydasız verileri silen, kısa ve uzun süreli bellek prensibine göre çalışan LSTM yapay zeka algoritması ile birlikte GB, ET, DT, kNR, RF gibi regresyon algoritmaları ile ısıtım tahmini için kullanılmış ve sonuçlar karşılaştırmalı olarak sunulmuştur. Tahmin hata oranları için R^2 , MAE ve MSE değerleri açısından en uygun değerler açık ara LSTM tarafından elde edilmiştir. Bu durum LSTM yönteminin ısıtım ve güneş enerjisi tahmininde regresyon algoritmalarına göre çok daha başarılı olduğunu göstermektedir.

Kaynaklar

Ahmed, R., Sreeram, V., Mishra, Y., & Arif, M. D. (2020). A review and evaluation of the state-of-the-art in PV solar power forecasting: Techniques and optimization. *Renewable and Sustainable Energy Reviews*, 124(March). <https://doi.org/10.1016/j.rser.2020.109792>

- Akhter, M. N., Mekhilef, S., Mokhlis, H., & Shah, N. M. (2019). *Review on forecasting of photovoltaic power generation based on machine learning and metaheuristic techniques*. 1009–1023. <https://doi.org/10.1049/iet-rpg.2018.5649>
- Bae, K. Y., Member, S., Jang, H. S., Member, S., & Sung, D. K. (2017). *Hourly Solar Irradiance Prediction Based on Support Vector Machine and Its Error Analysis*. 32(2), 935–945.
- Benali, L., Notton, G., Fouilloy, A., Voyant, C., & Dizene, R. (2019). Solar radiation forecasting using artificial neural network and random forest methods : Application to normal beam , horizontal diffuse and global components. *Renewable Energy*, 132, 871–884. <https://doi.org/10.1016/j.renene.2018.08.044>
- Bhamidi, L., & Sivasubramani, S. (2020). Optimal Planning and Operational Strategy of a Residential Microgrid with Demand Side Management. *IEEE Systems Journal*, 14(2), 2624–2632. <https://doi.org/10.1109/JSYST.2019.2918410>
- Dogan, A., & Cidem Dogan, D. (2022). A Review on Machine Learning Models in Forecasting of Virtual Power Plant Uncertainties. *Archives of Computational Methods in Engineering*, 30, 2081–2103. <https://doi.org/10.1007/s11831-022-09860-2>
- Elkazaz, M., Sumner, M., & Thomas, D. (2020). Energy management system for hybrid PV-wind-battery microgrid using convex programming, model predictive and rolling horizon predictive control with experimental validation. *International Journal of Electrical Power & Energy Systems*, 115, 105483. <https://doi.org/10.1016/J.IJEPES.2019.105483>
- Fan, C., Wang, J., Gang, W., & Li, S. (2019). Assessment of deep recurrent neural network-based strategies for short-term building energy predictions. *Applied Energy*, 236(November 2018), 700–710. <https://doi.org/10.1016/j.apenergy.2018.12.004>
- Yu, D., Choi, W., Kim, M., Liu, L. (2020). Forecasting Day-Ahead Hourly Photovoltaic Power Generation Using Convolutional Self-Attention Based Long Short-Term Memory. *Energies*, 13, 4017. doi:10.3390/en13154017
- IEA - International Energy Agency. (2022). *Renewables 2022, IEA, Paris. Analysis forecast to 2027*. 158. <https://www.iea.org/reports/renewables-2022>; License: CC BY 4.0
- Kumar, M., Majumder, I., & Nayak, N. (2018). *Engineering Science and Technology , an International Journal Solar photovoltaic power forecasting using optimized modified extreme learning machine technique*. 21, 428–438.
- Mahmoud, M. A. K. (2019). *Accurate photovoltaic power forecasting models using deep*. 2727–2740. <https://doi.org/10.1007/s00521-017-3225-z>
- Moreno, G., Martin, P., Santos, C., Rodríguez, F. J., & Santiso, E. (2020). *A Day-Ahead Irradiance Forecasting Strategy for the Integration of Photovoltaic Systems in Virtual Power Plants*. 8. <https://doi.org/10.1109/ACCESS.2020.3036140>
- Persson, C., Bacher, P., Shiga, T., & Madsen, H. (2017). Multi-site solar power forecasting using gradient boosted regression trees. *Solar Energy*, 150, 423–436. <https://doi.org/10.1016/j.solener.2017.04.066>
- Piazza, A. Di, Piazza, M. C. Di, Tona, G. La, & Luna, M. (2021). ScienceDirect An artificial neural network-based forecasting model of energy-related time series for electrical grid management. *Mathematics and Computers in Simulation*, 184, 294–305. <https://doi.org/10.1016/j.matcom.2020.05.010>
- Rahman, A., Srikumar, V., & Smith, A. D. (2018). Predicting electricity consumption for commercial and residential buildings using deep recurrent neural networks. *Applied Energy*, 212(October 2017), 372–385. <https://doi.org/10.1016/j.apenergy.2017.12.051>
- Solar-Radiation-Dataset* <https://www.kaggle.com/datasets/ibrahimkiziloklu/solar-radiation-dataset>

- Ullah, Z., Mokryani, G., Campean, F., & Hu, Y. F. (2019). Comprehensive review of VPPs planning, operation and scheduling considering the uncertainties related to renewable energy sources. *IET Energy Systems Integration*, 1(3), 147–157. <https://doi.org/10.1049/iet-esi.2018.0041>
- Vandeventer, W., Jamei, E., Sidarth, G., Seyedmahmoudian, M., Kok, T., Horan, B., & Mekhilef, S. (2019). *Short-term PV power forecasting using hybrid GASVM technique*. 140.
- Voyant, C., Notton, G., Kalogirou, S., Nivet, M. L., Paoli, C., Motte, F., & Fouilloy, A. (2017). Machine learning methods for solar radiation forecasting: A review. *Renewable Energy*, 105, 569–582. <https://doi.org/10.1016/j.renene.2016.12.095>
- Wang, H., Liu, Y., Zhou, B., Li, C., Cao, G., Voropai, N., & Barakhtenko, E. (2020). Taxonomy research of artificial intelligence for deterministic solar power forecasting. *Energy Conversion and Management*, 214(April), 112909. <https://doi.org/10.1016/j.enconman.2020.112909>
- Zhen, Z., Liu, J., Zhang, Z., Wang, F., Chai, H., Yu, Y., Lu, X., Wang, T., & Lin, Y. (2020). Deep Learning Based Surface Irradiance Mapping Model for Solar PV Power Forecasting Using Sky Image. *IEEE Transactions on Industry Applications*, 56(4), 3385–3396. <https://doi.org/10.1109/TIA.2020.2984617>

Infant Cry Classification by Using Adaptive Cepstral Features and Machine Learning Approach

Özkan Arslan^{*1} 

¹Department of Electronics and Communication Engineering, Tekirdağ Namık Kemal University, Tekirdağ, Turkey

(oarslan@nku.edu.tr)

Abstract—This study proposes an infant cry classification (ICC) system using adaptive cepstral features based on empirical mode decomposition (EMD). For this purpose, a publicly available dataset containing a total of 457 data divided into five different categories are used. In the proposed approach, mel-frequency cepstral coefficients and their derivatives, linear prediction coefficients, and linear prediction cepstral coefficient parameters are obtained by the EMD. The synthetic data are produced by applying the SMOTE technique to the feature sets and used in the training of the models. The features are selected with the ReliefF algorithm and used with the support vector machine (SVM) algorithm. Extensive results show that the proposed the adaptive cepstral features based on EMD and SVM-cubic provide 99% accuracy, 98.8% F1-score and 0.985 Matthews correlation coefficient values for the ICC system. It can be clearly said that the proposed approach is a computer-aided decision system that can help parents and healthcare professionals.

Keywords : *Infant cry classification, SMOTE algorithm, empirical mode decomposition, cepstral-domain acoustic parameters, ReliefF algorithm, support vector machine*

1. Introduction

The infant birth rate is increasing globally each year, and the World Health Organization reports approximately 134 million births in 2023. Infants perform communication and emotion transfer actions by cry (Chang et al., 2021). For first-time parents, taking care of newborn is quite difficult. Advice and suggestions from experienced parents may be insufficient to overcome this difficulty in practice. Although baby cry is audibly similar to each other, they contain differences in terms of physiological formation. The parents who have had children before can overcome some difficulties with their knowledge and experience. The distinctive and descriptive features obtained from baby cry signals provide information about various health problems (Ozseven, 2023).

In recent years, research on baby cry has been the subject of numerous studies (Aggarwal et al., 2023; Ji et al., 2021; Kulkarni et al., 2021). Baby cry recognition and classification applications are carried out using classical machine learning approaches consisting of pre-processing, feature extraction and classification stages or deep learning algorithms using direct signals. Acoustic parameters are used extensively for feature extraction in classical machine learning applications. Numerous studies use the following acoustic parameters extensively: pitch period, formant, jitter, shimmer, mel-frequency cepstral coefficients (MFCC), linear prediction coefficients (LPC) and linear prediction cepstral coefficient (LPCC). The results of many studies have shown that the cepstral feature set is quite effective among these acoustic parameters. In deep learning applications, mostly MFCC features, waveform of the signal or spectrogram images are used. It has been observed that the feature set of MFCC and its derivatives is used extensively in both acoustic and deep learning studies (Ji et al., 2021).

Many classifiers based on machine learning (ML) and deep learning (DL) are used to classify infant cry sounds. The prominent ML algorithms are as follows: artificial neural network (Bashiri & Hosseinkhani, 2020), probabilistic neural network (Hariharan et al., 2011), multi-layer perceptron (AhmedAl-Azzawi, 2014), support vector machine (Ashwini et al., 2021), Gaussian mixture model (Sharma et al., 2019) and k-nearest neighbors (Dewi et al., 2019). In addition, convolutional neural

network (Xie et al., 2021) and deep neural network (Alishamol et al., 2020) algorithms are used. Infant cry classification includes pre-processing, feature extraction, and classifier stages. Therefore, an approach that is a combination of optimal methods should be adopted to achieve the best performance.

In this study, the adaptive cepstral parameters based on empirical mode decomposition (EMD) is proposed as an alternative to the traditional cepstral feature extraction approach. To obtain the classification model, the publicly available Donate a Cry Corpus dataset, which includes baby cry sound data in five different categories, was used. MFCC and its derivatives, LPC and LPCC acoustic features were extracted from the coefficients in the first five modes of the cry signals that were decomposed by EMD. The highest distinctive features are obtained by ReliefF method. The selected features are used with SVM classifier and models are evaluated with performance metrics.

The remainder of the study can be summarized as follows: In Section 2, the methodology of the proposed approach is introduced. In Section 3, the results are presented and discussed. In Section 4, the results are summarized and concluded.

2. Proposed Method

In this study, EMD-based cepstral features and machine learning are used to classify infant cry. Figure 1 illustrates the proposed approach for classification of cry sounds categorized into five types. The proposed approach consists of the following steps: (1) Acquire cry sound data from a publicly available dataset. (2) Synthetic data are obtained by applying the SMOTE technique to the class-imbalanced dataset. (3) The raw signals are decomposed by EMD and cepstral features are obtained using IMFs in each mode. (4) The most effective features are selected by using the ReliefF feature selection algorithm. (5) All features and selected feature sets are used with the support vector machine classifier. (6) Models are evaluated with metrics to determine the highest classification performance.

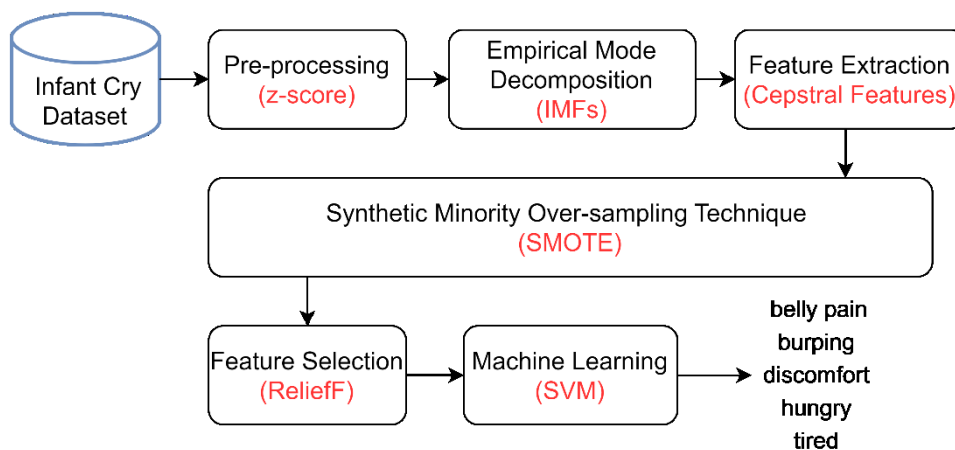


Figure 1. Block diagram of the proposed infant cry classification approach

2.1. Dataset

The baby cry acoustic sound data used in this study were obtained from the publicly available Donate-A-Cry corpus dataset (<https://github.com/gveres/donateacry-corpus>). The dataset includes sounds recorded for approximately 7 seconds with 8 kHz sampling rate and 128 kbps bit resolution. In the dataset, there are 457 signals from the categories of belly pain, burping, discomfort, hungry and tired. The dataset is class-imbalanced as the number of baby cry sound signals varies between classes. This class-imbalance creates a bias effect and severely affects classification performance. Therefore, the Synthetic Minority Oversampling Technique (SMOTE) algorithm is used to improve the imbalance rate

between classes (Wang et al., 2021). The SMOTE algorithm generates new samples by randomly interpolating between many samples and their nearby samples. The distribution of baby crying sounds in the original dataset and the artificial dataset created using the SMOTE algorithm according to classes is given in Table 1. In this study, the amplitudes of the signals were normalized with the z-score method.

Table 1. Distribution of original baby cry dataset and SMOTE applied dataset

	Original Data	Generated Data using SMOTE
Belly pain	16	256
Burping	8	64
Discomfort	27	351
Hungry	382	382
Tired	24	312

2.2. Empirical mode decomposition

The Hilbert-Huang transform (HHT) is a powerful method of signal analysis for decomposition non-stationary signals (Huang, 2005). HHT combines empirical mode decomposition (EMD) with Hilbert spectrum analysis techniques. EMD is widely used for decomposing time-varying signals such as voice/speech, and it decomposes signals from high frequency to low frequency modes (Arslan & Karhan, 2022). EMD employs the signal itself as the basis function in the decomposition process. The coefficients in each mode of the EMD decomposition are referred to as intrinsic mode functions (IMFs). An original $x(t)$ signal is mathematically decomposed by EMD as follows:

$$x(t) = \sum_{i=1}^k IMF_i(t) + r_k(t) \quad (1)$$

where k and $r(t)$ represent the total mode number and the residual signal, respectively.

The stopping criterion (SD) for the sifting process that yields all IMFs can be defined as:

$$SD_i = \sum_{t=0}^T \frac{|IMF_{i+1}(t) - IMF_i(t)|^2}{IMF_i(t)^2} \quad (2)$$

Figure 2 shows waveforms and Hilbert spectrum of infant cry signals. The first five level IMFs of the EMD are used to obtain the Hilbert spectrum of the signals. The Hilbert spectrums show that the spectral components of the cry sounds are different based on the modes and these differences can be used to distinguish the signals.

2.3. Feature extraction and selection

In voice processing applications, two techniques are usually prominent in acoustic information extraction. These parametric methods are based on the resonance structure of the vocal tract and include LPC, LPCC, and methods based on MFCC and their derivatives (Aggarwal et al., 2023; Zharif, 2015). In obtaining the cepstral domain-based parameters from the sound signal, pre-processing, framing and windowing operations are applied to the signals. Then, autocorrelation is carried out for LPC and LPCC features, while MFCC features are subjected to the FFT, mel-scale filter bank, logarithm, and DFT procedures, respectively (Chen et al., 2021; Fang et al., 2018; Ghoraani & Krishnan, 2011).

In this study, the EMD approach is used to decomposed infant cry signals into IMF modes. Considering the modes obtained with EMD, the first 5 IMF levels were used to extract features from cry

sound signals, since baby cry sounds contain different spectral components according to the modes. The raw signals and IMFs in each mode are used to extract features.

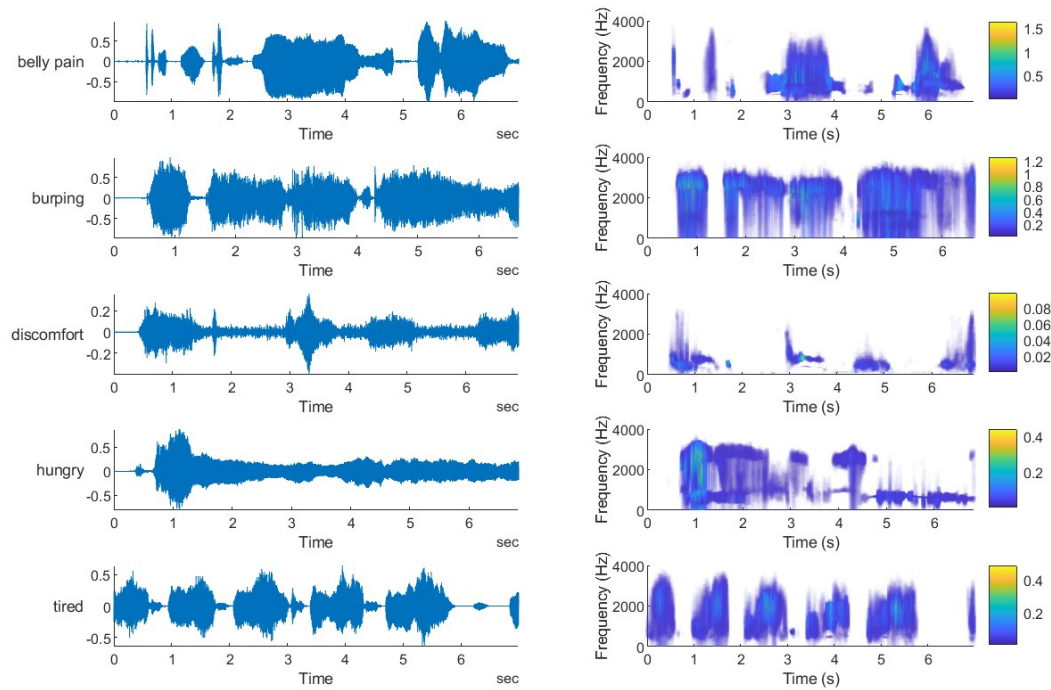


Figure 2. Infant cry signals and Hilbert spectrums

A feature vector of the MFCCs, its first and second derivatives (Δ -MFCC and $\Delta\Delta$ -MFCCs), LPC and LPCC is obtained from the signals and each IMFs. A total of 60 coefficients are obtained and the classification of cry signal is performed using both coefficients and derivatives, and mode frequency component differences. In obtaining the cepstral features within the scope of the study, 30 ms length with 20 ms overlap was used for framing and Hamming function was preferred for windowing. The features used in this study are given in Table 2.

Table 2. The features extracted from cry signals and their descriptions

Features	Description	Number of Features
MFCCs	Mel-frequency cepstral coeffs.	13
Δ -MFCCs	First derivative of MFCCs.	12
$\Delta\Delta$ -MFCCs	Second derivative of MFCCs.	11
LPCs	Linear prediction filter coeffs.	12
LPCCs	Linear prediction cepstral coeffs.	12
Total Features		60

Feature selection is an approach that improves classification performance in machine learning. Using the most distinctive features from the entire feature set will reduce both model complexity and execution

time. Therefore, the ReliefF feature selection algorithm (Kira & Rendell, 1992) was used in this study. The reduced selected feature set was obtained with the ReliefF algorithm using the 10 nearest neighbors.

2.4. Support vector machines

The SVM algorithm developed by (Vapnik, 1999) is based on a discriminative hyperplane principle. The optimum hyperplane is selected to optimize the separation between data points on each side of a plane in relation to the maximum margin. Support vectors are the data points that are most closely spaced from each separating hyperplane. SVM methods employ kernel functions, which take data as input and convert it into the necessary forms. The mathematical expressions of linear, quadratic and cubic kernel functions are given in Table 3. Also, the C parameter and kernel scale are set to 10 and 1, respectively.

Table 3. Linear and polynomial (quadratic and cubic) kernel functions for SVM

Kernel Function	Mathematical Expression
Linear Kernel	$k(x_1, x_2) = x_1 \cdot x_2$
Quadratic Poly. Kernel	$k(x_1, x_2) = (x_1 \cdot x_2 + 1)^2$
Cubic Poly. Kernel	$k(x_1, x_2) = (x_1 \cdot x_2 + 1)^3$

2.5. Performance metrics

In this study, accuracy (Acc), F1-score (F1), and Matthews correlation coefficient (MCC) metrics, which are heavily preferred especially in class-imbalanced conditions, were used in the performance evaluation of classification models. These metrics are calculated as follows (Chicco & Jurman, 2020):

$$Accuracy (Acc) = \frac{TP + TN}{TP + TN + FP + FN} \quad (3)$$

$$F1 - score (F1) = \frac{2TP}{2TP + FN + FP} \quad (4)$$

$$MCC = \frac{TP \times TN - FP \times FN}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}} \quad (5)$$

The four possible outcomes in the classification model evaluation are as follows: true positive (TP), false negative (FN), true negative (TN), and false positive (FP). K-fold cross-validation (CV) technique is employed in order to generate the best and most reliable classification model. In this study, 5-fold CV was preferred for classification models.

3. Results and Discussion

The adaptive cepstral features and the SVM algorithm were used to apply a classification method for infant cry sounds. In this approach, the first five-level IMFs were chosen from the modes that have been generated by the decomposition of baby cry signals into by EMD. Then, using these IMFs for each mode, the cepstral domain-based acoustic features such as MFCC, Δ -MFCC, $\Delta\Delta$ -MFCC, LPC and

LPCC were extracted. In proposed approach, a feature vector with a total number of 5x60 (1x60 per level) is obtained for each sound signal. ReliefF feature selection algorithm was applied to all features and a reduced selected feature set was obtained. All feature and selected feature sets were used with the SVM algorithm. The resulting model performances were evaluated with accuracy, F1 score, and MCC metrics, and Table 4 gives the results.

Table 4. Performance comparison of classification models for all features and selected features by ReliefF algorithm

Features/Algorithm	Perf. of all features			Perf. of selected features		
	Acc (%)	F1 (%)	MCC	Acc (%)	F1 (%)	MCC
Conventional Features/SVM Linear	67.5	63.7	0.557	72.7	72.3	0.653
Conventional Features/SVM Quadratic	91.0	90.8	0.887	94.1	93.7	0.922
Conventional Features/SVM Cubic	92.0	90.7	0.891	94.5	93.9	0.926
Mode #1 Features/SVM Linear	70.5	67.6	0.603	76.0	75.7	0.694
Mode #1 Features/SVM Quadratic	92.1	91.0	0.893	92.7	92.0	0.907
Mode #1 Features/SVM Cubic	92.2	91.4	0.898	93.0	92.3	0.910
Mode #2 Features/SVM Linear	70.4	64.9	0.584	74.1	73.3	0.666
Mode #2 Features/SVM Quadratic	91.3	90.3	0.885	93.3	92.7	0.914
Mode #2 Features/SVM Cubic	91.5	90.8	0.890	92.7	91.8	0.903
Mode #3 Features/SVM Linear	71.5	68.3	0.618	76.0	74.8	0.688
Mode #3 Features/SVM Quadratic	90.8	91.1	0.889	93.3	93.4	0.918
Mode #3 Features/SVM Cubic	91.0	90.5	0.886	93.0	92.8	0.913
Mode #4 Features/SVM Linear	70.1	67.9	0.605	73.9	72.8	0.661
Mode #4 Features/SVM Quadratic	88.9	89.8	0.870	90.2	90.4	0.879
Mode #4 Features/SVM Cubic	90.0	90.4	0.880	91.6	91.1	0.892
Mode #5 Features/SVM Linear	66.1	63.1	0.551	73.0	71.9	0.651
Mode #5 Features/SVM Quadratic	89.7	89.5	0.873	90.8	90.7	0.886
Mode #5 Features/SVM Cubic	90.5	89.9	0.879	91.4	90.3	0.885
All Modes Features/SVM Linear	88.9	88.4	0.856	91.4	91.8	0.901
All Modes Features/SVM Quadratic	94.7	94.8	0.937	96.4	96.6	0.957
All Modes Features/SVM Cubic	96.1	95.8	0.949	99.0	98.8	0.985

The results show that the selected feature set obtained by the ReliefF algorithm provides a higher performance than the conditions in which all features are used. It is concluded that the cepstral features extracted from all modes of EMD provide high classification performance. In addition, the cubic kernel function in the SVM algorithm provided more effective results than linear and quadratic functions. In the condition of using the SVM-cubic algorithm, 94.4% accuracy, 93.9% F1-score and 0.926 MCC values were obtained for the cepstral features obtained from the raw signal, while 99% accuracy, 98.8% F1-score and 0.985 MCC values were obtained for the cepstral features extracted from all modes.

The confusion matrix of classification models obtained by cepstral features extracted from the raw signals and IMFs provided by EMD is shown in Figure 3. In the confusion matrix for the raw signal cepstral features shown in Figure 3a, accuracy rates of 100%, 92.2%, 100%, 82.7%, and 98.7% were obtained for the classes of belly pain, burping, discomfort, hungry, and tired, respectively. Similarly, accuracy rates of 100% for belly pain, 95.3% for burping, 100% for discomfort, 97.1% for hungry, and 100% for tired were observed for the cepstral features obtained from EMD modes, as shown in Figure 3b. Extensive results showed that the EMD approach improved the classification of burping, hungry, and tired categories compared to traditional cepstral-domain approaches.

This study was compared with approaches in the literature using the same dataset, and the performances are summarized in Table 5. The results showed that increasing the numbers of data with

the SMOTE technique and using it in the training of the models contributed to the performance improvements and contributions. In addition, acoustic parameters such as formants, spectral features, MFCC, LPC, and LPCC directly from raw signals provide limited performance. In another study, the highest accuracy rate of 95.2% was obtained by scalogram and Resnet-18 CNN.

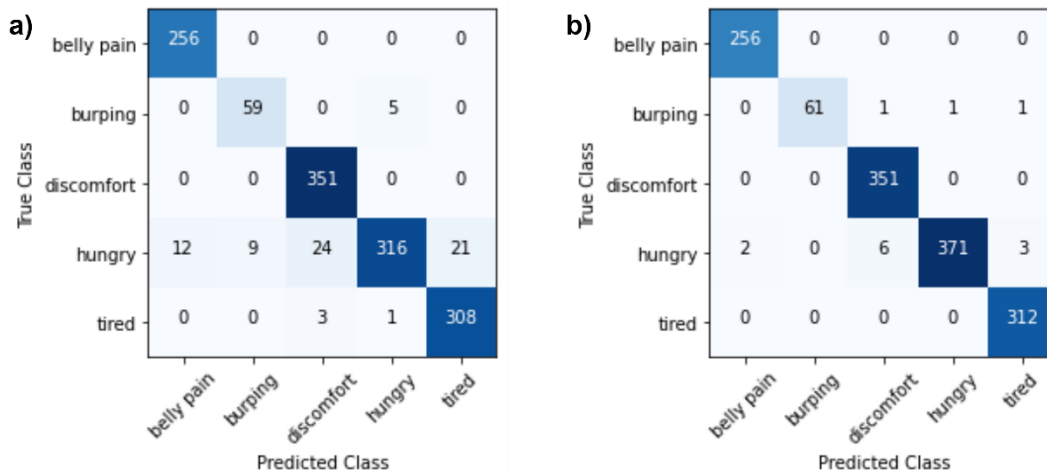


Figure 3. The confusion matrices of infant cry classification a) for conventional cepstral features b) for cepstral features based on EMD

Table 5. Performance comparison of the proposed approach with other studies

Author/Year	Features/Input	Algorithm	Accuracy (%)
(Sharma et al., 2019)	Frequency, Entropy, and Spectral	GMM	81.3
(Kulkarni et al., 2021)	MFCC, GFCC, and ZCR	RF	84
(Ozseven, 2023)	F1, F2, F3, MFCC, and LPCC	SVM	87.9
(Ozseven, 2023)	Scalogram	ResNet-18	95.2
This study	MFCC, LPC and LPCC	SVM	94.5
This study	MFCC, LPC and LPCC based on EMD	SVM	99.0

4. Conclusion

In this study, a cepstral feature vector obtained by EMD-based decomposition is proposed as an alternative to traditional cepstral-domain features for the classification of infant cry sounds. The SMOTE technique was applied to the class-imbalanced original dataset with a small number of data and a synthetic dataset with a high number of data was obtained. MFCC and its derivatives, LPC and LPCC features extracted using the IMFs in the first five modes obtained by EMD were used with the SVM algorithm. The results showed that the cepstral features obtained from all modes provide higher performance than the traditional cepstral features extracted from the raw signal. It has been observed that the proposed approach classifies burping, hungry and tired classes with higher accuracy than the traditional approach. Thus, a computer-assisted infant cry classification system was designed that could help inexperienced parents.

References

- Aggarwal, G., Jhajharia, K., Izhar, J., Kumar, M., & Abualigah, L. (2023). A Machine Learning Approach to Classify Biomedical Acoustic Features for Baby Cries. *Journal of Voice*, 1–10.
- AhmedAl-Azzawi, N. (2014). Automatic Recognition System of Infant Cry based on F-Transform. *International Journal of Computer*, 102(12), 28–32.

- Alishamol, K. S., Fousiya, T. T., Babu, K. J., Sooryadas, M., & Mary, L. (2020). System for infant cry emotion recognition using DNN. *Proceedings of the 3rd International Conference on Smart Systems and Inventive Technology, ICSSIT 2020*, 867–872.
- Arslan, Ö., & Karhan, M. (2022). Effect of Hilbert-Huang transform on classification of PCG signals using machine learning. *Journal of King Saud University - Computer and Information Sciences*, 34(10), 9915–9925.
- Ashwini, K., Vincent, P. M. D. R., Srinivasan, K., & Chang, C. Y. (2021). Deep Learning Assisted Neonatal Cry Classification via Support Vector Machine Models. *Frontiers in Public Health*, 9(June), 1–10.
- Bashiri, A., & Hosseinkhani, R. (2020). Infant crying classification by using genetic algorithm and artificial neural network. *Acta Medica Iranica*, 58(10), 531–539.
- Chang, C. Y., Bhattacharya, S., Raj Vincent, P. M. D., Lakshmana, K., & Srinivasan, K. (2021). An Efficient Classification of Neonates Cry Using Extreme Gradient Boosting-Assisted Grouped-Support-Vector Network. *Journal of Healthcare Engineering*, 2021.
- Chen, X., Hu, M., & Zhai, G. (2021). Cough Detection Using Selected Informative Features from Audio Signals. *Proceedings - 2021 14th International Congress on Image and Signal Processing, BioMedical Engineering and Informatics, CISP-BMEI 2021*.
- Chicco, D., & Jurman, G. (2020). *The advantages of the Matthews correlation coefficient (MCC) over F1 score and accuracy in binary classification evaluation*. 1–13.
- Dewi, S. P., Prasasti, A. L., & Irawan, B. (2019). Analysis of LFCC feature extraction in baby crying classification using KNN. *Proceedings - 2019 IEEE International Conference on Internet of Things and Intelligence System, IoTaIS 2019*, 86–91.
- Fang, S. H., Tsao, Y., Hsiao, M. J., Chen, J. Y., Lai, Y. H., Lin, F. C., & Wang, C. Te. (2018). Detection of Pathological Voice Using Cepstrum Vectors: A Deep Learning Approach. *Journal of Voice*, 1, 20–22.
- Ghoraani, B., & Krishnan, S. (2011). Time–Frequency Matrix Feature Extraction and Classification of Environmental Audio Signals. *IEEE Transactions on Audio, Speech, and Language Processing*, 19(7), 2197–2209.
- Hariharan, M., Yaacob, S., & Awang, S. A. (2011). Pathological infant cry analysis using wavelet packet transform and probabilistic neural network. *Expert Systems with Applications*, 38(12), 15377–15382.
- Huang, N. E. (2005). Introduction to the Hilbert Huang Transform. *Transform*, 5, 1–26.
- Ji, C., Mudiyanse, T. B., Gao, Y., & Pan, Y. (2021). A review of infant cry analysis and classification. *Eurasip Journal on Audio, Speech, and Music Processing*, 2021(1).
- Kira, K., & Rendell, L. A. (1992). A Practical Approach to Feature Selection. In *Machine Learning Proceedings 1992*. Morgan Kaufmann Publishers, Inc.
- Kulkarni, P., Umarani, S., Diwan, V., Korde, V., & Rege, P. P. (2021). Child Cry Classification - An Analysis of Features and Models. *2021 6th International Conference for Convergence in Technology, I2CT 2021*, 1–7.
- Ozseven, T. (2023). Infant cry classification by using different deep neural network models and hand-crafted features. *Biomedical Signal Processing and Control*, 83, 104648.
- Sharma, K., Gupta, C., & Gupta, S. (2019). Infant Weeping Calls Decoder using Statistical Feature Extraction and Gaussian Mixture Models. *2019 10th International Conference on Computing, Communication and Networking Technologies, ICCCNT 2019*, 1–6.
- Vapnik V. (1999). The nature of statistical learning theory. Springer science & business media.

- Wang, S., Dai, Y., Shen, J., & Xuan, J. (2021). Research on expansion and classification of imbalanced data based on SMOTE algorithm. *Scientific Reports*, 11(1), 1–11.
- Xie, J., Long, X., Otte, R. A., & Shan, C. (2021). Convolutional Neural Networks for Audio-Based Continuous Infant Cry Monitoring at Home. *IEEE Sensors Journal*, 21(24), 27710–27717.
- Zharif, A. (2015). LPCC Features. *Artificial Intelligence and Speech Technology*

Kablosuz Sensör Ağlarının Malatya Minimum Vertex Cover Yöntemiyle Konumlandırılması

Positioning of Wireless Sensor Networks with Malatya Minimum Vertex Cover Method

Cemalettin Sonakalan^{*1} , Furkan Öztemiz¹ 

¹Yazılım Mühendisliği Bölümü, İnönü Üniversitesi, Malatya, Türkiye

(cemalettin.sonakalan@inonu.edu.tr, furkan.oztemiz@inonu.edu.tr)

Özetçe— Minimum Vertex Cover (MVC) problemi, çizge teorisi alanında önemli bir optimizasyon problemidir. Verilen bir grafin en az sayıda düğüm içeren bir düğüm kümesini bulmayı amaçlar. Bu düğüm kümesi, grafin tüm kenarlarını kapsayan düğümleri içermektedir. MVC problemi birçok uygulama alanında önemli bir rol oynar. Örneğin, ağ tasarımı, iletişim ağları, bilgisayar ağları ve sosyal ağlar gibi alanlarda kullanılır. Bir ağdaki düğümlerin minimum sayıda kapsama alanına sahip olması, ağın verimliliğini artırırken, enerji tüketimini azaltır ve kaynakların daha etkin bir şekilde kullanılmasına olanak sağlar. Bu çalışmada, İnönü Üniversitesi yerleşkesindeki bütün bina ve tesislerin birbirleri ile haberleşmesi için konumlandırılması gereken kablosuz erişim noktalarının en az sayıda kullanılması hedeflenmektedir. Kablosuz erişim noktalarının kampüs içerisinde konumlandırılmaları için daha önce bu alanda ve bu özgün kampüs ağında kullanılmamış olan Malatya Vertex Cover yöntemi kullanılacaktır. Yöntemin birçok graf türünde başarılı sonuçlar verdiği literatürdeki araştırmalar sonucunda tespit edilmiştir. Çalışmada üniversite yerleşkesindeki bütün binalar Google Earth aracılığı ile işaretlenerek bir çizge örüntüsü oluşturulması sağlanmıştır. İşaretlenen her bir bina çizgenin bir düğümünü ifade etmektedir. R programlama dili ile çizge modelleneyecektir. Bir sonraki aşamada ise hangi düğümlerin seçileceğine Malatya Vertex Cover algoritması karar verecektir. Bu analize göre yerleşkedeki tüm tesislere optimal sayıda kablosuz erişim noktası ile konumlandırılacaktır. Bu sayede en az sayıda kablosuz erişim noktası kullanılacak olup maliyet konusunda daha avantajlı bir senaryo ortaya çıkacaktır.

Anahtar Kelimeler: *Minimum Vertex Cover, Kablosuz Sensör Ağları, Çizge Teorisi, Malatya Vertex Cover*

Abstract— The Minimum Vertex Cover (MVC) problem is an important optimization problem in graph theory. It aims to find a node set that contains the least number of nodes in a given graph. This node set contains nodes covering all the edges of the graph. The MVC problem plays an important role in many application areas. For example, it is used in areas such as network design, communication networks, computer networks and social networks. Having a minimum number of nodes in a network increases the efficiency of the network, reduces energy consumption and allows more efficient use of resources. In this study, it is aimed to use the minimum number of wireless access points that need to be located for all buildings and facilities in the İnönü University campus to communicate with each other. Malatya Vertex Cover method, which has not been used before in this area and in this unique campus network, will be used for positioning wireless access points within the campus. It has been determined as a result of the studies in the literature that the method gives successful results in many graph types. In the study, all the buildings in the university campus were marked via Google Earth and a graph pattern was created. Each marked building represents a node of the graph. The graph will be modeled with the R programming language. In the next stage, Malatya Vertex Cover algorithm will decide which nodes will be selected. According to this analysis, all facilities on the campus will be located with an optimal number of wireless access points. In this way, the least number of wireless access points will be used and a more cost-effective scenario will emerge.

Keywords : *Minimum Vertex Cover, Wireless Sensor Networks, Graph Theory, Malatya Vertex Cover*

1.Giriş

Kablosuz sensör ağları (WSN), belirli bir coğrafi bölgede bir ağ içinde gruplanarak oluşan küçük, kendi kendini organize eden, otonom olarak çalışabilen ve genellikle telsiz ile haberleşen akıllı sensör cihazlarıdır. Kullanım şekilleri ve amaçlarından kaynaklanan farklılıklara rağmen, bu cihazların ortak özelliği kaynaklarının sınırlı olmasıdır. Temel olarak, bu sınırlı özellikler küçük fiziksel boyutlar, küçük güç kaynakları, kısa radyo menzili, küçük bellek kapasitesi, ağı diğer parçaları hakkında bilgi eksikliği ve iletişim becerilerinin basitliği şeklinde özetlenebilir [1]. Çizgeler ise matematik ve bilgi sistemlerinde yaygın olarak kullanılan bir veri yapısı modelidir.[2] Optimizasyon problemlerinden birçoğu çizgeler ile modellenilebilir. Minimum Tepe Örtüsü Problemi (MVC), çizge teorisinin kilit sorunlarından biri olup, bir çizgenin tüm kenarlarının en az düğüm sayısı ile ne kadar iyi örtülebileceğinin belirlenmesi problemidir. Minimum Tepe Örtüsü Problemi bir optimizasyon problemi olduğundan, bu problemi çözmek için birçok algoritma ve yaklaşım önerilmiştir. Bu çalışmalardan bazıları;

Kablosuz iletişim, veri iletişimi ve radyo paketlerinin aktarımı yoluyla sağlandığı için veri manipülasyonu gibi çeşitli saldırılara açıktır. Bu soruna karşı bir önlem radyo paketlerini fiziksel olarak yakalayabilen ve denetleyebilen güvenli noktalar konuşlandırarak bağlantıları izlemektir. Çizge teorisi Kablosuz Sensör Ağlarındaki çeşitli problemleri çözmek için kritik bir rol oynamaktadır Minimum Tepe Örtüsü Problemi Kablosuz Sensör Ağları için önemli bir yapıdır ve Minimum Tepe Örtüsü sayesinde seçilen düğümler güvenli noktalar (monitörler) olarak ayarlandığında bağlantıları izlemek için mükemmel uyum sağlar [3].

Minimum Tepe Örtüsü Problemi için geliştirilmiş Feromon Düzeltme Stratejisine Sahip Bir Karınca Kolonisi Optimizasyon Algoritması da bir başka çalışma olarak karşımıza çıkmaktadır. Bu çalışmada, şüpheli unsurları hariç tutarak optimize çözüm ile bir feromon düzeltme sezgisel stratejisi önerilmiştir. Elemanlar, onları en iyi çözümün olası üyeleri yapmayan bazı istenmeyen özelliklere sahiplerse şüphelidirler. Bu ayrıştırma ile yerel yakınsamada karıncaların erken tuzağa düşmesini önleyerek saf karınca kolonisi optimizasyon algoritmasını geliştirmek amaçlanmıştır [4].

Kablosuz Sensör Ağlarında İletişim Probleminin Tepe Örtüsü Olarak Modellenmesi: İletişim problemi, bir hizmet alanına yerleştirilen sensör cihazlarının minimum kümesini seçmektir, böylece tüm hizmet alanına minimum sensör kümesi tarafından erişilebilir. Minimum sensör kümesini bulmak bir tepe noktası örtüsü problemi olarak modellenir; burada tepe noktası örtüsü kümesi, tüm sensörlerin sınırlı iletişim menzilini ve pil ömrünü göz önünde bulundurarak sensörler arasındaki iletişimi çok atlamalı bir şekilde kolaylaştırır [5].

Her bireyi bir nokta olarak kabul edelim. Eğer iki birey birbirini tanıyor ise, bu iki noktayı bir kırmızı hat ile birleştirelim. Ancak eğer iki birey birbirini tanımıyorsa, bu iki noktayı bir mavi hat ile birleştirelim. Sonuç olarak, birbirleriyle ya kırmızı ya da mavi hat ile bağlanmış olan sınırsız sayıda nokta elde edilmiştir. Bu noktalar içerisinde, yalnızca aynı renkte hatlarla (tamamı kırmızı veya tamamı mavi) bağlanmış olan sınırsız sayıda nokta bulabiliriz. Bu fikir, bu çalışmanın temelini oluşturmuştur [6].

Minimum Ağırlık Tepe Örtüsü Problemi İçin Geliştirilmiş Feromon Düzeltme Stratejisine Sahip Bir Karınca Kolonisi Optimizasyon Algoritması. Bu çalışmada, şüpheli unsurları hariç tutmak için en iyi bulunan çözüm hakkındaki bilgileri kullanan bir feromon düzeltme sezgisel stratejisi önerilmiştir. Elemanlar, onları en iyi çözümün olası üyeleri yapmayan bazı istenmeyen özelliklere sahiplerse şüphelidirler. Bu ayrıştırma ile yerel yakınsamada karıncaların erken tuzağa düşmesini önleyerek saf karınca kolonisi optimizasyon algoritmasını geliştirmek amaçlanmıştır [7].

Minimum Vertex Cover için Yeni Bir Yerel Arama: Bu çalışmada, Minimum Vertex Cover problemi için yeni bir yerel arama algoritması olan Edge Weighting Local Search (EWLS) tanıtılmaktadır. EWLS' nin kilit noktası, minimum köşe örtüsünün boyutu üzerinde sıkı bir üst sınır sağlayan bir köşe kümesi

bulmaktır. Bu amaçla EWLS, yerel optimumlarda takılıp kaldığında kenar ağırlıklarını güncelleyen bir kenar ağırlıklandırma şeması kullanan yinelemeli bir yerel arama prosedürü kullanır [8].

Kuantum Tavlayıcı Üzerinde Büyük Minimum Köşe Örtüsü Problemlerinin Çözümü: Kuantum tavlayıcılar, donanım gömülebildikleri takdirde bu tür NP-zor problemlerin optimum çözümünü bulabilirler. Ancak bu donanım kısıtlarından dolayı pek mümkün değildir. Bu yöntemde, minimum köşe örtüsü problemi için bir ayrıştırma algoritması sunulmaktadır: Algoritma, oluşturulan alt problemler tavlayıcıya gömülüp çözülebilece kadar rastgele bir problemi özyinelemeli olarak böler [9].

Bu çalışmada, Kablosuz Sensör Ağlarından minimum sayıda kullanılarak İnönü Üniversitesi yerleşkesindeki tüm binalar arasında bir network bağlantısı oluşturulması için cihaz kurulumlarının hangi binalarda olması gerektiğinin tespiti hedeflenmiştir. Tasarlanmış kampüs ağı çizge olarak modellenmiş ve özgün bir veri seti oluşturulmuştur. Problemin çözümü için literatüre yeni eklenmiş olan Malatya Vertex Cover algoritması tercih edilmiştir. Malatya Vertex Cover algoritmasının iki önemli aşaması bulunmaktadır. Düğüm merkezlik değeri hesaplanması ve örtü düğümleri seçimidir. Merkezlik değerinin hesaplanması Malatya Centrality algoritması ile yapılmaktadır. Algoritma çizgedeki her bir düğümün komşu düğümlerin derecelerine oranlarının toplamından oluşmaktadır. İkinci adım, tepe noktası örtüsü için bir düğüm seçim problemidir. Çizgedeki düğümlerden Malatya Merkezlik Değeri en büyük olan düğüm seçilir ve çözüm kümesine eklenir. İlgili tepe noktası ve kenarlar çizge üzerinden kaldırılır. Bu aşamalar iteratif olarak devam eder. İşlem tamamlandığında Minimum Tepe Örtüsü Problemi için çözümü oluşturan gerekli düğüm kümesi belirlenmiş olur [10].

Yöntem daha önce özel ve küçük yapay çizgeler üzerinde test edilmiştir. Gerçekleştirilen analizler neticesinde, Kablosuz Sensör Ağlarının hangi konumlara kurulacağı bilgisi tespit edilmiştir. Bu sayede her binada Kablosuz Sensör Ağı kurmak yerine minimum sayıda bina için kurulum yapılarak kampüs içerisindeki tüm binalar arasında network oluşturulmuştur.

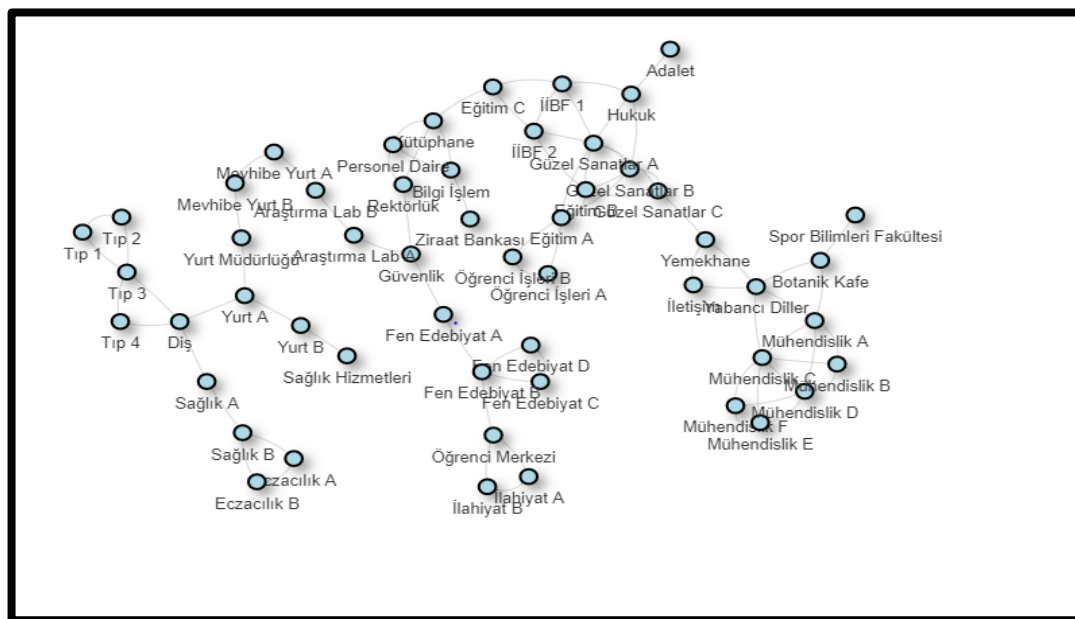
2. Materyal ve Metot

Bu çalışmada İnönü Üniversitesi yerleşkesindeki bütün binaların bir çizge modeli oluşturulmuştur. Bunun için görsel tarafta Google Earth programı kullanılırken arka planda ıgraph kütüphanesi ile sayısal olarak network tasarlanmıştır. Modellenen çizge üzerinde Malatya Vertex Cover algoritması uygulanmaktadır. Algoritma R programla dili ile kodlanmıştır. Yöntemlerin uygulanması sonrasında elde edilen vertex cover üyeleri binalarının sensor networklerin kurulumları için uygun lokasyonları ifade etmektedir. Sonuçlar kapsamlı şekilde analiz edilecek ihtiyaç halinde yöntemlere ek geliştirmeler yapılacak veya çizge ağının modellenmesi yenilenecektir. Tasarlanan yerleşke çizgesi ve kullanılan algoritmaların gerçek saha verilerine ilk defa uygulanması çalışmayı özgün kılan iki önemli etmendir. Şekil 1 'de verilerimiz Google Earth ile elde edilen model görülmektedir. Bu modelde her bina için 75m yarı çapında daireler ile işaretlenmiştir. 75m lik mesafe binalara konumlandırılacak olan kablosuz cihazların kapsama alanını ifade etmektedir. Bu model oluşturulduktan sonra R programlama dili ile bu model bir veri yapısı olan çizgeye aktarılmıştır.



Şekil 1. Google Earth üzerinde oluşturulan model

Şekil 2’de ise kapsama alanı kesişen düğümlerin ve bağıntılarının oluşturulduğu çizge gösterilmiştir.



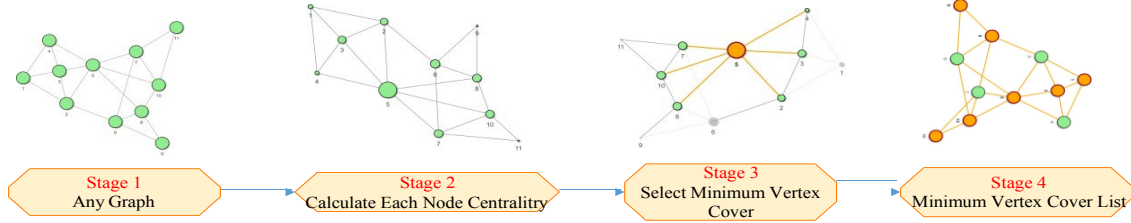
Şekil 2. Kampüsdeki binaların çizgesel hali

2.1. Malatya Vertex Cover Algoritması

Vertex-cover problemi, herhangi bir grafikteki düğümleri kullanarak grafikteki tüm kenarların kapsanabilir olup olmadığını göstermektir. Birçok gerçek dünya problemi VC problemleri olarak ifade edilebilir. Koridorlara aynı kattaki tüm odaları gören kameralar yerleştirmek veya tekrarlayan DNA dizilerini tespit etmek bu problem için örnek uygulamalardır. VC problemi için olası çözümler, gerçek dünya problemleri için de çözümler sunmaktadır [11].

Minimum Vertex Cover (MVC) problemi, graf teorisi alanında önemli bir optimizasyon problemidir. Verilen bir grafın en az sayıda düğüm içeren bir düğüm kümesini bulmayı amaçlar. Bu düğüm kümesi,

grafın tüm kenarlarını kapsayan düğümleri içermektedir. MVC problemi birçok uygulama alanında önemli bir rol oynar. Örneğin, ağ tasarımı, iletişim ağları, bilgisayar ağları ve sosyal ağlar gibi alanlarda kullanılır. Bir ağdaki düğümlerin minimum sayıda kapsama alanına sahip olması, ağın verimliliğini artırırken, enerji tüketimini azaltır ve kaynakların daha etkin bir şekilde kullanılmasına olanak sağlar.



Şekil 3. Malatya Vertex Cover Algoritmasının Aşamaları [10]

Çizge teorisi ve ağ analizi gibi bir dizi disiplinde merkezilik önemli bir konsepttir. Merkezilik, düğümlerin çizgedeki konumlarına dayalı olarak numaralandırılmasını ve sıralanmasını içerir [12]. Çoğu uygulamada hedef, çizge veya ağdaki en etkin düğümün tespitidir. Merkez düğüm veya düğümlerin belirlenmesi için bir dizi algoritma sunulmuştur. Ancak düğümün bağlantıları, çizgedeki belirlenmiş düğüm ile aynıdır. Bunlar, merkeziyetini ölçümünde önemli parametrelerdir. Bu yaklaşım, genellikle Derece Merkeziliği olarak bilinir ve PageRank gibi birçok yaygın algoritmanın temelini oluşturur. [13]

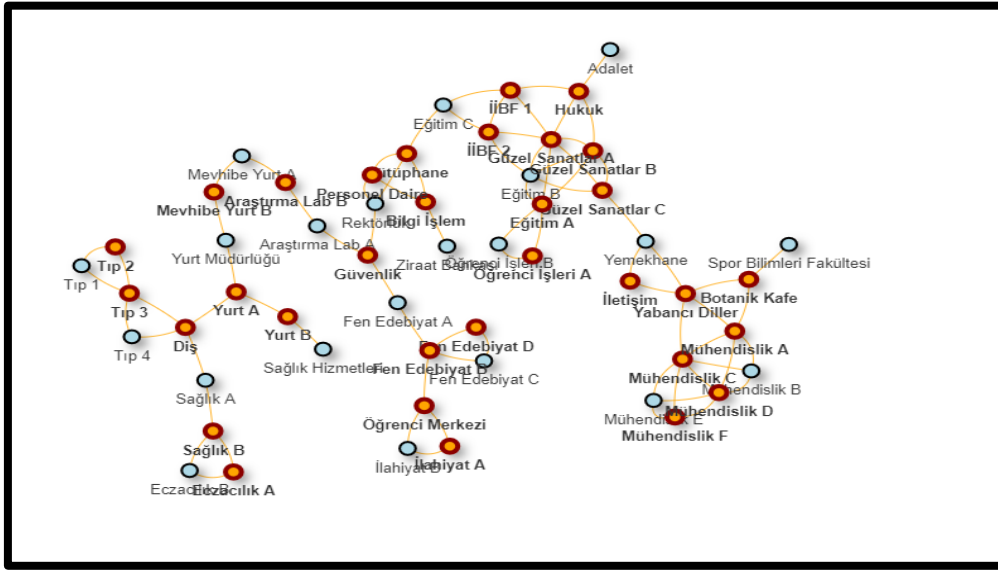
Malatya algoritması, merkezilik değerini Malatya merkezilik değeri olarak tanımlar ve bu değer hesaplanmasında düğüm bağlantıları kullanılır. Bir düğümün Malatya merkezilik değeri, düğümün derecesi ve komşu düğümlerin dereceleri ile hesaplanır. Malatya merkezilik değerinin belirlenmesinde, düğümün değeri yanı sıra komşu düğümlerin dereceleri de belirleyici faktördür. Malatya Vertex Cover algoritmasının sözde kodu Algoritma 1'de verilmiştir.

Algoritma 1. Malatya Vertex Cover algoritmasının matematiksel gösterimi[10]

Minimum Vertex-Cover Algorithm	
Input: Adjacency matrix of G is A and $G = (V, E)$	// G graph
Output: $V_c \subseteq V$, V_c is a set of nodes and it is a solution for vertex-cover problem	
1. $V_c \leftarrow \emptyset$	
2. While $E \neq \emptyset$ do	
3. $i \leftarrow 1, \dots, V $	
4. $\psi(v_i) = \sum_{v_j \in N(v_i)} d(v_j)$	
5. $V_c = V_c \cup \{\max(\psi(v_i))\}$	
6. $V = V - \{v_i\}$, and $E = E - \{(v_i, v_j) \in E\}$	
7. Output= V_c	

3. Uygulama

Şekil 4'te ise Malatya Vertex Cover algoritması uygulandıktan sonra çizgeden elde ettiğimiz düğümler gösterilmiştir. Çalışmanın başında 53 bina işaretlenmiş ve nodelar oluşturulmuştur. Bu nodelar arasında toplam 77 adet kenar belirlenmiştir. Algoritmamızı çizgeye uyguladıktan sonra sadece 32 adet düğümler bu özgün ağı tamamını kapsamış olduk. Bu 32 düğüm Şekil 4 'de turuncu renk ile belirlenmiştir.



Şekil 4. Algoritma sonrası filtrelenmiş çizge

Tablo 1.'de tüm ağı kapsayabilen düğümler belirtilmiştir. Bu seçilen düğümlerde routerlarımızı konumlandırarak üniversite yerleşkesinde kesintisiz bir hat sağlanmaktadır.

Tablo 1. Kampüs network'de seçilen lokasyonlar

Güzel Sanatlar A	Mühendislik C	Fen Edebiyat B	Mühendislik D	Tıp 3
Yabancı Diller	Eğitim A	Güvenlik	Kütüphane	Diş
Sağlık B	Hukuk	Güzel Sanatlar C	İİBF 2	Mühendislik A
Bilgi İşlem	Araştırma Lab B	Mevhibe Yurt B	Yurt A	Öğrenci Merkezi
Mühendislik F	Botanik Kafe	İletişim	Güzel Sanatlar B	İİBF 1
Öğrenci İşleri A	Personel Daire	Fen Edebiyat D	Yurt B	Eczacılık A
Tıp 2	İlahiyat A			

4. Sonuç ve Tartışma

Bu çalışmada İnönü Üniversitesi kampüsündeki binalar arasında kablosuz sensör ağları ile bir bağlantı oluşturulması ve minimum sayıda cihaz kullanarak bunun gerçekleştirilmesi hedeflenmiştir. Problemin çözümü için literatüre yeni kazandırılmış Malatya Vertex Cover algoritması kullanılmıştır. Tasarlanan kampüs çizgesi çalışmayı özgün kılan önemli bir veri setidir. Ayrıca literatüre yeni kazandırılan Malatya Vertex Cover (MVC) algoritması ilk defa bir gerçek dünya probleminin çözümü için kullanılmıştır. MVC 'den elde edilen sonuçlara dayanarak, algoritmanın başka grafik problemlerine uygulanabilirliği de değerlendirilmiştir. Son analizler sonucunda, yerleşkedeki tüm tesislere optimum sayıda kablosuz erişim noktası konumlandırılmıştır. Bu sayede en az sayıda kablosuz erişim noktası kullanılmış, maliyet ve enerji tüketimi açısından daha avantajlı bir senaryo oluşturulmuştur.

İlerleyen süreçlerde çalışmanın genişletilerek algoritmanın ulaşım ağı gibi başka çizge problemlerine de uygulanabilirliği araştırılacaktır.




Teşekkür

Bu çalışmaya desteklerinden dolayı İnönü Üniversitesi Bilimsel Araştırmalar Koordinasyon Birimine teşekkür ederiz. Proje No: 3136 ve kodu: FBG-2023-3136.

Kaynaklar

- [1] S. Gowrishankar, T. G. Basavaraju, D. H. Manjaiah & S. K. Sarkar, (2008) "Issues in wireless sensor networks", In Proceedings of the World Congress on Engineering (WCE '08), Vol. 1.
- [2] Cormen, TH, Leiserson, CE, Rivest, R., & Clifford, S. (2001). Introduction to algorithms (Introducti). London
- [3] Z. A. Dagdeviren, "Weighted Connected Vertex Cover Based Energy-Efficient Link Monitoring for Wireless Sensor Networks Towards Secure Internet of Things," in IEEE Access, vol. 9, pp. 10107-10119, 2021, doi: 10.1109/ACCESS.2021.3050930.
- [4] Jovanovic, Raka, and Milan Tuba. "An ant colony optimization algorithm with improved pheromone correction strategy for the minimum weight vertex cover problem." Applied Soft Computing 11.8 (2011): 5360-5366.
- [5] M. Safar, M. Taha and S. Habib, "Modeling the Communication Problem in Wireless Sensor Networks as a Vertex Cover," 2007 IEEE/ACS International Conference on Computer Systems and Applications, Amman, Jordan, 2007, pp. 592-598, doi: 10.1109/AICCSA.2007.370690.
- [6] Monien, Burkhard & Speckenmeyer, Ewald. (1985). Ramsey Numbers and an Approximation Algorithm for the Vertex Cover Problem.. Acta Inf.. 22. 115-123. 10.1007/BF00290149.
- [7] Jovanovic, Raka, and Milan Tuba. "An ant colony optimization algorithm with improved pheromone correction strategy for the minimum weight vertex cover problem." Applied Soft Computing 11.8 (2011): 5360-5366.
- [8] Cai, Shaowei, Kaile Su, and Qingliang Chen. "EWLS: A new local search for minimum vertex cover." Proceedings of the AAAI Conference on Artificial Intelligence. Vol. 24. No. 1. 2010.
- [9] Pelofske, Elijah, Georg Hahn, and Hristo Djidjev. "Solving large minimum vertex cover problems on a quantum annealer." Proceedings of the 16th ACM International Conference on Computing Frontiers. 2019.
- [10] Karci, A. , Yakut, S. & Öztemiz, F. (2022). A New Approach Based on Centrality Value in Solving the Minimum Vertex Cover Problem: Malatya Centrality Algorithm . Computer Science , Vol:7 (Issue:2) , 81-88 . DOI: 10.53070/bbd.1195501
- [11] Hossain, A., Lopez, E., Halper, SM, Cetnar, DP, Chief, AC, Strickland, D., ... Salis, HM (2020). Automated design of thousands of nonrepetitive parts for engineering stable genetic systems. Nature Biotechnology, 38(12), 1466–1475. Retrieved from <https://doi.org/10.1038/s41587-020-0584-2>
- [12] Borgatti, SP (2005). Centrality and network flow. Social Networks, 27(1), 55–71. Retrieved from <https://doi.org/10.1016/j.socnet.2004.11.008>
- [13] Kumar, G., Duhan, N., & Sharma, AK (2011). Page ranking based on number of visits of links of Web page. In 2011 2nd International Conference on Computer and Communication Technology (ICCCT-2011) (pp. 1114). IEEE. Retrieved from <https://doi.org/10.1109/ICCCT.2011.6075206>

Optimizing Shift Scheduling with Constraint Programming: A Practical Approach Using Google OR-Tools

Ali Nihat Uzunalioglu¹ , Behçet Mutlu¹ , Deniz Kılınç² 

¹Wingie Enuygun Group, Istanbul, Turkey

²Mühendislik ve Mimarlık Fakültesi, Bakırçay Üniversitesi, İzmir, Türkiye

(nihat.uzunalioglu@enuygun.com, behcet.mutlu@enuygun.com, deniz.kilinc@bakircay.edu.tr)

Özetçe— Çalışan memnuniyeti ve operasyonel verimlilik, bu sistemin farklı alanlarda ele aldığı iki önemli husustur. Vardiya çizelgeleme, birden fazla kısıtlama ve tercih içeren karmaşık bir görevdir ve bu da akıllı çözümler bulmayı gerekli kılar. Önerilen sistemimiz kısıt programlamanın teorik kavramlarına dayanmaktadır. Karmaşık çizelgeleme kısıtlarını etkili bir şekilde ele almasıyla bilinen çok yönlü bir açık kaynak paketi olan Google'ın OR-Tools kütüphanesini uyguluyoruz. Sistem, araçlar arasında vardiya dağılımını optimize ederek personel çizelgeleme problemine odaklanmaktadır. Bu sistem, her bir vardiyanın uygun bir şekilde doldurulmasını sağlarken, temsilci mevcudiyeti ve dinlenme süreleri gibi zor kısıtlamaları da yönetmektedir. Sistem, girdi olarak temsilciler hakkında bilgi, vardiya dağılımı meta verileri ve geçmiş atama verilerini içeren bir Excel dosyası alır. Çıktı, gelecekte kullanılmak üzere bir Excel dosyasına geri yazılan ve böylece hata ve sorun olasılığını azaltan optimize edilmiş bir programdır. Ayrıca sistem, gelecekteki atamaları etkileyen ve operasyonel verimliliği artıran temsilci atamalarının tarihsel bir kaydını tutar. Bu makale, Google OR-Tools kullanarak pratik uygulamaya odaklanarak, kısıt programlama kullanarak vardiya yönetiminin optimizasyonunu kapsamlı bir şekilde açıklamayı amaçlamaktadır. Bu araştırma, kısıt programlama çözümlerini pratik bir gerçek dünya sorununa uyguladığı ve bu araçların kurumsal bir ortamdaki potansiyelini sergilediği için benzersizdir.

Anahtar Kelimeler : Çalışan planlaması, Program optimizasyonu, Kısıtlama programlaması, Google OR-Tools.

Abstract— Employee satisfaction and operational efficiency are two crucial aspects that this system addresses in different domains. Shift scheduling is a complex task with multiple constraints and preferences, which makes it necessary to find intelligent solutions. Our proposed system is based on the theoretical concepts of constraint programming. We implement Google's OR-Tools library, a versatile open-source suite known for effectively handling intricate scheduling constraints. The system focuses on the problem of personnel scheduling by optimizing shift distribution among agents. This system ensures that each shift is appropriately staffed while managing hard constraints like agent availability and rest periods. The system receives an Excel file containing information on agents, shift distribution metadata, and historical assignment data as its input. The output is an optimized schedule that is written back into an Excel file for future use, thereby reducing the possibility of errors and issues. Moreover, the system maintains a historical record of agent assignments, which affects future assignments and improves operational efficiency. This paper aims to comprehensively explain the optimization of shift management using constraint programming, with a specific focus on practical implementation using Google OR-Tools. This research is unique because it applies constraint programming solutions to a practical real-world issue, showcasing the potential of these tools in a corporate environment.

Keywords : Employee scheduling, Schedule optimization, Constraint programming, Google OR-Tools.

1. Introduction

Effective shift management is crucial for both operational success and employee satisfaction in today's business landscape. Shift scheduling can be challenging due to its various constraints and preferences. This study examines the use of constraint programming - specifically Google OR-Tools - to enhance shift management efficiency. Google OR-Tools is an open-source software suite recognized for its robust capabilities in tackling intricate scheduling problems. In addition to its various programming tasks, Google OR-Tools provides a complete library of algorithms and tools for shift management.

Google OR-Tools have been effectively used in various sectors—from employee scheduling in multi-shift organizations to manufacturing process scheduling. In particular, these tools have been utilized in the healthcare sector to optimize nurse scheduling, a task that is known for its complexity. This paper explores the practical application of these tools in shift management, demonstrating how they have the potential to improve operational efficiency and employee satisfaction.

This paper proposes a system that utilizes Google OR-Tools to create an automated shift scheduling system. The system optimizes the assignment of shifts to a group of agents while adhering to a set of constraints, ensuring that each shift is adequately staffed. The system receives an Excel file containing agent information, shift distribution metadata, historical assignment data, and the schedule from the previous week. It utilizes Google's Constraint Programming solver from the OR-Tools library to find a feasible or optimal solution based on these constraints. The optimized schedule is subsequently recorded in a new Excel file for future scheduling.

Constraint programming is a widely used technique in various sectors, especially in the area of staff scheduling. For example, Kanet et al. (2004) define constraint programming as a systematic approach for formulating and solving discrete variable constraint satisfaction or constrained optimization problems. In our project, we adopt this definition, which involves the application of constraint programming to resolve the shift scheduling problem.

In healthcare, constraint programming has been utilized to tackle staff scheduling issues. Bourdais et al. (2003) introduced a constraint programming model and search strategy to formulate and resolve these problems. The deployment of constraint programming in healthcare, with the aim of optimizing shift scheduling, aligns with our project's objective, though applied in a different domain.

Triska and Musliu (2011) detailed another research work on CP-Rota, a novel constraint programming technique employed in rotating workforce scheduling. An application is currently under development at their institute to solve complex scheduling problems. This aligns with our project's approach of utilizing constraint programming to solve the problem of shift scheduling.

In addition, constraint programming has been applied in the context of inbound call centers. Türker and Demiriz's (2018) study focuses on shift scheduling and rostering problems in the context of inbound call centers. They compare their constraint programming approach with existing models.

These literature examples illustrate how constraint programming can effectively and flexibly solve complex scheduling problems in various sectors. The goal of our project is to leverage this versatility and effectiveness to develop an automated shift scheduling system. Our project aims to showcase the practical application of constraint programming in a real-world environment, by providing a solution to the complex problem of shift allocation. This contributes to the ongoing efforts for enhancing operational efficiency and employee satisfaction in various sectors through intelligent scheduling.

This paper has four main sections. Section 2 provides a comprehensive Literature Review delving into the existing literature on constraint programming and its diverse applications. Next, we present the Methodology section which provides a detailed description of the project methodology, including the application design. Section four is an experimental study where we describe our experimental setup and analyze the results. In conclusion, we summarize our findings and discuss the potential avenues for future research in the Conclusion and Future Works section.

2. Literature Review

The use of constraint programming in shift scheduling has been a topic of interest in multiple research studies. This section provides a literature review of the most relevant works on the utilization of Google OR-Tools in this field.

The research conducted by Van den Bergh et al. (2013) provides an in-depth analysis of issues and solution approaches related to personnel scheduling. The authors emphasize the complexity of these matters, often characterized by the need to reconcile a variety of constraints and preferences. Additionally, they discuss the use of constraint programming as a resolution strategy, highlighting its versatility and effectiveness in managing complex constraints. Türker and Demiriz (2018) aim to optimize shift schedules and rostering for call centers that have peak demands in short timeframes by utilizing overlapping shift systems. The article suggests utilizing an integer programming model for shift scheduling and both integer and constraint programming models for rostering. These models are tested on real data, demonstrating better efficiency, with the constraint programming model for the rostering problem showing superior computational performance.

Burke et al. (2004) highlight the importance of efficient nurse rostering in healthcare, which is a complicated task. Optimizing nurse rostering can enhance workforce contentment and effectiveness. The article provides an interdisciplinary review of existing solutions, from operations research to artificial intelligence. It concludes by emphasizing the need to address specific challenges in future nurse rostering research. Mladenovic et al. (2004) put forward a novel methodology to dynamically reschedule train trips. They consider it as a special case of the job shop scheduling problem and employ a constraint programming approach to solve it. The methodology introduces three heuristic categories - separation, bound, and search - for efficient time management. The authors validate the method by using a prototype of train rescheduling software, indicating its potential usefulness in operational railway control. In a separate study, Harjunkoski et al. (2000) introduce two approaches that integrate mixed integer programming (MIP) and constraint logic programming (CLP) to handle the combinatorial complexity of large discrete optimization problems. The authors demonstrate these methods through job-shop scheduling and trim-loss problems and compare them with direct MIP and CLP problem solutions. Russell and Urban (2006) consider a two-phase constraint programming approach to schedule sports competitions across multiple neutral venues. The method achieves optimal solutions for up to sixteen teams and near-optimal solutions for up to thirty teams. The results indicate its favorable performance when compared to an integer goal programming approach.

Zeballos et al. (2010) proposed an integrated constraint programming model to solve intricate issues in flexible manufacturing systems, such as tool allocation, machine loading, part routing, and scheduling. The high computational performance and effective solutions of this model are due to the unique use of two sets of two-index variables that simplify manufacturing activities and indirectly represent tool needs. As a result, the dimensionality is reduced, and tool copies are excluded. Topaloglu and Ozkarahan (2011) developed a mixed-integer programming model to optimize the scheduling of residents' duty hours, considering both the Accreditation Council for Graduate Medical Education (ACGME) regulations and residency program demands. A subsequent column generation model was tested with real hospital data. It uses both a master problem for individual schedules and an auxiliary problem for feasible schedules, achieving high-quality schedules within seconds.

The authors conducted an extensive review and observed that although there is a significant amount of research on optimizing shift scheduling using constraint programming and other techniques, none of the existing research directly correlates the use of Google OR-Tools with shift scheduling. The gap in the literature presents an opportunity for further exploration and study. Google OR-Tools, with its powerful solvers and accessibility, offers a promising avenue for applying constraint programming to shift scheduling. However, the lack of direct application in existing literature suggests that there may be unexplored challenges or opportunities in this specific area of application.

In conclusion, the literature review has shown the wide application of constraint programming and other optimization techniques to solve complex scheduling problems across various sectors. These techniques have shown their versatility and effectiveness in various studies, including call center scheduling, nurse rostering, train trip rescheduling, sports competition scheduling, and manufacturing systems. However, the direct use of Google OR-Tools in shift scheduling has not been thoroughly explored in the literature. The study aims to demonstrate the practical use of Google OR-Tools in developing an automated shift scheduling system, thus filling this gap. The study's findings will contribute to the current knowledge of shift scheduling and provide a basis for future research in this field.

3. Methodology

In this study, we utilize Google's OR-Tools, a powerful suite of operations research software, to implement and solve our constraint programming model. OR-Tools is an open-source software suite that provides a wide array of powerful algorithms for combinatorial optimization problems. The core of this suite is its support for constraint programming, which is a declarative method for defining combinatorial problems in terms of constraints that must be satisfied by the solution.

Constraint programming is particularly well-suited for solving complex scheduling, routing, and allocation problems, where the solution space is large and the constraints are intricate. Google's OR-Tools provide a robust and efficient constraint solver that can handle these types of problems. The solver uses a variety of techniques, including constraint propagation and domain reduction, to prune the search space and find optimal solutions more efficiently. It also supports a range of global constraints and decision strategies, which can be used to further guide the search process.

Figure 1 illustrates the basic operation of the search algorithm in OR-Tools, which consists of the following steps:

1. Initially, the search algorithm assigns variables, which are usually empty.
2. Then, the algorithm selects a variable and an untested value for that variable. The selection of the variable and value can be influenced by a search strategy.
3. Afterward, the algorithm tries to assign the chosen value to the chosen variable. When the assignment results in a conflict, that is, when a constraint violation occurs, the algorithm backtracks and makes an attempt with a different value.
4. If the assignment does not result in a conflict, the algorithm continues the search recursively with the new assignment.
5. The search ends when all variables are assigned and all constraints are satisfied, or when all possible assignments have been attempted, and none of them satisfy all constraints.

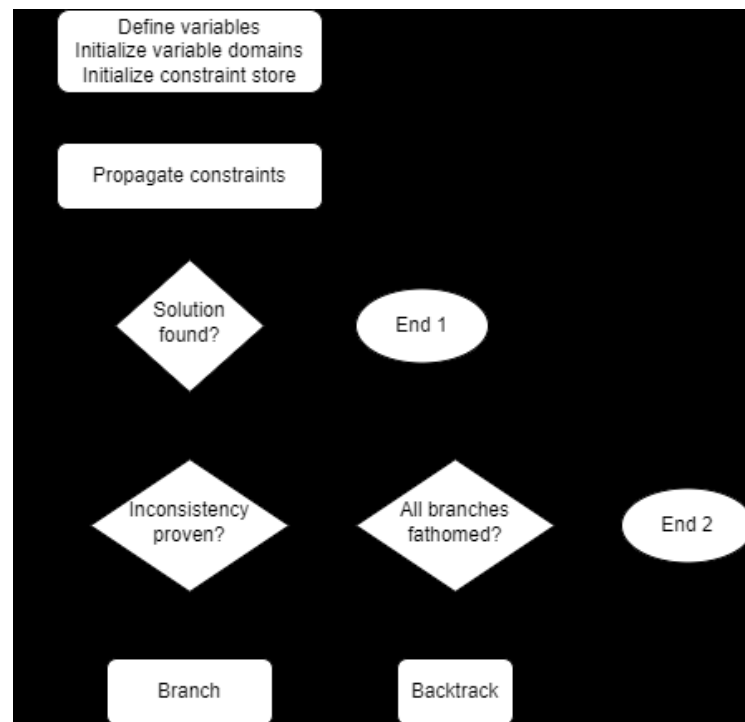


Figure 1. A general algorithm for solving constraint satisfaction problems (CSPs) (Kanet et al., 2004, 47-3).

Constraint programming provides a constraint-independent method to generate initial solutions via tree-based search with constraint propagation. This method can locate feasible first solutions that can later be enhanced via local search. Local search is an iterative heuristic method that improves a solution via small modifications. By combining complete and local search, we can utilize local search to find good solutions and then establish an upper bound on the cost function for proving optimality through complete search. Alternatively, we can run local and complete searches in parallel, exchanging the upper bound from local search to complete search and the solutions found in each direction. Elite solutions can be used in local search, for example, to provide good starting points for strategic restarting. The combination of tree-based and local searches can be a powerful optimization approach since it enables flexibility in generating and improving initial solutions.

OR-Tools solve combinatorial optimization problems, such as routing, scheduling, and vehicle routing problems, among others. The CP-SAT solver, which forms a part of Google's OR-Tools, solves problems for the Constraint Programming model. The algorithmic design and behavior of the CP-SAT solver are important to understand; here are some important points:

- **Modeling with Boolean Variables:** At the core, the CP-SAT solver models the problem as a Boolean Satisfiability Problem (SAT). This includes transforming integer variables and constraints into a set of Boolean variables and constraints. This is usually done by a process known as "domain encoding" or "integer encoding" where each possible value of an integer variable can be represented as a unique Boolean variable.
- **Presolve Phase:** During this phase, the solver tries to simplify the model and reduce the search space. This could include domain reduction, constraint simplification, and constraint expansion.
- **Constraint Propagation:** During the search process, the solver continuously applies the constraints to reduce the domains of the variables. This is known as "constraint propagation". It's one of the most important techniques used in CP solvers to efficiently prune the search space.
- **Explanations/Conflict Resolution:** Whenever the solver determines that a certain assignment of variables violates a constraint, it creates an "explanation" for that conflict.

These explanations are used to prune the search tree and avoid exploring similar infeasible solutions in the future.

- **Search Strategies:** CP-SAT solver uses various search strategies, including default search, fixed search, pseudo cost branching (inspired by Mixed Integer Programming solvers), and Large Neighborhood Search (LNS).
- **Multithreading:** CP-SAT solver also supports multithreading. It runs several workers in parallel, each with a different set of parameters.
- **Handling Large Domains and Complex Constraints:** For certain types of constraints or large domains, special techniques are used. For example, for "all different" constraints, it's common to treat them as a set of boolean variables with a sum equal to one to encode the domain constraints. Also, for constraints like the "circuit" or scheduling constraints, specific propagation techniques are used.

OR-Tools is implemented in C++, with interfaces in Python, Java, and .NET, making it accessible to a wide range of users. In this study, we utilized the Python interface to define and solve our model. The OR-Tools solver enables us to define our problem in a high-level, declarative manner, without needing to concern ourselves with the underlying search algorithms. Once defined, the solver locates the optimal solution, allowing the focus to remain on the problem, rather than the solution process itself.

3.1. Application Design

Our study aimed to develop an effective shift scheduling tool for the Wingie Enuygun Group call center. The tool was designed to optimize the distribution of shifts among agents, guaranteeing the call center's constant adequate staffing while also taking into account agents' preferences and historical working patterns. Implementation of the tool was performed by developing a web application using Flask, a Python-based web framework.

The application consists of two primary components. The first component handles user interactions, such as uploading and downloading files. Users are able to upload a file containing information about the present shift distribution and agent roster. Subsequently, the application produces an optimized shift schedule that can be downloaded in the same format as the initial file.

The optimization process for shift scheduling is performed in the second component of the application. The component utilizes an optimization suite to generate a shift schedule that optimizes available data. The optimization process takes various constraints into account, including the requirement that agents work precisely five days per week and rest for at least 11.5 hours between shifts. The tool also considers agents' previous working schedules to generate a shift schedule that closely aligns with their preferred working hours.

A random search algorithm is employed in the shift scheduling tool to optimize the schedule for agents. The algorithm explores the solution space iteratively by introducing random changes to the current solution in order to find a new one. The exploration continues until a solution is found that satisfies all constraints, has the lowest possible cost, or until a specified time limit is reached.

The tool incorporates a logging functionality that documents the specifics of every file upload request and the outcomes of the shift scheduling procedure. The inclusion of this feature is valuable for the purposes of resolving issues and comprehending how the tool is utilized.

In summary, our methodology comprised creating a web-based shift scheduling tool that implements an optimization algorithm to produce an optimal shift schedule founded on input data and different constraints. The design of the tool is focused on user-friendliness and efficiency, rendering it a valuable asset for the management of shifts within a call center environment.

4. Experimental Study

To evaluate the effectiveness of our shift scheduling tool, we conducted an experiment where the tool was utilized to generate a new weekly shift schedule for the call center of Wingie Enuygun

Group. The aim of the experiment was to monitor the duration taken to generate the new schedule and compare it with the duration taken using the previous manual approach.

For the experiment, the input data was an Excel file (as in Table 1) comprising the current shift distribution and agent information. The file was uploaded to the tool, which employed its optimization algorithm to produce a new shift schedule that fulfilled all constraints and requirements. We documented the duration taken to generate this schedule.

Table 1. This is a hypothetical example of the preferred shift distribution among various groups of agents. The table displays the essential metadata for the distribution of shifts throughout the week.

Group	Day	Shift 1	Shift 2	Shift 3	Shift 4
Group 1	Monday	3	2	2	9
Group 1	Tuesday	2	2	2	9
Group 1	Wednesday	2	2	2	8
Group 1	Thursday	2	1	2	9
Group 1	Friday	2	1	2	9
Group 1	Saturday	3	1	2	8
Group 1	Sunday	3	1	2	8

The role of the CP-SAT solver is crucial in the context of our experiment. This is illustrated in Table 1. The solver is designed to align as closely as possible with the metadata and strive for finding the optimal solution. Yet, due to the multitude of constraints and the number of available agents, finding an optimal solution within the allotted time may not always be possible. The solver outputs the most feasible solution it can find in such cases. This emphasizes the importance of metadata creation in the process. The metadata provides a framework that guides the solver in generating the most efficient shift schedule. Thus, creating metadata carefully and strategically is critical to effective shift scheduling.

4.1.1. Experimental Setup

To ensure reproducibility and comparability, it is necessary to note the system specifications on which the shift scheduling tool was executed. The tool was run on a system containing a 13th Generation Intel(R) Core(TM) i9-13900HX processor, featuring 32 cores, and operating at a base frequency of 2.20 GHz.

The system's processing power on which the tool is executed is likely to affect its performance. Thus, for future work aiming to reproduce our results or compare the performance of other shift scheduling tools, considering the system's specifications is recommended.

4.1.2. Experimental Results

In the initial stages of our experiment, we adjusted multiple parameters within the 'CpSolver class' through careful calibration. These parameters, specifically 'parameters.max_time_in_seconds', 'parameters.num_search_workers', and 'parameters.cp_model_probing_level', were identified as having the potential to influence the performance of our shift scheduling tool. In various trials, we adjusted the 'parameters.max_time_in_seconds' parameter which sets the time limit for the solver in an attempt to balance solution quality and computation time. Likewise, we adjusted the 'parameters.num_search_workers' parameter, which regulates the number of workers used in the search. Our observations showed that raising the number of workers notably increased the speed of finding solutions, especially on multi-core machines. Finally, we adjusted the 'parameters.cp_model_probing_level' parameter, which controls the level of probing conducted by the solver. We observed that higher levels of probing can enhance solution efficiency, but it also increases

computation time. The insights from these initial experiments were crucial in optimizing the performance of our shift scheduling tool.

After these initial trials, we proceeded to the main phase of our experiment. The results indicated a substantial improvement in the time taken to generate a shift schedule compared to the previous manual method. In the past, creating a schedule that met all constraints and requirements took more than 5 hours of manual effort. The process required a deep understanding of the agents' shift requirements and preferences and involved a significant amount of trial and error. By using our shift scheduling tool, the time it takes to produce a new schedule has been reduced to under a minute. This has resulted in a performance improvement of more than 300%. The drastic reduction in time is attributable to the efficiency of the tool's optimization algorithm, which can rapidly explore the solution space and locate a schedule that meets all constraints.

Apart from saving time, the tool is associated with other benefits. The tool eradicates the possibility of human error in the scheduling process, ensuring that the generated schedule adheres as closely as possible to the agents' preferred working hours. Additionally, the tool offers a consistent and repeatable process for shift scheduling, which can be conveniently modified to adapt to changes in shift requirements or agent preferences.

To conclude, our shift scheduling tool represents a significant improvement over the previous manual method of shift scheduling. The tool not only significantly reduces the time required to generate a new schedule but also offers a more reliable and flexible process for shift scheduling.

5. Conclusion and Future Works

This study has demonstrated the practical application of Google's OR tools in the development of an automated shift scheduling system. Within a call center, the system, based on constraint programming, significantly improved operational efficiency and employee satisfaction. The tool not only reduces generating time of a new schedule but also provides a more trustworthy and adaptable process for shift scheduling. It removes the possibility of human error in scheduling and guarantees that the generated schedule meets the agent's demands for preferred working hours.

The usage of OR-Tools by Google has addressed an area that has not been focused on in academic research, and it has shown the potential of these tools in a business environment. The adaptability and efficiency of constraint programming have been showcased, solving intricate scheduling complications in various industries. This has added to the continuous endeavor to boost operational efficiency and employee satisfaction by implementing intelligent scheduling techniques.

There are many potential areas for future research. Exploring other optimization algorithms in the OR-Tools suite is an interesting topic to consider. Other solvers, such as the linear solver or the routing solver, could also be used to solve shift scheduling problems, although the CP-SAT solver was utilized in this study. Furthermore, incorporating machine learning techniques to predict future shift requirements based on historical data could enhance the tool, improving the accuracy and efficiency of the scheduling process.

This tool could also be applied in other sectors, such as healthcare, manufacturing, or retail, where shift scheduling is also a critical aspect of operations. Additionally, an intriguing expansion to this research could be the inclusion of functionality that assesses the preferred working shifts of the agents. This inclusion may require the procurement of data on the agents' shift preferences and the integration of this data into the scheduling process. Implementing this modification would not only enhance the satisfaction of the agents but, in addition, could potentially improve productivity and efficiency in the call center.

C) References

Aksin, Z., Armony, M., & Mehrotra, V. (2007). The modern call center: A multi-disciplinary perspective on operations management research. *Production and operations management*, 16(6), 665--688. Wiley Online Library. <https://doi.org/10.1111/j.1937-5956.2007.tb00288.x>

- Bourdais, S., Galinier, P., & Pesant, G. (2003). HIBISCUS: A Constraint Programming Application to Staff Scheduling in Health Care. In *Principles and Practice of Constraint Programming - CP 2003: 9th International Conference, CP 2003, Kinsale, Ireland, September 29 - October 3, 2003, Proceedings* (Vol. 2833, pp. 153-167). Springer Berlin Heidelberg. https://doi.org/10.1007/978-3-540-45193-8_11
- Burke, E. K., De Causmaecker, P., Berghe, G. V., & Van Landeghem, H. (2004). The State of the Art of Nurse Rostering. *Journal of Scheduling*, 7, 441-499. <https://doi.org/10.1023/B:JOSH.0000046076.75950.0b>
- Harjunkski, I., Jain, V., & Grossman, I. E. (2000). Hybrid mixed-integer/constraint logic programming strategies for solving scheduling and combinatorial optimization problems. *Computers & Chemical Engineering*, 24(2), 337-343. [https://doi.org/10.1016/S0098-1354\(00\)00470-1](https://doi.org/10.1016/S0098-1354(00)00470-1)
- Kanet, J. J., Ahire, S. L., & Gorman, M. F. (2004). Constraint Programming for Scheduling. In *Handbook of Scheduling: Algorithms, Models, and Performance Analysis* (47-1 to 47-21). CRC Press. https://ecommons.udayton.edu/mis_fac_pub/1
- Mladenovic, S., Markovic, M., Cangalovic, M., Vujaklij, D., & World Conference on Transport Research Society. (2004). Constraint Programming Approach to Train Scheduling on Railway Network Supported by Heuristics. 10th World Conference on Transport Research. <https://trid.trb.org/view/844145>
- Russell, R. A., & Urban, T. L. (2006). A constraint programming approach to the multiple-venue, sport-scheduling problem. *Computers & Operations Research*, 33(7), 1895-1906. <https://doi.org/10.1016/j.cor.2004.09.029>
- Topaloglu, S., & Ozkarahan, I. (2011). A constraint programming-based solution approach for medical resident scheduling problems. *Computers & Operations Research*, 38(1), 246-255. <https://doi.org/10.1016/j.cor.2010.04.018>
- Triska, M., & Musliu, N. (2011). A constraint programming application for rotating workforce scheduling. In *Developing Concepts in Applied Intelligence* (Vol. 363, pp. 83-88). Springer Berlin Heidelberg. https://doi.org/10.1007/978-3-642-21332-8_12
- Türker, T., & Demiriz, A. (2018). An integrated approach for shift scheduling and rostering problems with break times for inbound call centers. *Mathematical Problems in Engineering*, 2018, 1-19. <https://doi.org/10.1155/2018/7870849>
- Van den Bergh, J., Beliën, J., De Bruecker, P., Demeulemeester, E., & De Boeck, L. (2013). Personnel scheduling: A literature review. *European journal of operational research*, 226(3), 367-385. <https://doi.org/10.1016/j.ejor.2012.11.029>
- Zeballos, L. J., Quiroga, O. D., & Henning, G. P. (2010). A constraint programming model for the scheduling of flexible manufacturing systems with machine and tool limitations. *Engineering Applications of Artificial Intelligence*, 23(2), 229-248. <https://doi.org/10.1016/j.engappai.2009.07.002>

Anomaly Detection-Based Web Application Firewall Using Machine-Learning Techniques

Özgün Karadeniz^{*1} , Erdem Yelken¹ 

¹Medianova R&D Center, Şehit Ahmet Sokak No:4 Mecidiyeköy İş Mer. 15th Floor, Şişli/İstanbul, Turkey

(ozgun.karadeniz@medianova.com, erdem.yelken@medianova.com)

Abstract—

Today, as the prevalence of web applications and services increases, so does their vulnerability to cyber-attacks. Web Application Firewalls are effective tools for protecting web applications, but they may not be able to keep up with the pace of innovation. While rules-based approaches are suitable tools for detecting known threats, they can generate large numbers of false positives and must be complemented with other security measures to detect new threats. This article explores the potential of machine learning-assisted anomaly detection as an alternative. Using a dataset of 1.6 million HTTP requests of real-time Medianova CDN traffic and CSIC 2010, this study aims to optimize WAF performance to both detect new attack threats and reduce the false positive rate. Machine learning models combined with NLP techniques are used to detect malicious HTTP requests in real-time weblogs. The result demonstrates that Logistic Regression has the best result for Medianova dataset, and Naïve Bayes has the best result for CSIC 2010 dataset in terms of accuracy. It promises a more secure and efficient web ecosystem.

Keywords: *Web Application Firewalls, Intrusion Detection Systems, Logistic Regression, Machine Learning.*

1.Introduction

With the increase in internet usage, there is a corresponding rise in both the number and the variety of web applications and web services (Statista, 2021). As a result of this surge, web applications and services are becoming attractive targets for cyber attackers. Therefore, preventing web attacks is one of the most important issues. Although Web Application Firewall (WAF) products are quite advanced today, there are still attacks that bypass WAFs.

Rule-based approaches continue to be actively employed for the detection and mitigation of malicious Hyper Text Transfer Protocol (HTTP) requests aimed at applications. While these methods are typically effective in blocking simple and known attacks, they may prove inadequate to prevent emerging attacks. Moreover, such approaches may classify benign HTTP requests as malicious mistakenly, leading to the generation of false positives (Follini, 2016). The operation or logic of some applications might appear suspicious, capable of potentially triggering a WAF rule. However, this arises from the fact that the applications are designed in such a way to serve a specific purpose. To overcome these limitations, the adoption of anomaly detection strategies can be implemented.

In previous studies, anomaly detection for HTTP traffic has been performed using statistical analysis (Denning, 1987; Smaha, 1988; Ye et al., 2002). However, this approach has limitations, as not all behaviours can be expressed using statistical models. Anomaly detection, harnessing the power of machine learning, has emerged as a powerful tool for identifying deviations from established norms. A survey in Chandola et al., (2009) outlines anomaly detection techniques that have been developed for many different applications. In anomaly detection approaches, the specific use of machine learning for network intrusion detection is crucial to detect new attacks and reduce the rate of false positives arising from rule-based WAF applications. This can be achieved only through the training of machine learning models with normal requests and the subsequent detection of anomalous requests.

HTTP traffic data is typically characterized by hundreds or thousands of features (Cai et al., 2018; Cui et al., 2018). However, among these features, some may be irrelevant and redundant features that can increase the complexity of the anomaly detection process and reduce its effectiveness. Feature selection is an important task for filtering HTTP traffic in WAFs. Effective feature selection can greatly improve the performance of WAFs. By reducing the number of features, WAFs can process traffic more quickly and efficiently. This can lead to improved detection accuracy and reduced system resources.

The rest of the paper is organized as follows: A brief review of related works that shows the previous and recent studies for web-based attack detection systems is presented in Section 2. In Section 3, the methodology encompasses discussions on the utilized dataset definitions, data pre-processing steps, the machine learning models used, experimental results, and comparisons. Finally, in Section 4, the obtained results are summarized.

2. Related Work

Anomaly-based detection is a crucial approach for Intrusion Detection Systems (IDSs) for preventing malicious requests. Feature selection and extraction are the crucial parts that reduce the complexity of the problem and increase the anomaly detection performance of the models (Nguyen et al., 2011; Gupta and Singh, 2017; Li et al., 2018; Li et al., 2020). There are studies in which NLP-based techniques are applied, beyond the traditional statistical feature selection and extraction methods (Torrano-Gimenez et al., 2011; Khreich et al., 2017; Pal and Chowdary, 2018; Vartouni et al., 2018; Zhang et al., 2019). In this work (Liu et al., 2020), they analyse and use feature selection methods and classify the HTTP request with the Support Vector Machine (SVM) algorithm. The study revealed that the model obtained using the the Spanish National Research Council (CSIC) 2010 data set detected the attacks with a high accuracy score.

Machine learning methods are one of the most noteworthy techniques for detecting anomalies. These methods provide algorithms that require training to be able to solve a problem. It also includes various techniques for solving many different classification and regression problems. Machine learning algorithms offer powerful methods for handling both supervised and unsupervised datasets. In intrusion detection systems, this approach can continually improve as new data is introduced. Therefore, machine learning algorithms for log-based anomaly detection have been proposed recently to fill the deficiency of rule-based systems (Meng et al., 2019; Zhang et al., 2019; Fält et al., 2021; Mandagondi, 2021).

Hoang (2021) proposed a decision tree-based machine learning model and the TF-IDF (Term-Frequency - Inverse Document Frequency) method to turn each row into a vector. The proposed model focused on web server logs that contain the HTTP requests and classified these logs into two categories: Attacked and Normal. As a result of the work, the model was able to detect common attack types like SQL injection (SQLi) and Cross-site Scripting (XSS). Duy et al., (2019) performed anomaly detection work by exploiting the attributes of user behaviour in HTTP requests. Their methodology involves the usage of both TF-IDF and common feature extraction approaches paired with a random forest algorithm. Notably, TF-IDF surpassed the common feature approach by achieving 4% better results.

In addition to these studies, more recent works have been performed in the literature. Shaheed et al. (2022) developed a web application firewall (WAF) that utilizes machine learning and feature engineering to enhance security against malicious traffic, focusing on HTTP requests, a methodology similar to our work. Further, Toprak et al., (2022) use deep learning to identify malicious requests. This study was combined with two layers. They parsed online HTTP requests and extracted their features with a pre-processing module. The Deep Neural Network (DNN) was trained to recognize and flag malicious requests. Compared to traditional techniques, using deep learning in this context sometimes offers improved accuracy and efficiency in detecting malicious activities.

3. Method

Rapid and precise identification of malicious requests is paramount to shield systems from substantial risks. In this research, our emphasis is on classifying incoming requests into 'normal' or 'attack' categories. We employ a variety of machine learning algorithms, including Logistic Regression, Naive Bayes, Decision Tree, and Random Forest. These models were meticulously trained and evaluated using data pre-processing techniques like Natural Language Processing (NLP). The data utilized for this study consists of real-time HTTP requests from Medianova Content Delivery Network (CDN) traffic and CSIC 2010.

Our methodology, presented in Figure 1, depicts both the training and real-time anomaly detection processes side by side. On the left of the figure, the training process shows the progression from "HTTP request storage" to "data pre-processing", and finally to "Training". This includes our systematic approach to preparing and pre-processing data before it is fed into our algorithms for learning.

On the right side of Figure 1, the anomaly detection process is showcased. It starts with "real-time network traffic", proceeds with "feature extraction" and then goes through the "anomaly detection score" stage using trained machine learning models, then results in "classification". The endpoint of this pipeline is the classification of traffic as either "abnormal" or "normal".

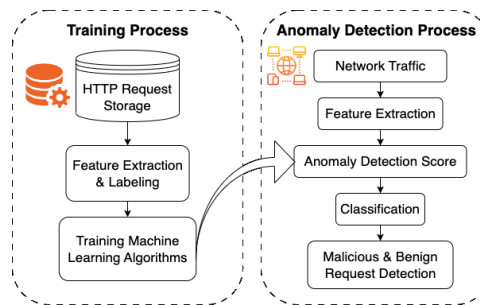


Figure 1. The Flowchart of Training and Anomaly Detection Processes

3.1. Data

The dataset utilized in this study was derived from real-time Medianova CDN traffic and consists of over 1.6 million HTTP requests. This extensive and diverse data source serves as a crucial foundation for developing effective mechanisms to detect web application attacks. The richness of the dataset enables the machine learning models to adapt to real-world attack vectors, and meticulous attention has been given to addressing class imbalance issues, ensuring fair and reliable outcomes. By leveraging this dataset, the trained machine learning models demonstrate their capability to successfully safeguard web applications, offering a robust solution to security challenges. A second dataset, CSIC 2010, was also employed in this study. It contains over 25,000 malicious requests and 36,000 benign requests, and was artificially generated using a semi-automatic tool by CSIC.

3.2. Data Pre-processing

To apply machine learning techniques for classifying abnormal or normal behaviour in web traffic, it is crucial to build labelled datasets for training. Pre-processing plays a vital role in preparing the collected dataset to be used effectively for training the machine learning models. In this step, we handle raw HTTP requests containing various headers and data fields. Given the complex and multifaceted nature of these requests, it is imperative to simplify data for efficient processing. We used a pre-processing module to reduce the complexity inherent in raw HTTP requests. Common header names, which are considered not necessary for the context of our study, were not considered, and headers such as "Method", "X-Forwarded-For", "URI", "Host", "Content-Length" and "User-Agent" were chosen.

In Figure 2, we illustrate the methodical sequence of data pre-processing steps vital for refining raw textual content and paving the way for more effective integration of NLP (Natural Language Processing) techniques. To commence, text data is uniformly transformed to lowercase to ensure uniformity and extraneous gaps are eliminated correspondingly. Moreover, HTML tags and punctuation marks are excluded from the text content. After this, superfluous components, including numerical figures and punctuation symbols, are filtered out. The software then employs a predetermined inventory to weed out frequent yet semantically inconsequential 'stop words' (such as 'and', 'however', 'whether'). This action significantly bolsters the calibre of subsequent textual analysis. The subsequent process of lemmatization ensues, whereby words are simplified to their elemental structures, culminating in a consistent representation of identical words formulated with diverse inflections. These pre-processing steps collectively simplify the text, reducing complexity and making it consistent, thus creating a strong base for in-depth analysis.

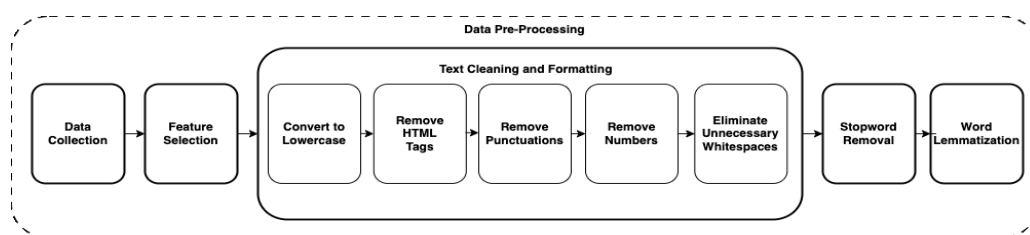


Figure 2. Flowchart of the sequential data pre-processing steps

In order to apply supervised machine learning techniques, the collected data needs to be labelled. This involves classifying each data instance into one of two classes: "abnormal" or "normal" behaviour. Experts or manual inspection might be required for accurate labelling.

Once the data is cleaned and labelled, the next step is to extract relevant features from each request in the dataset. Feature extraction is crucial as it transforms raw data into a format suitable for the machine learning model. In this step, we used the TF-IDF (Term Frequency-Inverse Document Frequency) vectorizer. Figure 3 shows a simple example of our data pre-processing steps. This approach ensures that our models are trained on data that encapsulates the core of HTTP requests while being free of the noise and redundancy typically found in raw requests. The pre-processed dataset is then divided into a training set (70%) and a test set (30%). The machine learning model is trained with the training dataset and evaluated on the test dataset to ensure its ability to generalize well to unseen data.

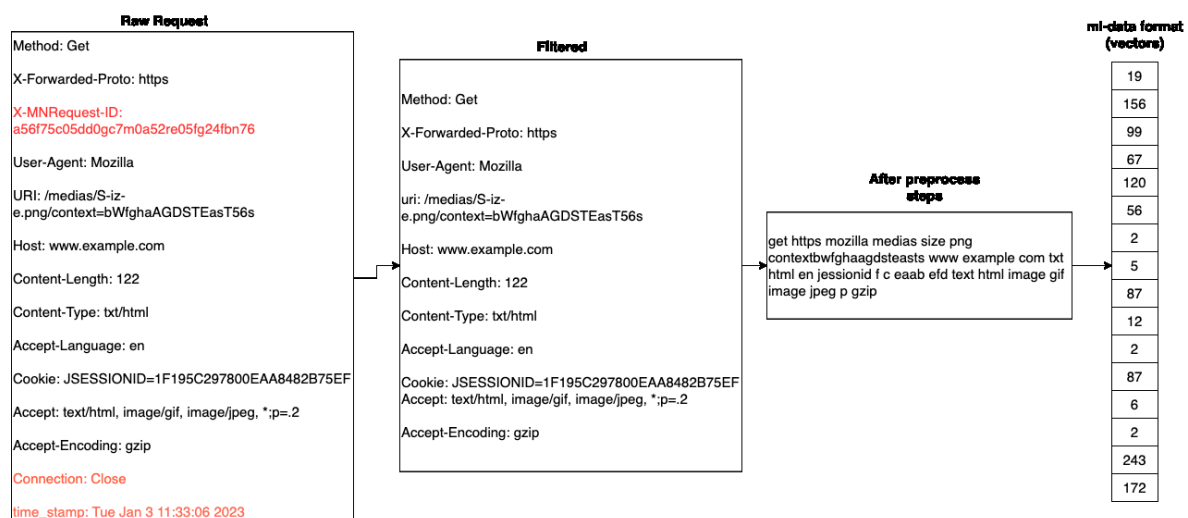


Figure 3. The workflow of data pre-processing: Transitioning to machine learning compatible formats

3.3. Models

In this study, we use four different machine learning models that are effective in binary classifier systems to perform WAFs anomaly detection Kumari et al., (2017). Below, we describe the key models used in the system and the corresponding methodologies. The first model utilized is Logistic Regression, a popular linear classifier ideal for binary classification tasks. In this work, using TF-IDF vectors as input features, the Logistic Regression model accurately predicts the probability that HTTP requests will be classified as malicious (1) or legitimate (0). It processes the relationship between the categorical dependent variable and one or more independent variables by estimating probabilities using a logistic function Hosmer et. al., (2013). The second model included in this study is the Naive Bayes (NB) model, a powerful probabilistic classifier based on Bayes' theorem, chosen for its effectiveness in computing probabilities where feature independence is a reasonable assumption Zhang (2004). It computes the probability that a data point belongs to a specific class with the assumption that all features in the data point are independent of each other. This assumption makes it particularly suitable for our classification task, where feature independence can be considered a reasonable approach. Next, The Random Forest model, an ensemble method using multiple decision trees, was chosen for this study due to its ability to produce accurate predictions by combining weak learners Cutler et. Al., (2012). Utilizing the bagging technique, it trains decision trees with different bootstrap samples, handling both numeric and categorical data. The capacity of this model to process both numerical and categorical data, combined with the ensemble approach, makes it a viable choice for WAF anomaly detection. Finally, we utilized the Decision Tree model for the classification of the HTTP request as normal or abnormal. This model, which has a flowchart-like structure, provides simple interpretability, and is especially used to break down complex decision-making processes into simpler steps Myles et al., (2004).

3.4. Experimental Results

The experiments for this study were conducted on a machine equipped with an AMD Opteron(tm) Processor 6386 SE CPU @ 2.79GHz, 64 CPUs, and 62GB of RAM.

Algorithm: Web Based Anomaly Detection

Input:

$RR = \{r_1, r_1, r_1, \dots, r_n\};$

RR denotes Real-time HTTP Requests; r_k is the number of HTTP request.

$ML = \{Logistic\ Regression, Naive\ Bayes, Decision\ Tree, Random\ Forest\};$

Output:

$R_{output} = \{Accuracy, Precision, Recall, F1 - Score\};$

1: Begin

2: $|r_k| = preprocessing(r_k)$

3: Initialize an empty set TFIDF_vectors, Labels;

4: **while** each term in $|r_k|$ **do**

5: TF = CalculateTermFrequency($|r_k|$)

6: IDF = CalculateInverseDocumentFrequency($|r_k|$, RR);

```

7:      TFIDF_vector = TF * IDF;

8:      Label = ClassifyRequest(TFIDF_vector);

9:  end

10: Initialize Training and Test dataset {X_train, y_train, X_test, y_pred}

11: Training data: (X_train, y_train), X_train: TF-IDF vectors, y_train: labels

12: for each ML in ML do:

13:     TrainML(X_train, y_train);

14:     y_pred=ML(X_test);

15: end

16: Calculate Training Results: Routput

17: End

```

Algorithm processes real-time HTTP requests and then calculates vectors of the text contents in each request via TF-IDF. Then it prepares the training and test data. Training data (X_train, Y_Train) contains TF-IDF vectors and labels. Then, training is carried out using 4 different machine learning algorithms that we prefer to use in the study. After the training is completed, label prediction is made using these trained models on the test data.

Table 1. Training Results of Machine Learning Algorithms

Study	Method	Dataset	Accuracy
Hoang (2021)	TF-IDF and decision tree	HTTP Param	98.5
Duy et al., (2019)	3 word n-gram and random forest	CSIC 2010	99.5
Işiker et al., (2022)	Character n-gram and linear support vector machines	CSIC 2010	99.5
Proposed Model	TF-IDF and Logistic Regression	CSIC 2010	90.8
	TF-IDF and Naive Bayes		93.0
	TF-IDF and Decision Tree		89.8
	TF-IDF and Random Forest		89.9
	TF-IDF and Logistic Regression	Our Dataset	99.6
	TF-IDF and Naive Bayes		94.7
	TF-IDF and Decision Tree		98.1
	TF-IDF and Random Forest		92.3

The comparison between our proposed model and existing studies in the literature is summarized in Table 1, highlighting several important insights. While previous studies such as Hoang (2021), Duy et al., (2019), and İşiker et al. (2022) have achieved high accuracies ranging from 98.5% to 99.53% on various datasets, our approach demonstrates competitive and even superior performance in certain cases. Specifically, on the CSIC 2010 dataset, our method achieved accuracies ranging from 0.90 to 0.91, which are slightly below the results reported in literature. However, on our dataset, the proposed model reached an impressive accuracy of up to 0.99 using TF-IDF and Logistic Regression, showcasing a significant advancement in the performance over the specific data domain.

Our study's contribution to the literature lies in the utilization of real-time CDN traffic, representing a dataset with diverse and unique user behaviours. Despite the complexity and variability inherent in such a real-world dataset, our proposed model has demonstrated highly competitive accuracies. This not only underscores the robustness and adaptability of our approach but also positions our work as a promising advancement in the field of web application security, offering insights and methodologies that can be leveraged across different data domains and real-world scenarios.

4. Conclusion

As the complexity and variety of web threats evolve, security tools such as rule-based WAFs face the rapid pace of these evolving threats. Aiming to address these challenges, our work highlights the integration of machine learning techniques, particularly anomaly detection, into WAF systems to improve threat detection capabilities.

Data pre-processing steps were applied both Medianova CDN traffic and CSIC 2010 datasets and machine learning models were trained and tested. We have demonstrated the efficiencies of various classifiers such as Logistic Regression, Naive Bayes, Decision Tree, and Random Forest. Especially Logistic Regression was the model that gave the most impressive results with Medianova dataset and Naïve Bayes for CSIC 2010 dataset with respect to accuracy score for anomaly detection. In addition, Naive Bayes, and Decision Tree models also performed as effectively as Logistic Regression.

Acknowledgments

This work was supported by the Scientific and Technological Research Council of Turkey Turkish (TUBITAK), Grand No: 3220054.

References

- Cai, H., Zheng, V. W., Chang, K. C.-C. J. I. t. o. k. ve engineering, d., 2018, A comprehensive survey of graph embedding: Problems, techniques, and applications, 30 (9), 1616-1637.
- Chandola, V., Banerjee, A. ve Kumar, V. J. A. c. s., 2009, Anomaly detection: A survey, 41 (3), 1-58.
- Cui, P., Wang, X., Pei, J., Zhu, W. J. I. t. o. k. ve engineering, d., 2018, A survey on network embedding, 31 (5), 833-852.
- Cutler, A., Cutler, D. R., & Stevens, J. R. (2012). Random forests. Ensemble machine learning: Methods and applications, 157-175.
- Denning, D. E. J. I. T. o. s. e., 1987, An intrusion-detection model, (2), 222-232.
- Duy, P. H., Thuy, N. T. T. ve Diep, N. N. J. S. A. J. o. S., 2019, Anomaly detection system of web access using user behavior features, 7 (2), 115-132.
- Fält, M., Forsström, S. ve Zhang, T., 2021, Machine learning based anomaly detection of log files using ensemble learning and self-attention, *2021 5th International Conference on System Reliability and Safety (ICSRS)*, 209-215.

- Follini, C. (2016). Apache Tutorial 8: Handling false positives with the ModSecurity Core Rule Set. Netnea. https://www.netnea.com/cms/apache-tutorial-8_handling-false-positives-modsecurity-core-rule-set/
- Gupta, J. ve Singh, J. J. I. J. o. A. R. i. C. S., 2017, Detecting Anomaly Based Network Intrusion Using Feature Extraction and Classification Techniques, 8 (5).
- Hoang, X. D., 2021, Detecting common web attacks based on machine learning using web log, *Advances in Engineering Research and Application: Proceedings of the International Conference on Engineering Research and Applications, ICERA 2020*, 311-318.
- Hosmer Jr, D. W., Lemeshow, S., & Sturdivant, R. X. (2013). Applied logistic regression (Vol. 398). John Wiley & Sons.
- IŞiker, B., & Soğukpınar, İ. (2021, December). Machine learning based web application firewall. In 2021 2nd International Informatics and Software Engineering Conference (IISEC) (pp. 1-6). IEEE.
- I. S. Institute, HTTP DATASET CSIC 2010, [On- line]. Available: <https://www.tic.itefi.csic.es/dataset/>, (Accessed on 18 December 2014), 2012.
- Khreich, W., Khosravifar, B., Hamou-Lhadj, A., Talhi, C. J. I. ve Technology, S., 2017, An anomaly detection system based on variable N-gram features and one-class SVM, 91, 186-197.
- Kumari, R., & Srivastava, S. K. (2017). Machine learning: A review on binary classification. *International Journal of Computer Applications*, 160(7).
- Li, H., Guo, W., Wu, G. ve Li, Y., 2018, A RF-PSO based hybrid feature selection model in intrusion detection system, *2018 IEEE Third International Conference on Data Science in Cyberspace (DSC)*, 795-802.
- Li, M., Wang, H., Yang, L., Liang, Y., Shang, Z. ve Wan, H. J. E. S. w. A., 2020, Fast hybrid dimensionality reduction method for classification based on feature selection and grouped feature extraction, 150, 113277.
- Liu, C., Yang, J., Wu, J. J. E. J. o. W. C. ve Networking, 2020, Web intrusion detection system combined with feature analysis and SVM optimization, 2020, 1-9.
- Mandagondi, L. G., 2021, Anomaly detection in log files using machine learning techniques.
- Meng, W., Liu, Y., Zhu, Y., Zhang, S., Pei, D., Liu, Y., Chen, Y., Zhang, R., Tao, S. ve Sun, P., 2019, Loganomaly: Unsupervised detection of sequential and quantitative anomalies in unstructured logs, *IJCAI*, 4739-4745.
- Myles, A. J., Feudale, R. N., Liu, Y., Woody, N. A., & Brown, S. D. (2004). An introduction to decision tree modeling. *Journal of Chemometrics: A Journal of the Chemometrics Society*, 18(6), 275-285.
- Nguyen, H. T., Torrano-Gimenez, C., Alvarez, G., Petrović, S. ve Franke, K., 2011, Application of the generic feature selection measure in detection of web attacks, *Computational Intelligence in Security for Information Systems: 4th International Conference, CISIS 2011, Held at IWANN 2011, Torremolinos-Málaga, Spain, June 8-10, 2011. Proceedings*, 25-32.
- Pal, R. ve Chowdary, N., 2018, Statistical profiling of n-grams for payload based anomaly detection for HTTP web traffic, *2018 IEEE International Conference on Advanced Networks and Telecommunications Systems (ANTS)*, 1-6.
- Shaheed, A., & Kurdy, M. H. D. (2022). Web Application Firewall Using Machine Learning and Features Engineering. *Security and Communication Networks*, 2022.)
- Smaha, S. E., 1988, Haystack: An intrusion detection system, *Fourth Aerospace Computer Security Applications Conference*, 37.

Statista (2021), Annual number of data breaches and exposed records in the United States from 2005 to 2020, Retrieved from <https://www.statista.com/statistics/273550/data-breaches-recorded-in-the-united-states-by-number-of-breaches-and-records-exposed/>.

TOPRAK, S., & YAVUZ, A. G. (2022). Web application firewall based on anomaly detection using deep learning. *Acta Infologica*, 6(2), 219-244.

Torrano-Gimenez, C., Nguyen, H. T., Alvarez, G., Petrović, S. ve Franke, K., 2011, Applying feature selection to payload-based web application firewalls, 2011 Third International Workshop on Security and Communication Networks (IWSCN), 75-81.

Vartouni, A. M., Kashi, S. S. ve Teshnehlab, M., 2018, An anomaly detection method to detect web attacks using stacked auto-encoder, 2018 6th Iranian Joint Congress on Fuzzy and Intelligent Systems (CFIS), 131-134.

Ye, N., Emran, S. M., Chen, Q. ve Vilbert, S. J. I. T. o. c., 2002, Multivariate statistical analysis of audit trails for host-based intrusion detection, 51 (7), 810-820.

Zhang, X., Xu, Y., Lin, Q., Qiao, B., Zhang, H., Dang, Y., Xie, C., Yang, X., Cheng, Q. ve Li, Z., 2019, Robust log-based anomaly detection on unstable log data, Proceedings of the 2019 27th ACM Joint Meeting on European Software Engineering Conference and Symposium on the Foundations of Software Engineering, 807-817.

Zhang, H. (2004). The optimality of naive bayes, 2004. American Association for Artificial Intelligence (www.aaai.org).

Breiman, L. (2001). Random forests. *Machine learning*, 45, 5-32.

Differential Evolutionary Optimal Architecture Deep Neural Network Model for Heating Load Estimation of Buildings

Nida Tekedereli Özçelik^{*1} , Baris Baykant Alagoz² 

¹Computer Engineering Department, Inonu University, Malatya, Türkiye

(ntekedereli@gmail.com, baykant.alagoz@inonu.edu.tr)

Abstract— Optimality of neural network architecture according to training datasets is a main concern for data-driven modeling of real-world systems. This study presents a differential evolution (DE) based data-driven optimal architecture artificial neural network (ANN) model design scheme for the heating load prediction from building features. For neuroevolution of ANN architecture, we implemented a multi-objective function that considers the sum of mean absolute error (MAE) performances for training and test datasets. An advantage of using absolute error instead of sum square error, absolute-value norm is closely related with the L1-regularization that is preferred to suppress weight of less significant parameters in the model. Thus, we investigated effects of the MAE based objectives on the model complexity and optimal neural architecture search for a given dataset. Effectiveness of the proposed optimal ANN models is demonstrated in the heating load estimation problem of residential buildings and compared with results of other methods.

Keywords : *neuroevolution, optimal architecture, artificial neural network, energy, building.*

1.Introduction

Data-driven modeling has been widely preferred to identify a mathematical model for computational representation of real-world system responses, and it enables prediction or simulation of the modeled system behavior. Over two decades, artificial neural networks have been a widely used for the black-box modeling in many engineering applications such as hydroinformatics (Solomatine et al., 2009), energy (Ertugrul, 2016; Chou et al., 2014), process control (Hou et al., 2013; Alimohammadi et al., 2020, Medi et al., 2021) etc. Today's intelligent system paradigms, which aim to achieve optimal and adaptive system responses, are shaping around the data-driven modeling of real-world systems.

In data-driven modeling, determination of ANN model architecture is an important problem, and if the model architecture, namely model complexity, is not optimal, this drawback may easily lead to underfitting or overfitting problems in the training stage (Gavrilov et al., 2018). Therefore, the optimal architecture ANN design becomes an essential topic to maintain practical performance of ANN models in real-world applications. Moreover, autonomous and intelligent systems require self-design of neural architectures according to the nature of data. At this point, neuroevolution can be an effective solution to deal with overfitting and underfitting problems in artificial neural network training tasks, and evolutionary optimization methods are implemented to address optimal architecture design problems (Benardos et al. 2007).

The optimal training performance depends on a suitable neural architecture selection for a given training dataset (Baker et al., 2016; Kandasamy et al., 2018). Selection of a suitable neural network architecture involves determination of the number of hidden layers and the neuron counts at each hidden layer. When this process is carried out manually, it is very time consuming and requires expertise on dataset and neural networks. To deal with this issue, Benardos et al. addressed optimization of ANN architecture according to mean absolute relative error of training and test dataset and considered neural complexity objectives (Benardos et al. 2007). Then, Carvalho et al. and Anochi et al. have developed

multi-objective functions for the self-tuning ANN models, and this function focused on minimizing sum of square errors ($sse = \sum (y_d - y_{nn})^2$), total neuron numbers in the neural network and the training epoch counts. These selected objectives consider reducing neural network complexity and improving computational efficiency in the training process. Simsek et al. implemented and modified their objectives to utilize popular metaheuristics in optimal architecture ANN design problems for several real-world applications (Simsek et al., et. 2022; Simsek et al., et. 2023).

In this study, we implemented Simsek et al.'s optimal architecture ANN design procedure for data-driven optimal architecture model design for the residential heating load prediction problem. To address overfitting and underfitting problems, we suggested a multi-objective function with mean absolute error (MAE, $mae = \sum |y_d - y_{nn}|$) performance of training and test datasets. Thus, properties associated with the absolute-value norm are utilized to simplify the objective function in neural architecture optimization. Modeling performance of the optimal ANN model is investigated in the heating load prediction of building according to 8 building features, and results are compared with other methods that were used to model generation from the energy efficiency dataset. Two main contributions of this study can be highlighted as follows:

- (i) Data-driven neuroevolution of deep neural networks is demonstrated by adapting a differential evolution algorithm (Storn et al., 1997, Qin et al. 2009) according to Simsek et al.'s neural architecture optimization procedure (Simsek et al., et. 2022; Simsek et al., et. 2023), and effects of MAE based objective function on neural architecture optimization is studied. Benefits of absolute-value norm based objectives for ANN architecture optimization is discussed.
- (ii) The differential evolutionary optimal architecture ANN modeling is implemented for optimal data-driven modeling of the energy efficiency dataset, and the heating load estimation performance of the optimal architecture deep ANN model is reported and compared with performance of other models in literature.

2. Methodology

2.1. Preliminary Works and Problem Statement:

This section briefly explains the metaheuristic based ANN architecture optimization scheme that was implemented by Simsek et al. Figure 1 shows details and adaptation of this scheme for the solution of heating load prediction problems. Modified blocks are indicated by blue asterisk (*).

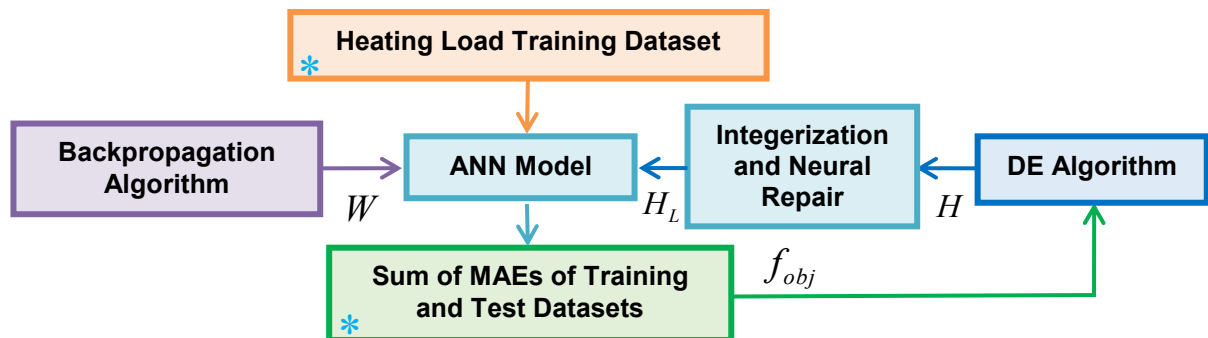


Figure 1. Some details of metaheuristic based ANN architecture optimization scheme (Simsek et al., et. 2022; Simsek et al., et. 2023) and modifications on the scheme for heating load prediction problem

This scheme performs the backpropagation algorithm for parameter based learning of ANN model and metaheuristic architecture optimization to manage structural learning of ANN from the dataset. Candidate solution vectors (H) of the DE algorithm is decoded into a valid architecture description (H_L), and the vector H_L is used to define architecture of ANN model. To obtain a valid architecture description, a number of processes is carried out on the H vector: Firstly, the H vector turns into a integer value vector by rounding the elements to the nearest integer number (Integerization process) and then zero value elements in the H vector are removed (Neural repair process) to avoid neural disconnection between the hidden layers (Simsek et al., et. 2022; Simsek et al., et. 2023). The model, which is described by means of the vector H_L , is trained and tested several times by using the backpropagation algorithm. To select the best model, a minimum value policy ($\arg \min_i \{f_{obj}(i)\}$) is preferred to deal with uncertainty of backpropagation training (Simsek et al., et. 2022). Accordingly, f_{obj} value of the best model is sent to the DE algorithm and this value is used for differential evolution of the architecture description vector H to reach better solutions.

To optimize the neural architecture, Simsek et al. implemented the multi-objective function ($fobj_1$), which has been progressively developed in works of Carvalho et al. and Anochi et al.

$$fobj_1 = \left[C_1 (\varepsilon_{neu})^2 \times C_2 (epochs) + 1 \right] \times \frac{(\rho_1 \times sse_{train} + \rho_2 \times sse_{gen})}{\rho_1 + \rho_2}, \quad (1)$$

where ρ_1 and ρ_2 are coefficients to adjust weights of the square error sum ($\sum (y_d - y_{nn})^2$) term for the training set (sse_{train}) and square error sum term for the test dataset (sse_{gen}) in the optimization process. C_1 and C_2 are coefficients to adjust weights of total neuron counts (ε_{neu}) and total training epoch counts ($epoch$).

In the current study, we investigate the effects of the objective function, which is the weighted sum of MAE of training and test sets, on the neural architecture design. Therefore, the following simplified objective function ($fobj_2$) is used for the architecture optimization of ANNs:

$$fobj_2 = \gamma_1 \times mae_{train} + \gamma_2 \times mae_{gen}, \quad (2)$$

where γ_1 and γ_2 are coefficients that are used to adjust weights of the mean absolute error term for training dataset (mae_{train}) and the mean absolute error term for test dataset (mae_{gen}). Through the optimization process, an advantage of using absolute error in the objective function can be computational properties associated with the absolute-value norm. The absolute-value norm has been used in L1-regularization in order to suppress less important parameters in optimization processes, and this property can be helpful for self-tuning the ANN model complexity during the optimization process. For this reason, we simplified objective function structure and removed parameters that have been used to govern neuron counts (ε_{neu}) and epoch number in equation (1). In a previous work, Benardos et al. used the mean absolute relative error objectives (Benardos et al. 2007), however they didn't investigate the mean absolute error to manage the neural architecture complexity for heating load prediction problem.

3. Numerical Experiments

3.1. Energy Efficiency Dataset and Configuration of Optimal Architecture ANN Designs

Energy efficiency data set have been collected by Tsana et al. 2012 for heating or cooling load estimation from 8 different features of residential buildings (Tsanas et al., 2012). These features are the

relative compactness (x_1), the surface area (x_2), the wall area (x_3), the roof area (x_4), the overall height of building (x_5), the orientation of building (x_6), the glazing area (x_7) and glazing area distribution (x_8). Using these building features, the goal of this regression problem is to obtain a heating load prediction model and a cooling load prediction model. By using these models, one can design energy efficient buildings that can optimize energy consumption for heating and cooling of the designed houses. In the literature, several machine learning methods have been implemented to solve this modeling problem (Al-Rakhami et al., 2019). In the current study, to simplify our analysis and limit the length of this paper, we worked on the heating load data to predict the heating load of buildings based on those eight features. Depending on these features, the ANN model output (y_{nn}) can be expressed in the form of

$$y_{nn} = F(x_1, x_2, x_3, x_4, x_5, x_6, x_7, x_8) \quad (3)$$

We used 8 features from 614 buildings in the energy efficiency dataset. Randomly selected 430 data instances were used for the training set and 92 data instances were used as the validation set for the backpropagation algorithm. The 92 data instances were allocated for test data at each run and they were also used for test data in objective functions. Parameters of the objective function (Eq. (1)) were configured to $\rho_1 = 1$, $\rho_2 = 0.1$, $C_1 = 1$ and $C_2 = 0.1$. (The model output for the objective function with equation (1) is denoted by y_{nn1}) Parameters of the second objective function (equation (2)) were configured to $\gamma_1 = 1$ and $\gamma_2 = 1$. (The model output for the objective function with equation (2) is denoted by y_{nn2}) Maximum number of hidden layers is set to 5 layers, and maximum neuron numbers in the hidden layers limited to 30 neurons. That is why, the parameter number of the DE vector is set to 5 and the range of each parameter is set to 30 for the DE algorithm (Storn et al., 1997, Qin et al. 2009).

It is useful to see an example decoding of a DE solution vector to neural network architecture (Simsek et al., et. 2022; Simsek et al., et. 2023). For instance, a given valid architecture description $H_L = [24 \ 12 \ 5 \ 16 \ 7]$ implies that the neural architecture has 5 hidden layers: The first hidden layer has 24 neurons, the second hidden layer has 12 neurons, the third hidden layer has 5 neurons, the fourth hidden layer has 16 neurons and the fifth one has 7 neurons. The output layer is fixed in the regression model and it has one neuron with a linear activation function to fit the output data that can take values from a large interval of real numbers. The input layer has 8 inputs corresponding to 8 building features from the dataset. Matlab *fitnet()* function was used to configure the ANN model according to the valid candidate architecture description H_L that comes from the DE algorithm. For the training, the Levenberg-Marquardt backpropagation algorithm was implemented in the objective function of the differential evolution algorithm. The repeated training numbers are 3 to calculate mean absolute errors according to three independent ANN training tasks, and we employ the minimum value policy, which selects the best performance trained ANN model that presents a minimum value of the objective function.

3.2. Numerical Results and Discussions

Table 1 and 2 show the root mean square (RMSE) performances for training and test datasets, respectively. Table 3 and 4 show the mean absolute error (MAE) performances for training and test datasets. The data-driven neuroevolution algorithm was performed three times for each objective functions $fobj_1$ (Eq.1) and $fobj_2$ (Eq.2). One can see in these tables that the optimal architecture ANN model by using equation (1) can yield lower RMSE and MAE for the training dataset; however the RMSE and the MAE for the test dataset can be higher than those of optimal architecture ANN model by using equation (2). This is a possible indication of an overfitting problem and less generalization. On the other hand, we observed in a few tests that ANN models according to equation (1) may yield low performance models. These results indicate to us that the performance of modeling with equation (2)

can be more consistent (robust) than the performance of the modeling with equation (1). Lower standard deviation of performance data for equation (2) also supports this observation. Figure 2 also reveals the performance variability of the model in the case of using equation (1). A reason for this performance variability in modeling performance of equation (1) may be that the complex structure of equation (1) can further complicate the parameter search space for metaheuristic optimization algorithms, and this factor may reduce performance consistency in the repeated training and testing case. Also, more hyperparameters (user-defined constants) in equation (1) may require finer tuning to maintain modeling consistency. As a consequence, we concluded that simplification of the objective function and use of absolute value metrics can contribute to the modeling consistency in neural architecture optimization.

Table 1. RMSE performances for training set

	Average RMSE	Minimum RMSE	Maximum RMSE	Standard Deviation
Equation (1)	1.023	0.142	2.776	1.518
Equation (2)	0.171	0.146	0.182	0.021

Table 2. RMSE performances for test set

	Average RMSE	Minimum RMSE	Maximum RMSE	Standard Deviation
Equation (1)	1.373	0.457	3.138	1.528
Equation (2)	0.409	0.357	0.511	0.088

Table 3. MAE performances for the training set

	Average MAE	Minimum MAE	Maximum MAE	Standard Deviation
Equation (1)	0.703	0.081	1.893	1.030
Equation (2)	0.121	0.108	0.127	0.011

Table 4. MAE performances for the test set

	Average MAE	Minimum MAE	Maximum MAE	Standard Deviation
Equation (1)	0.957	0.310	2.192	1.070
Equation (2)	0.277	0.271	0.288	0.009

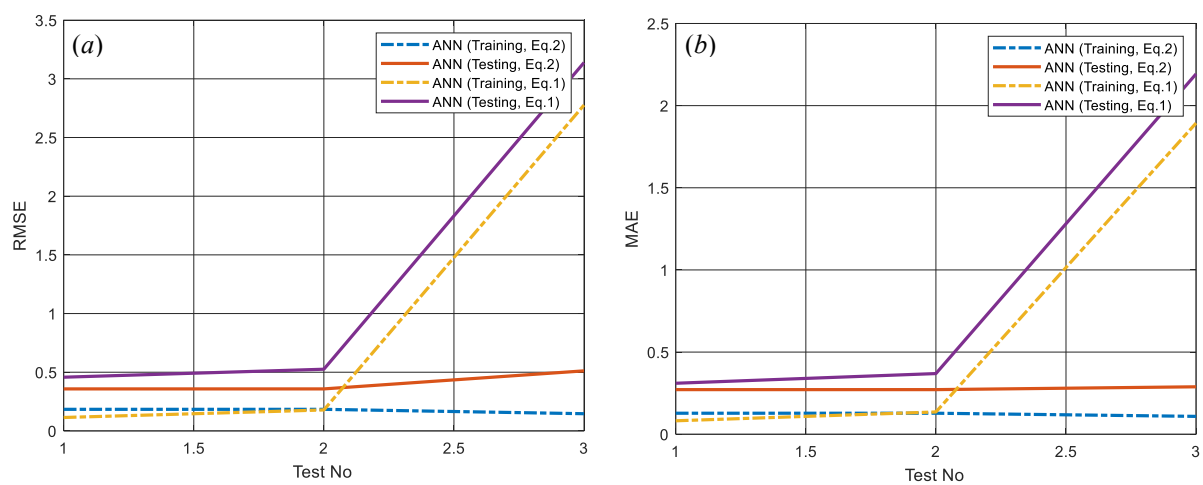


Figure 2. RMSE and MAE performances for three repeated training and test phases

We selected the best models that presented the lowest RMSE and MAE for training and test datasets according to performance data in Figure 2. Figure 3 shows convergence of objective functions and fitting of model estimates to the datasets. Decrease of the objective function validates evolution of deep neural network architecture during the differential evolution optimization in Figure 3(a) and (b). Fitting of predictions of the ANN models to the real heating load data is significant for the training dataset (Figure 3(c) and (d)) and the test dataset (Figure 3 (e) and (f)).

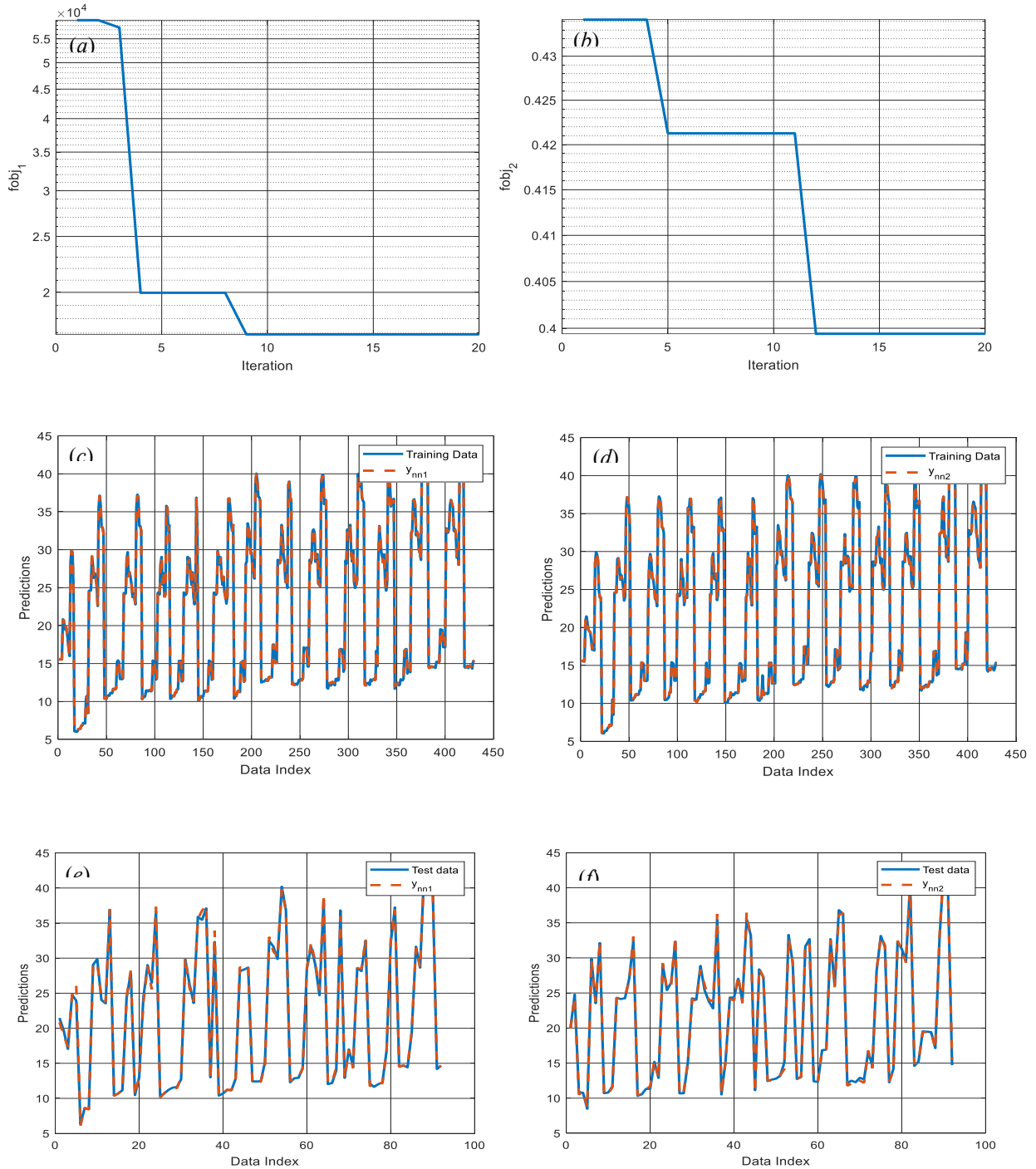


Figure 3. (a) and (b) show convergence of objective functions during differential evolution of neural architecture; (c) and (d) show the fitting of head load estimations to training data; (e) and (f) show the fitting of head load estimations to test data. (Left hand side figures is for equation (1) and right hand side figures is for equation (2))

The best neural architecture descriptions are given at Table 5. According to these architecture descriptions, one can observe that the best models did not use all of the allowed 5 hidden layers. They used two layers according to the objective function equation (2) and three layers according to the objective function equation (1). Even though there is not any objective to reduce neuron counts in equation (2), it can yield a lower complexity model with two hidden layers. This result can be an indication for the parameter suppression effect of the MAE and supports the expectation of inherent model complexity reduction effect via MAE based objectives. Also, these findings indicate that ANN based modeling of this dataset may not need a deeper neural network, and results of optimization process suggested two or three layers for improved modeling performance in this study. This result comes out of a discussion whether a deep neural network is always necessary for ANN modeling of every datasets. According to our findings, the optimal architecture, which can improve the modeling performance, can be accessible via the neural architecture optimization, and determination of optimal neural architecture can be performed according to model complexity requirements of datasets. This property is referred to as data-driven optimization of neural architecture, namely the data-driven neuroevolution. Table 6 compares performances of the best optimal architecture ANN models with those of other published works. The best ANN model according to the equation (2) can produce better results than the majority of other state-of-art methods in the table.

Table 5. Valid architecture descriptions (H_L) of the best ANN models from repeated neuroevolution

Objective Functions	Hidden Layers				
	1 th layer	2 th layer	3 th layer	4 th layer	5 th layer
Equation 1	23	16	8	-	-
Equation 2	9	21	-	-	-

Table 6. Performances of the best optimal architecture ANN models and performance data from other published works (Al-Rakhami et al., 2019)

Related Works	Methods	MAE	RMSE	MAPE(%)	R ² -score
Tsanas et al.(2012)	Random Forest	0.51	-	2.2	-
Cheng et al. (2014)	Ensemble Model	0.34	0.46	-	0.998
Chou, et al. (2014)	Ensemble Model	0.23	0.35	1.1	0.999
Castelli et al. (2015)	Genetic Program	0.38	-	0.43	-
Duarte et al. (2017)	Random Forest	0.315	0.22	1.4	0.998
Goliatt et al. (2018)	Gaussian Process	0.25	0.38	1.3	0.999
Al-Rakhami et al.(2019)	Extreme Gradient Boosting	0.17	0.26	0.91	0.999
According to equation (1)	Optimal architecture ANN	0.31	0.45	1.4	0.997
According to equation (2)	Optimal architecture ANN	0.27	0.35	1.3	0.998

3. Conclusions

This study illustrated an application of differential evolutionary optimal architecture deep neural network design approach for prediction of heating load in buildings according to eight building features. We modified the multi-objective function by using MAE error instead of the sum of square errors (SSE). This modification aims to investigate the effects of two modifications on the neural architecture optimization problem: (i) simplification of multiobjective function in order to facilitate search of metaheuristic methods, (ii) the parameter suppressing effect of absolute value norm to reduce complexity of the resulting ANN models.

Our numerical experiments on the energy efficiency dataset reveal that deep neural networks may not be a necessity for high performance solutions of regression problems. An optimal neural architecture, which deals with overfitting and underfitting problems, can be found by using data-driven neuroevolution methods. Optimization of neural architecture to improve learning performance of ANNs is known as structural learning (Cortes et al.,2017). To this end, this study investigated a differential

evolutionary neuroevolution mechanism, and performance of the suggested neuroevolution method was demonstrated in comparison with the performance of other methods that have been published for the solution of heating load prediction problem. Results indicate that differential evolutionary optimal classical neural network models can present an acceptable prediction performance compared to performances of other state-of-art methods. The data-driven neuroevolution and benefiting from structural learning play a role in these results.

References

- Alimohammadi, H., Alagoz, B. B., Tepljakov, A., Vassiljeva, K., & Petlenkov, E. (2020). A NARX model reference adaptive control scheme: improved disturbance rejection fractional-order PID control of an experimental magnetic levitation system. *Algorithms*, 13(8):201.
- Anochi JA, Velh, HFC, Furtado HC and Luz EF (2015) Self-configuring two types of neural networks by mpca. *Journal of Mechanics Engineering and Automation* 5: 112-120.
- Al-Rakhami, M., Gumaei, A., Alsanad, A., Alamri, A., & Hassan, M. M. (2019). An ensemble learning approach for accurate energy load prediction in residential buildings. *IEEE Access*, 7: 48328-48338.
- Baker B, Gupta O, Naik, N and Raskar R (2016) Designing neural network architectures using reinforcement learning. *arXiv preprint arXiv:1611.02167*.
- Benardos, P. G., & Vosniakos, G. C. (2007). Optimizing feedforward artificial neural network architecture. *Engineering applications of artificial intelligence*, 20(3), 365-382
- Carvalho AR, Ramos FM and Chaves AA (2011) Metaheuristics for the feedforward artificial neural network (ANN) architecture optimization problem. *Neural Computing and Applications* 20(8): 1273–1284.
- Castelli, M., Trujillo, L., Vanneschi, L., & Popović, A. (2015). Prediction of energy performance of residential buildings: A genetic programming approach. *Energy and Buildings*, 102: 67-74.
- Cheng, M. Y., & Cao, M. T. (2014). Accurately predicting building energy performance using evolutionary multivariate adaptive regression splines. *Applied Soft Computing*, 22: 178-188.
- Chou, J. S., & Bui, D. K. (2014). Modeling heating and cooling loads by artificial intelligence for energy-efficient building design. *Energy and Buildings*, 82: 437-446.
- Cortes, C., Gonzalvo, X., Kuznetsov, V., Mohri, M., & Yang, S. (2017). Adanet: Adaptive structural learning of artificial neural networks. In *International conference on machine learning* (pp. 874-883). PMLR.
- Duarte, G. R., da Fonseca, L. G., Goliatt, P. V. Z. C., & de Castro Lemonge, A. C. (2017). Comparison of machine learning techniques for predicting energy loads in buildings. *Ambiente Construído*, 17(3): 103-115.
- Ertugrul, Ö. F. (2016). Forecasting electricity load by a novel recurrent extreme learning machines approach. *International Journal of Electrical Power & Energy Systems*, 78: 429-435.
- Gavrilov, A. D., Jordache, A., Vasdani, M., & Deng, J. (2018). Preventing model overfitting and underfitting in convolutional neural networks. *International Journal of Software Science and Computational Intelligence (IJSSCI)*, 10(4): 19-28.
- Goliatt, L., Capriles, P. V. Z., & Duarte, G. R. (2018,). Modeling heating and cooling loads in buildings using Gaussian processes. In *2018 IEEE Congress on Evolutionary Computation (CEC)* (pp. 1-6). IEEE.
- Hou, Z. S., & Wang, Z. (2013). From model-based control to data-driven control: Survey, classification and perspective. *Information Sciences*, 235: 3-35.
- Kandasamy K, Neiswanger W, Schneider J, Poczos B and Xing E (2018) Neural architecture search with bayesian optimisation and optimal transport. *arXiv preprint arXiv:1802.07191*

- Medi, B., & Asadbeigi, A. (2021). Application of a GA-Optimized NNARX controller to nonlinear chemical and biochemical processes. *Heliyon*, 7(8): e07846.
- Qin, A. K., Huang, V. L., & Suganthan, P. N. (2009). Differential evolution algorithm with strategy adaptation for global numerical optimization. *IEEE Transactions on Evolutionary Computation*, 13(2): 398–417.
- Simsek, O. I., & Alagoz, B. B. (2022). Model Based Demand Order Estimation by Using Optimal Architecture Artificial Neural Network with Metaheuristic Optimizations. *Iğdır Üniversitesi Fen Bilimleri Enstitüsü Dergisi*, 12(3):1277 -1291.
- Simsek, O. I., & Alagoz, B. B. (2023). Optimal architecture artificial neural network model design with exploitative alpha gray wolf optimization for soft calibration of CO concentration measurements in electronic nose applications. *Transactions of the Institute of Measurement and Control*, 45(4): 686-699.
- Solomatine, D., See, L. M., & Abraham, R. J. (2009). Data-driven modelling: concepts, approaches and experiences. *Practical hydroinformatics*, 17-30.
- Storn, R., & Price, K. (1997). Differential evolution—a simple and efficient heuristic for global optimization over continuous spaces. *Journal of global optimization*, 11: 341-359.
- Tsanas, A., & Xifara, A. (2012). Accurate quantitative estimation of energy performance of residential buildings using statistical machine learning tools. *Energy and buildings*, 49: 560-567.