



# JISTA

*Journal of Intelligent Systems: Theory  
and Applications*

MARCH 2021

ISSN: 2651-3927



**VOL 4 NO 1**

ARTIFICIAL INTELLIGENT > MACHINE LEARNING > MULTI-AGENTS

<https://dergipark.org.tr/tr/pub/jista>



# Journal of Intelligent Systems: Theory and Applications

## Volume: 4 Issue: 1

### Editorial Boards

#### Honorary Editors

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

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

#### Editor-In-Chief

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

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

#### Editors

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

Ayten Yılmaz Yalçiner, ayteny@sakarya.edu.tr, Turkey

Alper Kiraz, kiraz@sakarya.edu.tr, Sakarya University, Turkey

M. Fatih Adak, fatihadak@sakarya.edu.tr, Sakarya University, Turkey

Fuat Şimşir, fuatsimsir@karabuk.edu.tr, Karabük University, Turkey

Zafer Albayrak, zalbayrak@karabuk.edu.tr, Karabük University, Turkey

Caner Erden, cerden@subu.edu.tr, Sakarya University of Applied Sciences, Turkey

#### Language Editor

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

#### Editorial Advisory Board

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

## **Editorial Assistants**

---

Enes Furkan Erkan, eneserkan@sakarya.edu.tr, Sakarya University, Turkey

Elif Yıldırım, elifyildirim@sakarya.edu.tr, Sakarya University, Turkey



# Journal of Intelligent Systems: Theory and Applications

## Volume: 4 Issue: 1

### Contents

#### Research Articles

- |  |       |
|--|-------|
| <b>Performance Evaluation of WBANs MAC Protocols in Different dBm and OMNet++</b><br><i>Zafer ALBAYRAK, Hatem MUSA, Muhammet ÇAKMAK</i>  | 1-7   |
| <b>Tüketici Yorumları Üzerine Bir Metin Madenciliği ve Veri Boyutu İndirgeme Yaklaşımı</b><br><i>Ahmet YÜCEL</i>   | 8-16  |
| <b>Recognition of Sign Language Letters Using Image Processing and Deep Learning Methods</b><br><i>Ali ÖZTÜRK, Melih KARATEKİN, İsa Alperen SAYLAR, Nazım Bahadır BARDAKCI</i>   | 17-23 |
| <b>A Tabu Search and Hybrid Evolutionary Strategies Algorithms for the Integrated Process Planning and Scheduling with Due-date Agreement</b><br><i>Halil İbrahim DEMİR, Caner ERDEN, Abdullah KÖKÇAM, Alper GÖKSU</i> | 24-36 |
| <b>Classifier Selection in Resource Limited Hardware: Decision Analysis and Resolution Approach</b><br><i>Atilla ÖZGÜR</i>   | 37-42 |
| <b>Küme Birleşimli Sırt Çantası Probleminin Adaptif Yapay Arı Kolonisi Algoritması ile Çözümü</b><br><i>Rafet DURGUT, İlim YAVUZ, Mehmet AYDIN</i>   | 43-54 |
| <b>Makine Öğrenmesi Algoritmaları Kullanarak Erken Dönemde Diyabet Hastalığı Riskinin Araştırılması</b><br><i>Gürkan BİLGİN</i>  | 55-64 |



# Performance Evaluation of WBANs MAC Protocols in Different dBm and OMNet++

Zafer Albayrak<sup>1\*</sup> , Hatem Musa<sup>2</sup> , Muhammet Çakmak<sup>3</sup> 

<sup>1</sup>Karabuk University, Faculty of Engineering, Computer Engineering, Karabük/Turkey

<sup>2</sup>Tobruk University, Faculty of Engineering, Computer Engineering, Tobruk/Libya

<sup>3</sup>Karabuk University, TOBB Tech. Sciences Vocational School, Computer Technologies, Computer Programming, Karabük/Turkey  
zalbayrak@karabuk.edu.tr, hatem.musa@tu.edu.ly, muhammetcakmak@karabuk.edu.tr

## Abstract

In present days, wireless sensor networks (WSN) have involved considerable attention of both academy and industry because of the varied range of contexts in which they could be used. The has wireless body area network (WBAN) become the most important standard for WSN, and several software and hardware platforms are built on it. The implementation and performance analysis of this standard is essential to understand the important limits of it. The simulation is one of the greatest valuable tools for protocol evaluation and prototyping design. Furthermore, network simulators play an important part to test new algorithms and other protocols built on this specification. In this paper, the performance of the WBAN MAC standard protocols has been tested. The performance of the protocols regarding power consumption, delay and packets congestion are compared using OMNet++ simulator.

**Keywords:** wireless, sensor network, WBANs, algorithm, simulation.

## 1 INTRODUCTION

The wireless connection is used to connect different devices without any physical connection like cables. The wireless type of connection reduces the cost and difficulties of using the traditional wired network. In wireless networks, the connected devices use the radio frequencies to send data between source and destination. The physical layer in the wireless network devices is responsible for getting connected to each other [1].

Current developments in micro-electro-mechanical systems (MEMS) technology, wireless communications, and digital electronics have allowed the progress of low-cost, low-energy, multiuse sensors that are minor and connect free in small areas. These little sensors, that contain sensing, data processing, and communicating gears, force the knowledge of sensor networks grounded on cooperative energy of several nodes [2].

The Institute of Electrical and Electronic Engineers (IEEE) confirmed the structure of a working collection for IEEE 802.15.4 (IEEE 2003) to outline a foundation to Body Area Network [3]. The 802.15.4 defines both the

physical & media access control layer. The physical layer can work in different bandwidths, the first one is the frequency band 2.4 to 2.4835 GHz using 16 different channels, the second one is the frequency band from 902 to 928 MHz using ten different channels, and the third one is one channel in the frequency band 868.0 MHz to 868.6 MHz [4]. There are different features of the media access control layer managing. They are beacons, channel contact, managing of GTS, proof of the frames, and others. There are two methods of process of the media access control layer contingent on the topology that used and the need for certain bandwidth; they are beaconless approach and beacon approach. In the beaconless approach, the sink node is the only state waiting for information. The expedient that wants to send info, it will first check if the channel is empty. If it is empty, then it will send the info. If it is not empty, it will wait for an arbitrary time that defined in the ordinary. If the sink node has info that must be sent to an expedient, it will wait till the nodes demand for the information. After that, the sink node must send the acknowledgment to reaction of the demand. The sink node will transmit the info if they are pending, using the exact

\* Corresponding Author.  
E-mail:zalbayrak@karabuk.edu.tr

procedure of CSMA/CA. In the case of no info that waits, the sink node sends an empty info frame. The beaconless approach is naturally used in the nodes that sleep for a long time (99%). When an incident occurs, the nodes will wake-up and immediately will direct a frame of aware. In this kind of work, the sink node will not supply any synchronization for the nodes, no Guaranteed Time Slot (GTS) will be coming, and only arbitrary access is taken over for medium spreading because of no superframe and the space of synchronization. There are many kinds of sensors used in sensor networks: low sampling rate magnetic, thermal, seismic, visual, infrared, acoustic and radar, which can sense different environment circumstances such as humidity, temperature, lightning condition, vehicular movement, noise levels, pressure, soil makeup and the current features like speed, course, and size of an item.

Sensor nodes could be beneficial to place sensing, continuous sensing, control of actuators, and event recognition. The idea of wireless connection and micro-sensing of these nodes open the door for many different application zones [5]. In current times, there has been growing awareness from researchers, application developers and system designers, on a novel kind of network construction commonly known as wireless body area networks (WBANs) or body sensor networks (BSNs), made possible by new developments on frivolous, ultra-low-power, small-size, and smart monitoring wearable sensors. In WBANs, sensors constantly watch human's physical actions and activities, like motion pattern and fitness situation [6].

We studied five different MAC layer protocols which are used broadly in WBANs. Two different scenarios were created using Castalia simulator under the OMNET++ platform. We calculate, delay, power consumption, and packets congestion with the using of two different sensor dBm.

The rest of the paper is organized as follows: In Section 2, we will review the literature. Section 3 contains MAC protocols in WBANs. Section 4 contains OMNET++ simulation framework of mac protocols and finally, section 5 concludes the paper.

## 2 RELATED WORK

Recently, many articles have reviewed and detailed the aspects of wireless sensor and WBANs.

In [7], a new frame structure design is proposed for a new MAC in WBAN by prioritizing patient data traffic. Later, critical patient data were measured effectively and threshold-based slot allocation technique was developed. Also, packet size, mark spacing, super frame time efficient frame design are presented.

Hybrid MAC protocols for WBAN are compared in [8]. First, the design challenges of WBAN MAC are listed. Second, hybrid MAC protocols are compared to standard MAC protocols. Third, a comprehensive comparison of hybrid MAC protocols is presented in terms of QoS and WBAN specific values. Finally, important open research areas neglected in hybrid MAC design have been identified.

A new QoS-based cross-layer MAC protocol is proposed in [9], capable with the ISO / IEEE 11073 standard, using a socket allocation scheme, priority mechanism, multi-channel architecture, and cross-layer solution. The proposed MAC protocol has been modeled and simulated. The proposed method gives better results than other MAC protocols in terms of end-to-end delay, packet loss rate and throughput parameters.

In [10], a TDMA-based MAC protocol is proposed to dynamically adjust the transmission order and transmission time of the nodes according to the channel status of the WBAN and application context. The working slot allocation optimizes by minimizing the energy consumption of the nodes, subject to delivery probability and production constraints. It has also proposed a new synchronization scheme to reduce synchronization overhead.

In [11], reviewed protocols for sensor wireless networks and grants cataloguing for the several methods pursued. The Datacentric, the hierarchical and position-based are three important classifications that are inspected in this paper. In addition, the quality of network flow and service modeling is also discussed.

In [12], reviewed the synchronization of the time issue and the requirement for synchronization of the time in sensor networks, after that presents in detail, the synchronization of the time in the basic form approaches designed and proposed for sensor networks.

In [13], suggested the unwanted EA's performance when dealing with grouped routing problem in WSN by framing a new fitness role that incorporates two clustering parts, viz. cohesion and separation error.

In [14], offers sensor-MAC (SMAC), a new medium access control protocol planned for the networks of the wireless sensors. While dropping power, feeding is the primary goal in the plan; the protocol has the decent ability to change the size and capable of escaping from the collision. It achieves these by using a collection of scheduling and contention scheme.

In [15], the availability of WSN nodes are considered that can be addressed by indulging the distant testing and fixing the substructure for separate sensor nodes using COTs components, they built and evaluated the system level examination interface for distant testing repair and software update. This also covers contents regarding the plan methods which were carried to explore the difficulty using the projected infrastructure. The wireless broadcast can be used efficiently in various testing with optimum cost.

In [16], the modified superframe structure of IEEE 802.15.4 based MAC protocol is proposed which addresses the problems and improves the energy consumption efficiency. Moreover, priority guaranteed CSMA/CA mechanism is used where different priorities are assigned to body nodes by adjusting the data type and size.

In [17], presented an energy-efficient cooperative MAC (EECO-MAC) protocol using power control in mobile ad hoc networks. Cooperative communications improve network performance by taking full advantage of the broadcast nature of wireless channels.

### 3 WBANs MAC PROTOCOLS

At the MAC layer, there is an interchange among latency, reliability, and energy feeding that must be fixed. The QoS needs, i.e., latency and reliability, create from applications, and power feeding mirrors the overall protocol complexity and appropriate duty cycle [18]. Resource efficiency is a very important factor when developing a MAC protocol for WBAN. Comparing to wireless networks for more wide areas, WBANs experience much fewer power consuming which explains into more long times by getting an actual little duty cycle and a basic protocol job. Regularly, the body sensor has a partial battery volume, particularly for these sensors that are located in the body. For raise, the lifetime of those sensors, power effective MAC protocols will be a significant part. In contrast, some WBANs grounded applications require a very dependable connection, little delays, and little energy feeding [19].

To report the serious problem of spreading sensor time, many low energy MAC protocols have been projected for general WSNs [20]. In those protocols, the radio is switched on and off occasionally to maintain power [21].

S-MAC usages three new methods to decrease power feeding and provision self-formation. To decrease power feeding in hearing to a silent network, nodes occasionally snooze. Adjacent nodes procedure practical groups to auto-match on snooze timetables. Enthused by PAMAS, SMAC likewise puts the radio to snooze through communications of further nodes. Different from PAMAS, it solitary utilizes in-network signing. Lastly, SMAC puts on note transitory to decrease argument dormancy for sensor-network applications which need to keep and advance dispensation as information transfer over the network [22].

TMAC protocol is a medium access control protocol planned particularly for wireless sensor networks. TMAC allows wireless sensor node switch on its wireless at harmonized periods and switches it off later of a firm time-out once no message happens through some period. Messages are spread in bursts. This arrangement lets active alteration of the wireless-on period to altering message rates. TMAC protocol keeps additional power comparing to its predecessor SMAC in a network which message rates change. SMAC protocol allows the node to switch the wireless on for a static period. S-MAC needs change to the message rate, while T-MAC does not [23].

BMAC is a carrier sense media access (CSMA) protocol for wireless sensor networks. Driven by ecological monitoring applications, BMAC structures extreme low energy work, actual impact escaping, minor code scope, and expectable implementation. To reach little energy work, BMAC hires an adaptive little energy wireless selection arrangement to decrease working sequence, minimize idle hearing, and remove the overhead of harmonization. BMAC lets facilities to rearrange the MAC protocol for the best act, whether it be for productivity, dormancy, or energy preservation [24].

BanMAC is IEEE 802.15.6 for wireless body area transportations. That standard aims to stipulate numerous physical layers (PHY) and medium access control (MAC)

layer protocols for the diversity of requests with numerous QoS needs [25].

ZigBeeMAC is IEEE 802.15.4 standard stipulates that MAC layer is largely accountable for retrieving of the physical layer wireless channel, that is to reach networks active admission grounded on the physical layer interface purposes [26]. There are mostly two types of information sending style in ZigBee networks: with-beacon connection and without-beacon connection. In with-beacon networks, the network director occasionally transmits beacon frames, gear in PAN network is harmonized per the beacon frames from the director. For without-beacon networks, the network director arbitrarily broadcast beacon frames from period to period. When the node is around to transmit info, initially, it must pause for an arbitrary distance of time, and after that start to sense the network situation, if free, the node begins to transmit info; if not free, the node must pause for additional time, and re-sensing network till the network is free to transmit info. To shorten the understanding of the protocol, the plan utilizes without-beacon info sending model. Figure 1 shows the wireless body area networks' MAC protocols.

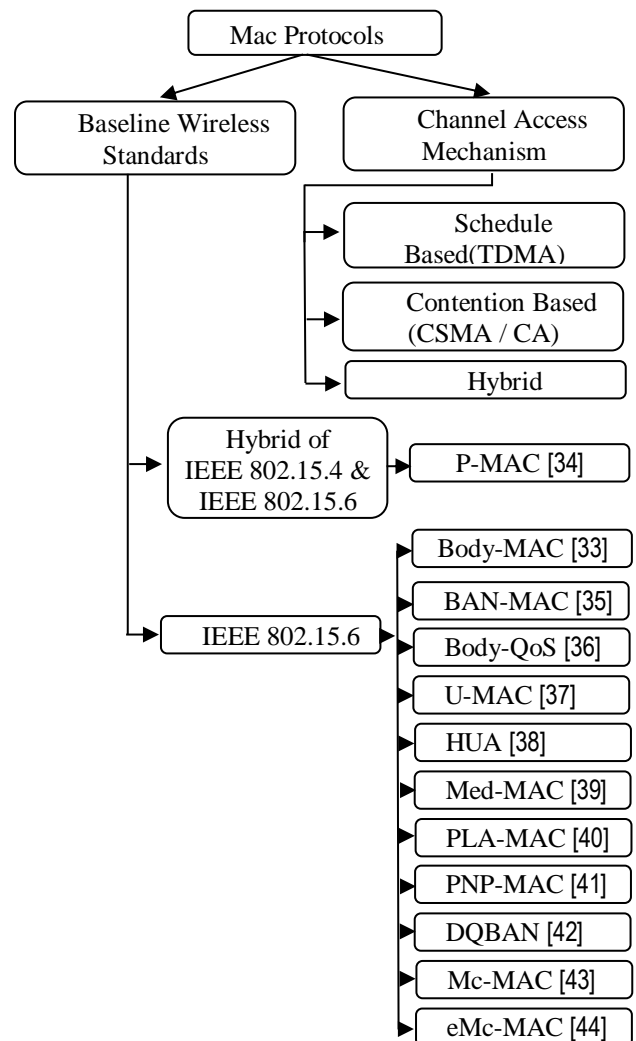


Figure 1. WBAN MAC Protocols.

## 4 SIMULATION FRAMEWORK

Network simulators attempt to model the actual networks [27]. The main knowledge is that if a system could be modelled, structures of the model could be altered, and the conforming outcomes could be examined. As the procedure of model adjustment is inexpensive comparing the whole actual operation, an extensive variety of scenarios could be examined at small charge [28] [29].

Presently there are a lot of network simulators which have various structures in dissimilar features [30]. A small list of the present network simulators contains OPNET, NS-2, NS-3, REAL, OMNet++, J-Sim, SSFNet, and QualNet. Though, in this chapter, we do not aim to shelter all the presented network simulators. We only choose some characteristic ones and do some study and compare some from the others a little to grow a good opinion of the key structures of a specific network simulator. OMNet++ is widely used WBANs open-source network simulator that has an influential modular core design and graphical interface [31] [32].

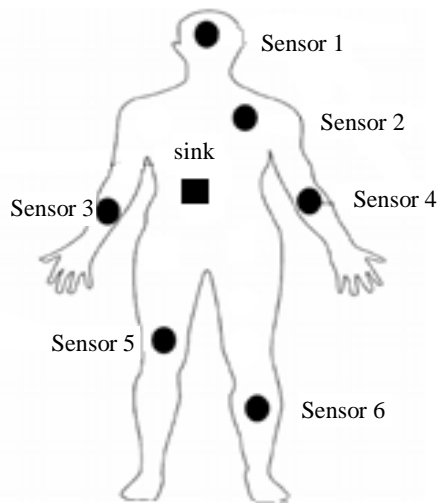


Figure 2. Distribution of the sensors.

We take four different considerations to compare among; delay, the power consumed, and collision. To evaluate the performance of the wireless sensor network; we created a simulation scenario using Castalia simulator based on OMNet++ platform. We created two different scenarios in Castalia simulator, the first scenario with six nodes and one sink node and the second scenario with 24 nodes and one sink node. The simulation parameters that we used are shown in table 1. In 6 nodes scenario, we assumed that all the 6 nodes are attached to one person only plus the sink node. as shown in Figure 22. For the 24 nodes scenario, we assumed that there are 4 persons in one room and every person have 6 nodes plus sink node attached for each one. We applied five different MAC layer protocols; TMAC, SMAC, BMAC, BANMAC (802.15.6), and ZigBeeMAC (802.15.4). In both scenarios, the application in each node generate 10 packets per second, and each packet size is 105 Byte as listed in table 1. We calculated the average delay from the delayed histogram by taking the average time for each interval in the histogram and multiply it by the number of packets received during this

interval, then we take the summation of them and divided it by the total number of the received packets.

Table 1. Simulation parameters

Parameters	Scenario 1	Scenario 2
Topology	Star	
Number of Nodes	6 + 1	24 + 1
Field Area	2 x 2 meter	6 x 4 meter
Mobility	Static Nodes	
Simulation Time	60 s	
Startup Delay Time	1 s	
Application Packets Rate	10 packets per second	
Application Packets Size	105 Byte	
Node TX Power	-10 dBm, -20 dBm	
MAC Protocols	TMAC, SMAC, BMAC, BANMAC, and ZigBeeMAC	
Max Packet Size for MAC	No Limit	
Buffer Size for MAC Packet	32 Packets	
Overhead for MAC	11 Byte	

We calculated the power consumption for the nodes by assuming that each node will consume 3.0 mW per second during the transmission and receiving with -10dBm sensor power and 2.9 mW per second during the transmission and receiving with -20dBm sensor power. Then calculate the time the consumed to transmit and receive all the packets in each node and multiply it by the power rate for each sensor power that assumed above.

We calculated the packets congestion by computing the number of packets that failed to be reached the sink node from each sensor due to the interference.

The average power consumed for six and twenty-four nodes. -10 dBm and -20 dBm power is shown in figure 3 and figure 4. In the six nodes, the SMAC protocol gives the least average power consumption level in both -20 dBm and -10 dBm power. While in the 24 nodes, the BMAC protocol gives the least average power consumption level in both -20 dBm and -10 dBm power. This is because BMAC does not have the RTS-CTS mechanism or synchronization requirements of other MAC protocols like SMAC and TMAC, the implementation is both simpler and smaller.

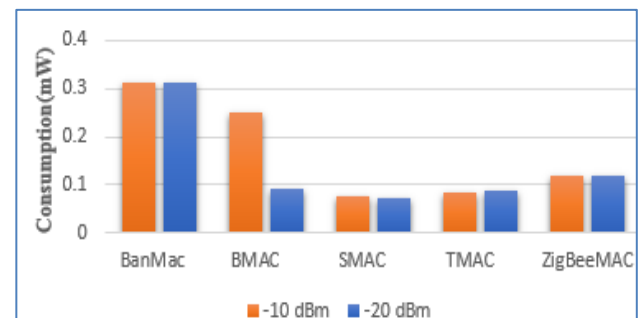
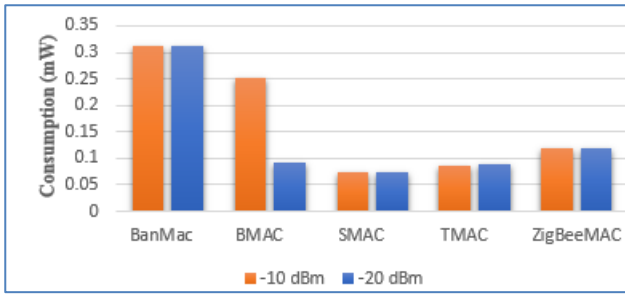


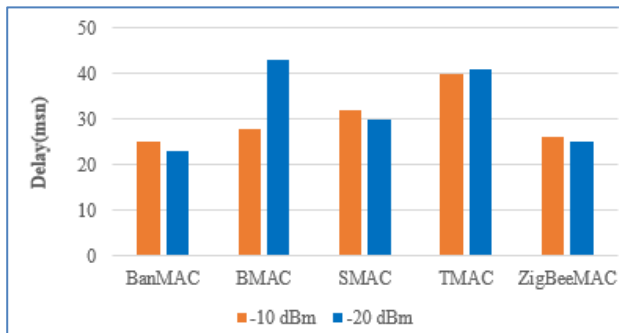
Figure 3. Average power consumption for six nodes.



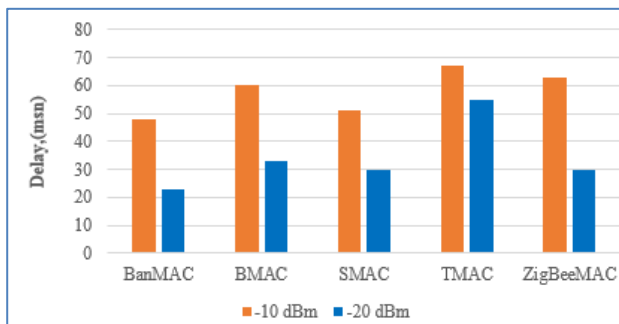


**Figure 4.** Average power consumption for twenty-four nodes.

The average delay for six nodes, -10 dBm and -20 dBm power is shown in figure 5 and The average delay for twenty-four nodes, -10 dBm and -20 dBm power is shown in figure 6. In the six nodes, the BanMAC and ZigBeeMAC protocols give the least delay in application level in both -20 dBm and -10 dBm power. While in the 24 nodes, the BanMAC protocol gives the least delay in application level in both -20 dBm and -10 dBm power. This is because BanMAC standard in a beacon mode with superframe boundaries, a hub divides the time into multiple superframes.

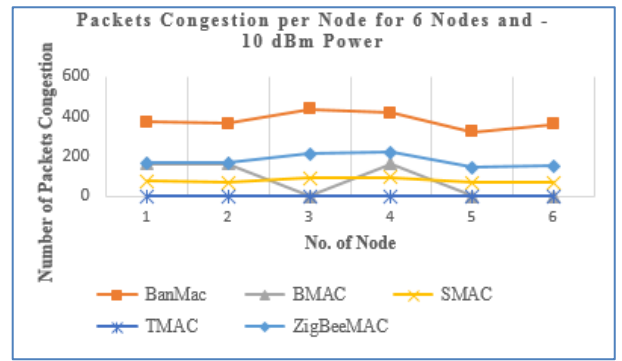


**Figure 5.** Average delay for six nodes

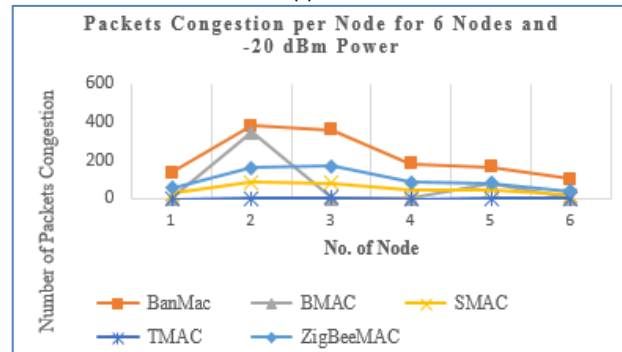


**Figure 6.** Average delay for twenty-four nodes

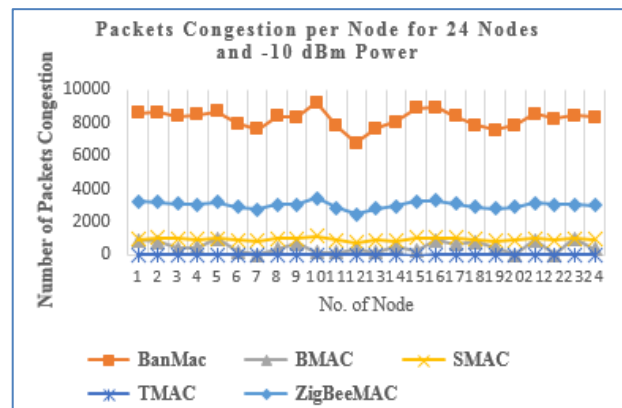
The congestion of the packets for six and twenty-four nodes, -10 dBm and -20 dBm power is shown in figure 7. In the six nodes, the TMAC protocol gives the least average packets congestion level in both -20 dBm and -10 dBm power. While in the 24 nodes, the TMAC and BMAC protocols give the least average packets congestion level in both -20 dBm and -10 dBm power. This is because T-MAC allows wireless sensor node switch on its wireless at harmonized periods and switch it off later of a firm time-out once no message happens through some period.



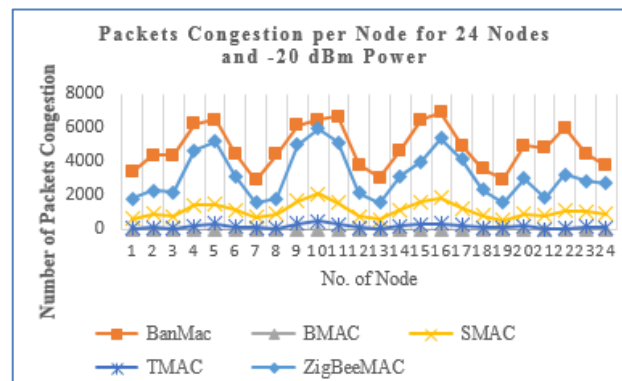
(a)



(b)



(c)



(d)

**Figure 7.** Congestion of the packets

To be clearer to understand the performance differences between the MAC protocols, we make a comparison as shown in table 2.

**Table 2.** Comparison of MAC Protocols

Power	Average Delay				Power Consumption				Packets Congestion			
	-10 dBm		-20 dBm		-10 dBm		-20 dBm		-10 dBm		-20 dBm	
Number of Nodes	6	24	6	24	6	24	6	24	6	24	6	24
<b>BanMAC</b>	Very Good	Good	Very Good	Very Good	Bad	Bad	Bad	Bad	Bad	Bad	Bad	Mid
<b>BMAC</b>	Mid	Mid	Mid	Good	Bad	Very Good	Good	Very Good	Good	Good	Mid	Very Good
<b>SMAC</b>	Good	Good	Good	Mid	Good	Good	Very Good	Good	Good	Good	Good	Good
<b>TMAC</b>	Very Bad	Very Bad	Very Bad	Very Bad	Very Good	Good	Very Good	Good	Very Good	Very Good	Very Good	Very Good
<b>ZigBee MAC</b>	Very Good	Very Good	Very Good	Good	Very Good	Good	Good	Good	Mid	Mid	Mid	Mid

## 5 CONCLUSION

WBANs deliver talented applications in health monitoring systems to amount stated physiological information and deliver position-based info. In this paper, we offered an overview of the present MAC protocols for Body Area Network namely ZigBee (IEEE 802.15.4), BanMAC (IEEE 802.15.6), TMAC, SMAC, and BMAC. We also studied the performance of these protocols under two different numbers of nodes in terms of power consumption, packets congestion, and average delay using Castalia under OMNET++ simulator.

The analysis shows that ZigBeeMAC and SMAC show the high number of delays in high traffic. TMAC and SMAC show better average power consumption than the other protocols. TMAC gives the best results of congestion avoidance in different traffic load comparing to the rest four protocols.

## 6 REFERENCES

[1] Z. Albayrak and A. Zengin, "Bee-MANET: A new swarm-based routing protocol for wireless ad hoc networks," *Elektron. ir Elektrotehnika*, 2014, doi: 10.5755/j01.eee.20.3.3421.

[2] Q. Gao, K. J. Blow, D. J. Holding, I. W. Marshall, and X. H. Peng, "Radio range adjustment for energy efficient wireless sensor networks," *Ad Hoc Networks*, 2006, doi: 10.1016/j.adhoc.2004.04.007.

[3] A. N. Ozalp, Z. Albayrak, and A. Zengin, "Expansion of Wireless Networks using IEEE 802.3af Protocol in Protected Areas," in *5th International Symposium on Innovative Technologies in Engineering and Science*, 2017, pp. 1421–1427.

[4] K. Hasan, K. Biswas, K. Ahmed, N. S. Nafi, and M. S. Islam, "A comprehensive review of wireless body area network," *Journal of Network and Computer Applications*. 2019, doi: 10.1016/j.jnca.2019.06.016.

[5] S. K. Dash, S. Mohapatra, and P. K. Pattnaik, "A Survey on Applications of Wireless Sensor Network Using Cloud Computing," *Int. J. Comput. Sci. Emerg. Technol.*, 2010.

[6] K. Pahlavan, M. D. Advisor, and Y. Massoud, "Radio Propagation for Localization and Motion Tracking In Three Body Area Network Applications," no. September, 2016.

[7] F. Ullah, A. H. Abdullah, O. Kaiwartya, J. Lloret, and

M. M. Arshad, "EETP-MAC: energy efficient traffic prioritization for medium access control in wireless body area networks," *Telecommun. Syst.*, 2020, doi: 10.1007/s11235-017-0349-5.

[8] A. Saboor, R. Ahmad, W. Ahmed, A. K. Kiani, Y. Le Moullec, and M. M. Alam, "On Research Challenges in Hybrid Medium-Access Control Protocols for IEEE 802.15.6 WBANs," *IEEE Sens. J.*, 2019, doi: 10.1109/JSEN.2018.2883786.

[9] A. Sevin, C. Bayilmis, and I. Kirbas, "Design and implementation of a new quality of service-aware cross-layer medium access protocol for wireless body area networks," *Comput. Electr. Eng.*, 2016, doi: 10.1016/j.compeleceng.2016.02.003.

[10] B. Liu, Z. Yan, and C. W. Chen, "Medium Access Control for Wireless Body Area Networks with QoS Provisioning and Energy Efficient Design," *IEEE Trans. Mob. Comput.*, 2017, doi: 10.1109/TMC.2016.2549008.

[11] K. Akkaya and M. Younis, "A survey on routing protocols for wireless sensor networks," *Ad Hoc Networks*, 2005, doi: 10.1016/j.adhoc.2003.09.010.

[12] F. Sivrikaya and B. Yener, "Time synchronization in sensor networks: A survey," *IEEE Netw.*, 2004, doi: 10.1109/MNET.2004.1316761.

[13] B. A. Attea and E. A. Khalil, "A new evolutionary based routing protocol for clustered heterogeneous wireless sensor networks," *Appl. Soft Comput. J.*, 2012, doi: 10.1016/j.asoc.2011.04.007.

[14] W. Ye, J. Heidemann, and D. Estrin, "An energy-efficient MAC protocol for wireless sensor networks," 2002, doi: 10.1109/INFCOM.2002.1019408.

[15] M. W. Chiang, Z. Zilic, K. Radecka, and J. S. Chenard, "Architectures of increased availability wireless sensor network nodes," 2004, doi: 10.1109/test.2004.1387396.

[16] M. B. Rasheed, N. Javaid, M. Imran, Z. A. Khan, U. Qasim, and A. Vasilakos, "Delay and energy consumption analysis of priority guaranteed MAC protocol for wireless body area networks," *Wirel. Networks*, 2017, doi: 10.1007/s11276-016-1199-x.

[17] X. Zhang, L. Guo, A. Anpalagan, and A. S. Khwaja, "Performance of Energy-Efficient Cooperative MAC Protocol with Power Backoff in MANETs," *Wirel. Pers. Commun.*, 2017, doi: 10.1007/s11277-016-3580-6.

[18] G. Yang, "A non-cooperative game approach for power control MAC in wireless sensor networks," *Teh. Vjesn. - Tech. Gaz.*, vol. 22, no. 2, pp. 303–310, 2015, doi: 10.17559/TV-20150312155401.

[19] "Processing missing power data in wireless sensor

- networks,” *Teh. Vjesn. - Tech. Gaz.*, vol. 24, no. 4, Aug. 2017, doi: 10.17559/TV-20170320123738.
- [20] Z. ALBAYRAK and H. Musa, “Kablosuz Vücut Alan Ağları için Enerji Verimli MAC Protokolleri (KVAA),” *Acad. Platform-Journal Eng. Sci.*, pp. 127–113, Oct. 2017, doi: 10.21541/apjes.334940.
- [21] M. Chen, S. Gonzalez, A. Vasilakos, H. Cao, and V. C. M. Leung, “Body area networks: A survey,” *Mobile Networks and Applications*. 2011, doi: 10.1007/s11036-010-0260-8.
- [22] W. Ye, J. Heidemann, and D. Estrin, “Medium access control with coordinated adaptive sleeping for wireless sensor networks,” *IEEE/ACM Trans. Netw.*, 2004, doi: 10.1109/TNET.2004.828953.
- [23] T. Van Dam and K. Langendoen, “An adaptive energy-efficient MAC protocol for wireless sensor networks,” 2003, doi: 10.1145/958491.958512.
- [24] J. Polastre, J. Hill, and D. Culler, “Versatile low power media access for wireless sensor networks,” 2004, doi: 10.1145/1031495.1031508.
- [25] P. Khan *et al.*, “Performance analysis of different backoff algorithms for WBAN-based emerging sensor networks,” *Sensors (Switzerland)*, 2017, doi: 10.3390/s17030492.
- [26] M. Zhou and Z. L. Nie, “Analysis and design of ZigBee MAC layers protocol,” 2010, doi: 10.1109/FITME.2010.5654824.
- [27] M. Çakmak and Z. Albayrak, “Performance Analysis of Queue Management Algorithms Between Remote-Host and PG-W in LTE Networks,” *Acad. Platf. J. Eng. Sci.*, pp. 456–463, Sep. 2020, doi: 10.21541/apjes.662677.
- [28] C. D. Guerrero and M. A. Labrador, “On the applicability of available bandwidth estimation techniques and tools,” *Comput. Commun.*, 2010, doi: 10.1016/j.comcom.2009.08.010.
- [29] V. Jha, K. Khetarpal, and M. Sharma, “A survey of nature inspired routing algorithms for MANETs,” 2011, doi: 10.1109/ICETECH.2011.5942042.
- [30] Z. Albayrak and M. Çakmak, “A Review: Active Queue Management Algorithms in Mobile Communication,” *Int. Conf. Cyber Secur. Comput. Sci.*, pp. 180–184, 2018.
- [31] D. Wangerin, C. DeCoro, L. M. Campos, H. Coyote, and I. D. Scherson, “A modular client-server discrete event simulator for networked computers,” *Proc. - Simul. Symp.*, 2002, doi: 10.1109/SIMSYM.2002.1000138.
- [32] Z. Albayrak and C. Torun, “Recent LTE simulation tools,” in *International Conference on Engineering and Natural Sciences (ICENS) 2016*, 2016, pp. 2007–2014.
- [33] G. Fang and E. Dutkiewicz, “BodyMAC: Energy efficient TDMA-based MAC protocol for wireless body area networks,” 2009, doi: 10.1109/ISCIT.2009.5341045.
- [34] N. Bradai, L. C. Fourati, S. Boudjit, and L. Kamoun, “New priority MAC protocol for wireless body area networks,” 2013, doi: 10.1145/2491148.2491149.
- [35] K. S. Prabh, F. Royo, S. Tennina, and T. Olivares, “BANMAC: An opportunistic MAC protocol for reliable communications in body area networks,” 2012, doi: 10.1109/DCOSS.2012.37.
- [36] G. Zhou, J. Lu, C.-Y. Wan, M. D. Yarvis, and J. A. Stankovic, “BodyQoS: Adaptive and Radio-Agnostic QoS for Body Sensor Networks,” 2008, doi: 10.1109/infocom.2008.105.
- [37] K. A. Ali, J. H. Sarker, and H. T. Mouftah, “Urgency-based MAC protocol for wireless sensor body area networks,” 2010, doi: 10.1109/ICCW.2010.5503911.
- [38] C. Li, J. Li, B. Zhen, H. B. Li, and R. Kohno, “Hybrid Unified-slot Access protocol for wireless body area networks,” *Int. J. Wirel. Inf. Networks*, 2010, doi: 10.1007/s10776-010-0120-2.
- [39] N. F. Timmons and W. G. Scanlon, “An adaptive energy efficient MAC protocol for the medical body area network,” 2009, doi: 10.1109/WIRELESSVITAE.2009.5172512.
- [40] I. Anjum, N. Alam, M. A. Razzaque, M. Mehedi Hassan, and A. Alamri, “Traffic priority and load adaptive MAC protocol for QoS provisioning in body sensor networks,” *Int. J. Distrib. Sens. Networks*, 2013, doi: 10.1155/2013/205192.
- [41] J. S. Yoon, G. S. Ahn, S. S. Joo, and M. J. Lee, “PNP-MAC: Preemptive slot allocation and non-preemptive transmission for providing QoS in body area networks,” 2010, doi: 10.1109/CCNC.2010.5421718.
- [42] B. Otal, L. Alonso, and C. Verikoukis, “Highly reliable energy-saving mac for wireless body sensor networks in healthcare systems,” *IEEE J. Sel. Areas Commun.*, 2009, doi: 10.1109/JSAC.2009.090516.
- [43] M. M. Monowar, M. M. Hassan, F. Bajaber, M. Al-Hussein, and A. Alamri, “McMAC: Towards a MAC protocol with multi-constrained QoS provisioning for diverse traffic in Wireless Body Area Networks,” *Sensors (Switzerland)*, 2012, doi: 10.3390/s121115599.
- [44] S. Pandit, K. Sarker, M. A. Razzaque, and A. M. Jehad Sarkar, “An energy-efficient multiconstrained QoS aware MAC protocol for body sensor networks,” *Multimed. Tools Appl.*, 2015, doi: 10.1007/s11042-014-1999-x.

**Contact information:**

**Zafer ALBAYRAK, Dr., Lecturer**

Faculty of Engineering, Department of Computer Engineering  
Karabuk University, Karabuk TURKEY  
[zalbayrak@karabuk.edu.tr](mailto:zalbayrak@karabuk.edu.tr)

**Hatem MUSA, MSc,**

Faculty of Engineering, Tobruk University Libya,  
[htoph@yahoo.com](mailto:htoph@yahoo.com).

**Muhammet ÇAKMAK, Ph.D.,**

Faculty of Engineering, Department of Computer Engineering  
Karabuk University, Karabuk TURKEY  
[muhammetcakmak@karabuk.edu.tr](mailto:muhammetcakmak@karabuk.edu.tr)



# Tüketici Yorumları Üzerine Bir Metin Madenciliği ve Veri Boyutu İndirgeme Yaklaşımı

Ahmet YÜCEL\* 

Ankara Yıldırım Beyazıt Üniversitesi, Ankara, Türkiye  
ayucel@ybu.edu.tr

## Öz

Veri boyutunun artmasıyla orantılı olarak değişkenler arası bağlantıların analizi daha karmaşık bir hale gelmiştir. Yapısal olmayan veri kümelerinde, yapısal forma dönüştürme ön işlemleriyle birlikte, analiz süreçleri daha karmaşık hale gelecektir. Konuşma dilinin doğası gereği, sıradan bir doküman dahi yüzlerce farklı terim içermektedir. Bu durum veri çıkarımı ve verinin yapısal forma dönüştürme süreçlerini oldukça uzatmaktadır. Bu çalışmada kullanılan veri, E-ticaret olarak adlandırılan, çevrimiçi alışveriş işlemleri sırasında ve sonrasında, gerçek kişiler tarafından yazılan yorumlardan oluşmaktadır. Alternatif bir alışveriş yöntemi olan e-ticaret platformlarında, tüketiciler istedikleri ürüne/hizmete ait birçok farklı seçeneği eşzamanlı inceleyebilmektedir. Tüketiciler bu sayede almış oldukları ürünle/hizmetle alakalı tecrübelerini/düşüncelerini kolayca ifade edebilirken, diğer tüketicilerin yorumlarına ulaşma fırsatını da bulabilmektedir. Bu durum metin veri açısından, sürekli büyüyen bir kaynak oluşturmaktadır. Veri boyutunun sürekli olarak artması, veri analizindeki zorluğu da aynı oranda arttırmaktadır. Boyut problemini aşmak için metin madenciliği (MM) çalışmalarında oldukça popüler olan veri boyutu indirgeme yöntemlerinden biri, Tekil Değer Ayrışımı (TDA) kullanılmaktadır. Bu çalışmada, sınıflandırmaya dayalı polarite yöntemi, kompozit (bileşik) bir değişken oluşturma sürecinde kullanılmaktadır. Oluşturulan kompozit değişken, veri içinde yer alan tüm kelime ve kelime gruplarının matematiksel olarak bir araya getirilmesiyle oluşmaktadır. Dolayısıyla ilgili değişken bir veri boyutu indirgeme fonksiyonu da sağlamaktadır. TDA ve kompozit değişkenin, veri boyutu indirgeme performansları kıyaslanmaktadır. Modelleme yöntemi olarak, Genelleştirilmiş Lineer Model (GLM) kullanılmaktadır. Modellerin performansları 5-katmanlı-çapraz-doğrulama yöntemiyle değerlendirilmektedir. TDA skorları ve kompozit değişken kullanılarak GLM modelleri oluşturulmaktadır. Sonuçlar, 5-katmanlı tamamında kompozit değişkenin TDA skorlarından ortalama %6 civarında daha iyi performans sağladığını göstermektedir. Bu yaklaşım, MM'nin veri analizi sürecini kolaylaştırmada ve doğruluk performansını arttırmada önemli bir katkı sağlayacaktır.

**Anahtar kelimeler:** Metin madenciliği, Polarite, Tekil Değer Ayrışımı.

## Text Mining and Data Dimension Reduction Approach on Consumer Comments

### Abstract

In proportion to the increase in data size, the analysis of connections between variables has become more complex. In unstructured datasets, analysis processes would become even more complex with transformation preprocessing to structural form. Due to the nature of natural speaking language, even an ordinary document contains hundreds of different terms. This situation extends the processes of data extraction and transformation. The data used in this study consists of comments written by real persons during and after online shopping, called e-commerce. On e-commerce platforms, which are an alternative shopping method, consumers can simultaneously examine many different options of a product / service. In this way, consumers can easily express their experiences / thoughts about the product / service they have purchased, and have the opportunity to access other consumers' comments. This situation creates a continuously-growing source of text data. The continuous increasing in data size increases the difficulty in data analysis at the same rate. One of the most popular data size reduction methods in text mining (MM) studies, Singular Value Decomposition (SVD) is used to overcome the size problem. In this study, the classification-based polarity method is used in the process of creating a composite variable. The composite variable is formed by mathematically combining all words and phrases in the document. Therefore, the relevant variable also provides a data size reduction function. Data size reduction performances of the SVD and the composite variable are compared. Generalized Linear Model (GLM) is used as the modeling method. The performances of the models are evaluated with 5-fold-cross-validation method. GLM models are created by using TDA scores and composite variables. The results show that the

\* Sorumlu yazar: Ahmet YÜCEL  
E-posta adresi: ayucel@ybu.edu.tr

Alındı : 28 Eylül 2020  
Revizyon : 24 Kasım 2020  
Kabul : 26 Kasım 2020

composite variable outperformed the SVD scores on average by about 6% in the all 5-layers. This approach will make a significant contribution in facilitating the data analysis process of MM and increasing its accuracy performance.

**Keywords:** Text Mining, Polarity, Singular Value Decomposition.

## 1. Giriş (Introduction)

Bilgisayar ve iletişim teknolojileri büyük bir hızla gelişimini sürdürürken, sosyal hayatın her aşamasında da aynı hızla yerini almaktadır. Günlük hayatın doğal akışı, gittikçe daha da artan bir şekilde elektronik bir form kazanmaktadır. Ekonomiden sağlığa, hemen her alanda devam eden süreçler, anlık olarak bilgisayar ortamında kayıt altına alınmaktadır. Bu durum, anlık kayıtlı veri oluşumunun büyüklüğünü ve hızını insan hayal gücünün de ötesine taşıırken, veri madenciliği ve veri analizi yapmak için mükemmel bir imkân sağlamaktadır. Sürekli olarak artan veri boyutları, veri analizi süreçlerini oldukça güçleştirmektedir. Bu amaçla mevcut yöntemlere ek olarak birçok yeni veri boyutu indirgeme yöntemi geliştirilmektedir (Varghese, 2012).

Bir verinin kalitesini belirleyen unsurların başında verinin ulaşılabilirliği, uygun büyüklüğe sahip oluşu, gerçekliği, ilgili alanı kapsayıcılığı, objektifliği ve güncelliği gelmektedir (Pipino, 2002). Bu unsurlar dikkate alındığında, sosyal platformlarda oluşan verinin kıymeti ve sağlayabileceği bilginin kalitesi açıkça anlaşılmaktadır. Elektronik ortamda oluşan verinin %80'den fazlasının yapısal olmayan formda olduğu bilinmektedir (Rajalakshmi, 2015). Bu durum, yapısal olmayan bir veriyi yapısal forma dönüştürmede kullanılan metin madenciliğinin de önemini ortaya koymaktadır. Bu çalışmamızda, e-ticaret alanında, tüketici yorumlarına dayalı bir metin madenciliği uygulamasını göstereceğiz. Bu sayede, TDA'nın ve polariteye dayalı geliştirilen bir kompozit değişkenin veri boyutu indirgemedeki performanslarını inceleyeceğiz.

Elektronik ticaret (E-ticaret) elektronik iletişim yollarıyla, ürün ve hizmetlerin verilmesi anlamına gelmektedir. Her ne kadar 2000'li yıllardan sonra yaygınlığı artsa da, e-ticaretin ortaya çıkışı 30 yıl geriye kadar uzanıyor. Elektronik veri değişimi (EVD), elektronik ortamda belgelerinin bir bilgisayardan diğerine standart bir formatta ulaştırılabilmesi yöntemi, e-ticaretin yapılabilmesini mümkün kılan temel başlangıç noktası olmuştur. EVD, 1960'lı yıllarda, bazı taşımacılık ve perakendecilik sektörlerinin kırtasiye masraflarını azaltma çabasıyla ortaya çıkmıştır. 1990'lara gelindiğinde, halen Avrupa'daki ve Amerika Birleşik Devletleri'ndeki şirketlerin çok azı, EVD yöntemini kullanmaya başlamıştı. Ancak bu daha sonra geliştirilen World Wide Web (WWW) ve bununla ilişkili olarak kaynak adres bilgisi veren URL ve HTML dili e-ticaretin bugünkü seviyesine ulaşmasının önünü açmıştır (Tian, 2008). E-ticaretle birlikte kişi ve işletmelerin alış/satış, tedarikçileri ve müşterileri ile ilişkilerini yönetme, lojistik ve envanter düzenleme süreçleri hızlanmış ve kolaylaşmıştır. Dolayısıyla

maliyetler de önemli ölçüde azalmıştır. Tüm bu sağladığı kolaylık ve hızla, e-ticaret yeni bir rekabet alanı oluşturmuştur. Bu amaçla Joyo Amazon, DangDang gibi birçok şirket e-ticaret sektörüne güçlerini birleştirerek girmiştir. Tekstil endüstrisi de geleneksel pazarlama yöntemini, e-ticaretin ortaya çıkması nedeniyle değiştirmiştir. Zhenxiang ve Lijie çalışmalarında Çinli Zara ve Vancı tekstil firmalarının online perakendecilik alanında bir araya gelmelerini ve bu alanda gösterdikleri başarılarını incelemiştir (Zhenxiang, 2011).

Uzun süre sınırlı alanlarda kullanılan e-ticaret, bugün bütün sektörlerde kullanılmaya başlanmıştır. Bu durum müşteri etkileşimlerini kayıt altına alma imkânı sağlamıştır. Bu amaçla hizmet etkileşimlerinde müşterilerin yorum ve görüşlerini incelemek için duygu analizi teknikleri geliştirilmiştir. 2018'de Yom-Tov tarafından yapılan bir çalışmada müşterilerin duygu analizini otomatik olarak yapacak bir algoritma geliştirilmiştir. Bu sayede, müşteri görüş ve duyguları dinamik bir biçimde takip edilmiştir. Başlangıçta negatif müşteri yorumlarına konu olan aksaklıklar belirlenmiş ve sonrasında ilgili firmaların bu alanlar üzerine yaptıkları düzeltmelerle, müşteri yorumlarının pozitive döndüğü gözlemlenmiştir. Bu sayede, hizmet etkileşimi sırasında müşteri duygu dinamikleri ile hizmet başarısızlığı ve iyileştirme kavramları arasındaki ilişki unsurları belirlenmiştir. Bu durum gelecekteki hizmet kalitesini arttırmaya yardımcı bir bilgi kaynağı olarak değerlendirilmiştir. Yom-Tov çalışmasında, müşteri etkileşimi ile hizmet kalitesi arasındaki bağlantıyı net bir biçimde ortaya koymuştur ve web tabanlı hizmet kalitesinin gerçek zamanlı izlenebilmesi ve kontrolü için duyarlılık analiz araçlarının kullanılmasını önermiştir (Yom-Tov, 2018). Elbette e-ticaret alanında faaliyet gösteren firmalar müşterilerin duygu ve eğilimlerini takip etmek için geleneksel yöntemlerin yanı sıra, sosyal medyayı da aktif olarak kullanmaktadır. Bu amaçla Twitter gibi sosyal medya platformları önemli bir bilgi kaynağı oluşturmaktadır. Firmalar kendilerini ilgilendiren bir konuda mevcut olan bir başlık etiketi (hashtag) altına yapılan paylaşımları takip ederek veya kendilerinin belirleyeceği bir başlık etiketi yardımıyla, insanların firma tarafından paylaşılan bazı sorulara cevaplarını veya genel yorumlarını alması, firmanın müşterileriyle dinamik ve eş zamanlı bir etkileşim kurmasını sağlayacaktır. Buna ek olarak müşteriler tarafından paylaşılan bu veri, diğer tüketiciler için hizmet ve ürünler hakkında fikir sahibi olmalarını sağlayacaktır. Yani, bu değerli veriler hem firmanın hem de tüketicinin kararlarını desteklemek için kullanılabilir. Al-Otaibi konuyla alakalı çalışmasında, metin formatında Twitter verilerini, destek vektör makinesi (Support Vector Machine

(SVM)) algoritması ile pozitif veya negatif şeklinde sınıflandırmıştır (Al-Otaibi, 2018). İnternet kullanımının yaygınlaşması ve internet ortamında kişisel fikir ifade etme kolaylığı, internet üzerinden algı çalışması yapabilmek için çok verimli bir alan olmuştur. Son on yılda söz konusu metin verilerin otomatik şekilde toplanması ve işlenmesi üzerine birçok duygu analizi algoritması geliştirilmiştir. Bu algoritmaların temel amacı bir metnin öznel olup olmadığını belirlemek ve eğer öznel ise, olumlu veya olumsuz bir görüş olduğunu tespit etmektir (Pajupuu, 2016).

Duygu analizinin temel amacı bireylerin bir konu hakkındaki eğilimlerini tespit etmek ve gelecek eğilimleri hakkında öngörü geliştirmektir. Pajupuu yaptığı çalışmada öznel metinlerin polaritesini otomatik olarak belirleyen makine öğrenimine ve kelime karakter dizilimine (lexical sequence) dayalı algoritmalar geliştirmiştir ve iki modelin performanslarını kıyaslamıştır. İlgili çalışmada makine öğrenimi yöntemi daha başarılı sonuç üretmiştir (Pajupuu, 2016). Benzer bir polarite çalışması da Singh tarafından yapılmıştır. Yapılan çalışmada Twitter'dan blog paylaşımlarına birçok farklı kaynaktan veri toplanmış ve duygu analizinde makine öğrenimine ve kelime karakter dizilimine (lexical sequence) dayalı algoritmaların performansları incelenmiştir. Ayrıca yapılan çalışmada gözetimli (supervised) ve gözetimsiz (unsupervised) algoritmalar da karşılaştırılmıştır. Çeşitli kaynaklardan elde edilen metinlerin duygu analizi ve sezgisel tabanlı duyarlılık modellemesi için makine öğrenimi algoritmalarına ek olarak, sözlüğe dayalı yöntemlerin performansı da değerlendirilmiştir. Bu sayede sinema filmi veya benzeri ürünler için yapılan yorumların polarizasyonu sağlanmıştır. Her iki yaklaşımın kıyaslaması da ek olarak yapılmıştır (Singh, 2014). Singh'in sözlüğe dayalı yöntemlerin polarizasyon performansını makine öğrenimi algoritmalarının performanslarıyla kıyaslamıştır. Bunun yanında, makine öğrenimi algoritmalarının da kendi içinde performans olarak ayrıştığını tespit etmiştir. Zubrinic, beş farklı makine öğrenimi algoritmasını duygu analizi modellerinde uygulamıştır. Çevrimiçi müşteri yorumlarının polarizasyonuna dayalı bu çalışmada, Naive Bayes, Destek Vektör Makinesi, Yapay Sinir Ağları ve Maksimum Entropi C4.5 algoritmalarının performansı gözlemlenmiştir. Sonuçlar Destek Vektör Makinesi ve Maksimum Entropi C4.5 yöntemlerinin daha yüksek doğruluğa ulaştığını göstermektedir (Zubrinic, 2018). Zhao ve Xu ise çalışmalarında Yapay Sinir Ağlarının özel bir yaklaşımını doğal dil işleme sürecinde uygulamıştır. Çalışmada duygusal polarite ve yapay sinir ağları bir metin içinde yer alan çeldirici/manipüle edici ifadelerin tespitinde kullanılmıştır (Zhao, 2018).

Metin madenciliğinin en temel amacı, bilgi kaybetmeden metin içinde yer alan sınıfsal ve anlamsal bağlantıları belirlemektir. Ancak veri boyutunun hızla artması bu konuda önemli bir zorluk oluşturmaktadır.

Arunachalam yaptığı çalışmada Bayesian sınıflandırmadan genetik sınıflandırmaya birçok sınıflandırma teknikleri üzerine fikirler geliştirmiştir (Arunachalam, 2017). Boling ve Das yaptıkları metin sınıflaması çalışmasında veri boyutunun oluşturduğu zorlukları aşmak için Tekil Değer Ayrışımı (TDA) veri indirgeme yöntemini kullanmıştır. TDA yönteminin kullanılmasının en temel sebebi, büyük boyutlu bir verinin getirdiği veri önileme sürecinin zorluklarını azaltmak ve bunu yaparken orijinal veriden bilgi kaybını en az seviyede tutmaktır (Boling, 2015).

Bu çalışmada çevrimiçi alışveriş işlemleri sırasında ve sonrasında, paylaşılan müşteri yorumları üzerine duygu analizi yapılmıştır. Yöntem olarak sınıflandırmaya dayalı polarite ve kompozit değişkene dayalı boyut indirgeme yaklaşımları uygulanmıştır. Genelleştirilmiş Lineer Modeller (GLM) üzerinden, kompozit değişkenin sınıflandırma performansı ve veri boyutu indirgeme performansı TDA ile kıyaslanmıştır. Kompozit değişkenin TDA skorlarından ortalama %6 civarında daha iyi performans sağladığı gözlemlenmiştir. Bu yaklaşımla, metinsel veri analizi sürecinin kolaylaştırılması ve doğruluk performansının artırılması hedeflenmiştir.

## 2. Materyal ve Yöntem (Material and Method)

### 2.1. Polarite (Polarity)

Polarite kavramı, çalışmaların içeriğine bağlı olarak farklı amaçla kullanıma sahip olsa da, bu çalışmada, bir dokümanı oluşturan öznel terimlerin olumlu ya da olumsuz olarak sınıflandırılması anlamında kullanılmıştır. Tomar yaptığı çalışmada polariteyi iki aşamalı olarak ele almıştır. Bunlar, bir terimin ön polaritesi ve bağlamsal polaritesi olarak ifade edilebilir. Ön polarite, terimin yalın halde taşıdığı anlama göre, bağlı polarite ise terimin bulunduğu cümle içinde taşıdığı anlama göre ele alınması olarak özetlenebilir (Tomar, 2016). Bağlı polarite sürecine analizcinin sezgisel olarak katılımı gerekmektedir. Bu da tamamen otomatik bir süreci imkânsız kılmaktadır. Bu sebeple çalışmada ön polariteye dayalı bir metin işleme süreci uygulanmıştır.

Çalışmanın bu kısmı, kişilerin duyguları ile terimler arasında ön polariteye dayalı bir ilişki tespiti hakkındadır. Temel amaç, her bir dokümanın içerdiği pozitif (“tavsiye edilen” recommended) ve negatif (“tavsiye edilmeyen” not recommended) terimlerin tespit edilmesi ve daha sonra mevcut bilgiyi matematiksel olarak dönüştürüp ilgili dokümanın genel pozitif oranının hesaplanmasıdır. Bu hesabın temel amacı ise, verinin toplandığı alanla ilgili polariteye dayalı bir sözlük oluşturmak ve her bir terimin taşıdığı polarite ağırlığının büyüklüğünü ve yönünü belirlemektir. Böylece ilgili alanda yer alan benzer verilerin otomatik duygu analizi yapılabilir olacaktır. Bu amaçla, her bir dokümanın pozitif polarite oranını içeren

kompozit (bileşik) bir değişken hesaplanmıştır. Polariteye dayalı oluşturulan kompozit değişkeninin matematiksel ifadesi aşağıdaki verilmiştir: Kullanılan veride mevcut bağımlı ve bağımsız değişkenler ikili (binary) formdadır. Yani,

$$a_{ij} = \begin{cases} 1, & \text{j. terim i. dokümanda mevcut} \\ 0, & \text{j. terim i. dokümanda mevcut değil} \end{cases} \quad (1)$$

Ayrıca, modelin başarısını en iyi şekilde test edebilmek için, bağımlı değişkenin her iki kategorisi için de eşit sayıda durum alınmıştır. Bu sebeple, yapısal olmayan veriden elde edilen terim sayısı  $n$  olmak üzere, oluşturulan yapısal verinin matris boyutu  $(2k) \times n$  şeklinde ifade edilmiştir.

$$\mathbf{X}_{(2k) \times n} = \begin{bmatrix} a_{11} & a_{12} & \dots & a_{1n} \\ a_{21} & a_{22} & \dots & a_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ a_{k1} & a_{k2} & \dots & a_{kn} \\ a_{(k+1)1} & a_{(k+1)2} & \dots & a_{(k+1)n} \\ \vdots & \vdots & \ddots & \vdots \\ a_{(2k)1} & a_{(2k)2} & \dots & a_{(2k)n} \end{bmatrix} \quad (2)$$

$\mathbf{X}_{(2k) \times n}$  metin veriden elde edilen yapısal veri olsun öyle ki, ilk  $k$  adet satır bağımlı değişkenin ‘0’ (“Not Recommended”) kategorisine, ikinci  $k$  adet satır ise bağımlı değişkenin ‘1’ (“Recommended”) kategorisine karşılık gelsin. Yani,

$$\mathbf{A}_{k \times n} = \begin{bmatrix} a_{11} & a_{12} & \dots & a_{1n} \\ a_{21} & a_{22} & \dots & a_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ a_{k1} & a_{k2} & \dots & a_{kn} \end{bmatrix} \quad (3)$$

$$\mathbf{B}_{(2k) \times n} = \begin{bmatrix} a_{(k+1)1} & a_{(k+1)2} & \dots & a_{(k+1)n} \\ a_{(k+2)1} & a_{(k+2)2} & \dots & a_{(k+2)n} \\ \vdots & \vdots & \ddots & \vdots \\ a_{(2k)1} & a_{(2k)2} & \dots & a_{(2k)n} \end{bmatrix} \quad (4)$$

olmak üzere,  $\mathbf{A}_{k \times n}$  matrisi bağımlı değişkenin ‘0’ kategorisine,  $\mathbf{B}_{(2k) \times n}$  matrisi ise bağımlı değişkenin ‘1’ kategorisine karşılık gelen kısım olsun. Ayrıca  $\mathbf{V}_j$ ,  $\mathbf{X}_{(2k) \times n}$  matrisinin  $j$ . sütunudur öyle ki, yapısal olmayan veriden elde edilmiş  $j$ . değişkene (terime) karşılık gelmektedir ve değişkene ait birim değerler sırasıyla  $\{a_{1j}, a_{2j}, \dots, a_{kj}\}$  şeklinde verilmektedir.

$$\mathbf{0}_j = \sum_{i=1}^k a_{ij} \quad (5)$$

Öyle ki  $\mathbf{0}_j$   $j$ . terimin yer aldığı ‘0’ kategorili doküman sayısıdır. Benzer olarak,

$$\mathbf{1}_j = \sum_{i=k+1}^{2k} a_{ij} \quad (6)$$

Öyle ki  $\mathbf{1}_j$   $j$ . terimin yer aldığı ‘1’ kategorili doküman sayısıdır. Buna göre,  $j$ . terimin tüm veri için,

‘1’ kategorili dokümanlar içinde yer alma oranı  $\mathbf{1R}_j$  olarak hesaplanır.

$$\mathbf{1R}_j = \frac{\mathbf{1}_j}{\mathbf{0}_j + \mathbf{1}_j} \quad (7)$$

Her bir dokümanın ( $i = 1, 2, \dots, 2k$ ) bireysel ‘kategori 1’ eğilimini tespit etmek için  $\mathbf{Oran1}_i$  değerini hesaplanacaktır. Bu amaçla öncelikle her dokümandan elde edilen toplam terim sayısı ( $\mathbf{TS}_j$ ) hesaplanmıştır.

$$\mathbf{TS}_j = \sum_{i=1}^n a_{ij} \quad (8)$$

$$\mathbf{Oran1}_i = \frac{\sum_{j=1}^n (\mathbf{1R}_j \cdot a_{ij})}{\mathbf{TS}_j} \quad (9)$$

Polariteye dayalı değişken  $\mathbf{Review\_Rec1\_Rate\_N}$ ’e (N Fold (katman) Sayısı (k-fold cross-validation)) ait birim değerleri  $\mathbf{Oran1}_1, \mathbf{Oran1}_2, \dots, \mathbf{Oran1}_{2k}$  şeklinde ifade edilmektedir.

## 2.2. Tekil değer ayrışımı (TDA)

Tekil Değer Ayrışımı (TDA) çok popüler bir matris boyutu küçültme tekniğidir ve istatistikte büyük boyutlu veri kümelerinin boyutlarının bilgi kaybı oluşmadan, daha makul seviyelere indirgenmesi için kullanılmaktadır. Metin madenciliğinin doğası gereği, küçük sayılabilecek metinlerden dahi binlerce özel terim çıkarımı olabildiğinden, TDA metin madenciliği alanında çok yaygın kullanıma sahip bir yöntemdir. TDA indirgeme işlemi özetle şu şekilde yapılmaktadır: A matrisi  $m \times n$  boyutunda, metin veriden üretilmiş ikili (binary) formda bir (terim-doküman) frekans dağılım tablosu olsun öyle ki  $m$  doküman (tüketici yorum) sayısı ve  $n$  çıkarılan (seçilmiş) terimlerin sayısıdır.

$$\mathbf{A} = \mathbf{UDV}' \quad (10)$$

Öyle ki  $\mathbf{U}$ ,  $m \times r$  boyutunda ortogonal bir matris,  $\mathbf{V}$ ,  $n \times r$  boyutunda ortogonal bir matris,  $\mathbf{V}'$  matrisi  $\mathbf{V}$  matrisinin eşlenik transpozesi ve  $\mathbf{D}$   $r \times r$  boyutunda kare bir matristir. Burada  $r$   $\mathbf{A}'\mathbf{A}$  çarpımının özdeğer (eigenvalue) sayısıdır öyle ki  $\mathbf{A}'$  matrisi  $\mathbf{A}$  matrisinin eşlenik transpozesidir. Modellere yerleştirdiğimiz kavram matrisi, doküman skor matrisi olarak da işlev gören,  $\mathbf{U}$  matrisidir (Yucel, 2016).

## 2.3. Genelleştirilmiş Lineer Model (Generalized Linear Model)

$Y$  bir bağımlı değişken ve  $X_1, X_2, \dots, X_n$  rastgele değişkenler olsun öyle ki,  $P(Y | X_1, X_2, \dots, X_n)$  şartlı dağılımına sahip genelleştirilmiş lineer model, genel olarak şu bileşenlerden oluşur (Levy, 2012):

- 1-  $X_i$ 'lerin  $Y$  üzerindeki etkileri  $\eta$  fonksiyonu ile ifade edilir öyle ki  $\eta$   $X_i$ 'lerden oluşan lineer bir kombinasyondur.
- 2- Model bir link fonksiyonu üzerine kurulur.
- 3-  $l: \eta \rightarrow E(Y) = \mu$  tersinir bir fonksiyon olmak üzere,

$$\eta = \beta_0 + \sum_{i=1}^n \beta_i X_i \quad (10)$$

ifadesi bir genelleştirilmiş lineer modeldir öyle ki  $\eta = l(\mu)$  link fonksiyonudur.

### 3. Uygulama ve Bulgular (Application and Findings)

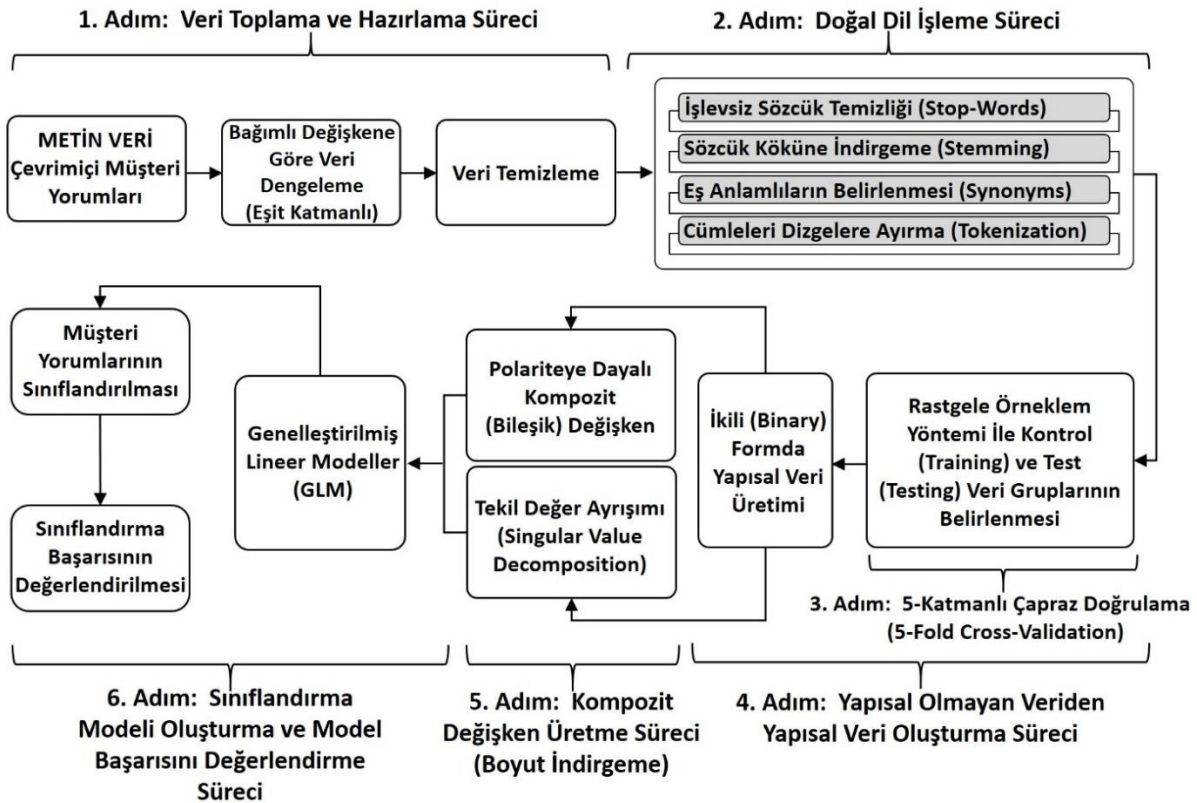
#### 3.1. Veri (Data)

Çalışmada kullanılan [verinin](#) orijinal adı 'Women's Clothing E-Commerce'. Veri ikili (binary) formda olup, internet üzerinden bayan kıyafetleri satışı yapan bir firmanın müşterilerinin, ürünler hakkındaki yorumlarını ve beğenme düzeylerini içermektedir. Verinin bağımlı değişkeni 'Recommend' (tavsiye) (0: No, 1: Yes) olarak belirlenmiştir. Veri içinde toplam 7480 tüketici yorumu

bulunmaktadır ve her iki kategori (0/1) için tüketici yorum sayıları eşittir. Veri gerçek bir ticari işletmeleri konu aldığından, veri içinde yer alan firma isimleri 'perakendeci' olarak değiştirilmiştir. Veri CC0 1.0 Universal lisansına sahiptir. Lisansın detaylarına [creativecommons.org/publicdomain/zero/1.0/](https://creativecommons.org/publicdomain/zero/1.0/) adresinden ulaşabilirsiniz.

#### 3.1. Uygulama Adımları (Implementation steps)

Çalışma 6 temel adımdan oluşmaktadır. Birinci adım üç aşamalı olarak uygulanır. İlk aşama, uyumlu bir algoritma ile direkt olarak veri kaynağından veya hazır paket olarak ilgili bir veri tabanından veri teminidir. Çalışmada kullanılan veriseti, bir online veri paylaşım platformu olan Kaggle.com'dan alınmıştır. İkinci aşamada, belirlenen bağımlı değişkenin kategorilerine göre dağılımın dengeli olması için, rastgele seçim yöntemini kullanarak her iki kategorinin de (0 veya 1) eşit sayıda doküman sayısına sahip olması sağlandı. Verinin ham halinde alınan kaynağıyla alakalı olarak, bazı link, dosya uzantıları, adres bilgileri, sayfa/sıra numaraları gibi çalışmayla ilgisi olmayan unsurlar bulunabilir. Bu tarz unsurlar üçüncü aşamada veriden temizlenir ve böylece ilk adım tamamlanmış olur.



Şekil 1: Yöntem ve uygulama adımları(Method and implementation steps)

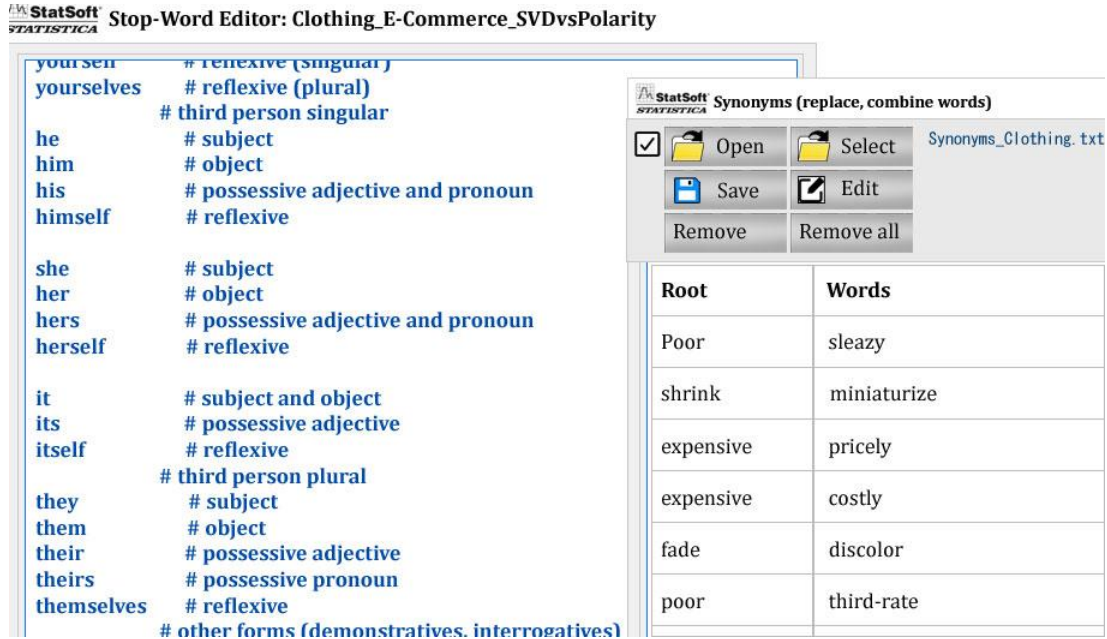
İkinci adım genel olarak doğal dil işleme süreci şeklinde ifade edilebilir. İlgili veri İngilizce metinlerden oluşmaktadır. Metin işlemede kullanılan metin

madenciliği aracı StatSoft Statistica12, İngilizce için hazır ve genel bir 'işlevsiz sözcükler' (stop-words) listesi sağlamaktadır. Mevcut veride tespit edilen diğer



işlevsiz sözcükler de aynı listeye eklenmiştir. Bunun yanında, Statistica İngilizce dili için, kelimeleri köklerine indirgeyen (stemming) ve cümleleri dizgelere ayıran (tokenization) algoritmaları da sağlamaktadır. Ek

olarak, veri içinde yer alan eş anlamlı terimler belirlenip, dil analizinden önce programa yüklenmiştir. Şekil 2’de Statistica metin madenciliği sözlük oluşturma sürecine ilişkin bir ekran görüntüsü paylaşılmıştır.



Şekil 2. Metin madenciliği sözlük oluşturma süreci (Text mining dictionary creation process)

Üçüncü adımda, veri eşit sayıda doküman içeren, birbirinden bağımsız ve kesişimi boş küme olan 5 katmana rastgele seçim yöntemiyle ayrılmıştır öyle ki her katmanda yer alan dokümanların, bağımlı değişkenin kategorilerine göre dağılımı da eşit sayıdadır. Şekil 3 ‘de GLM tanımsal kodlama altında yer alan ‘SAMPLE’ kodu ile ifade edilen test grupları belirlenmektedir.

Dördüncü adım temel olarak iki aşamadan oluşmaktadır. İlk aşamada, oluşturulan beş katmandan biri (%20) test verisi olarak, geriye kalan dört katman ise kontrol verisi olarak düzenlenmiştir. Bu işlem, her bir katmanın, sırasıyla bir kez test verisi olarak düzenlenmesini ve böylece beş adımda verinin tamamının test edilmesini sağlamak için uygulanmıştır. İkinci aşamada ise, veriden ikili (binary) formda bir matris elde edilmiştir. Bu matris terim-doküman frekans tablosu olarak adlandırılmaktadır. Genel olarak dördüncü adım, yapısal olmayan (metin) veriden, yapısal ve ikili (binary) formda bir veri elde etme süreci olarak ifade edilebilir.

Beşinci adımda, temelinde yukarıda verilen matematiksel formülleri çalıştıran algoritmalar yardımıyla, polariteye dayalı kompozit değişken ve TDA’ya dayalı doküman skorları hesaplanmıştır. Kompozit değişken ve doküman skorlarının bulunduğu değişkenler, sonraki adımda oluşturulacak olan modellerde bağımsız değişken olarak yer almaktadır.

```
Distribution : BINOMIAL
Link function: LOGIT
Response variable : "Recommended{1:Y/0:N}"

Codes of dependent variable
0 : Primary code (1)
1: Secondary code (0)

Design Effects:
Continuous effects : "REVIEW_REC1_RATE_"
Categorical effects:

Model specification:
GLZ ;
RESPONSE = "Recommended{1:Y/0:N}" (0 1) ;
GROUPS = none;
COVAR ATE = ""REVIEW_REC_RATE_1";
DESIGN = ""REVIEW_REC_RATE_1;
INTERCEPT = include :
PARAM = sigma;
SDELTA = 7;
SURFACE = none;
MIXTURE = none;
SAMPLE = "CV_TESTING1" { 102.} ;
COUNTY = none;
MBUILD = all; |
CONVERGE = 7:I
MAXITER = 100 ;
IN TIALS = none;
OFFSET = none;
OUTPUT = none;
```

Şekil 3. GLM tanımsal kodlama (GLM definitive coding)

Çalışmanın altıncı ve son adımında ise, beşinci adımda üretilen iki grup bağımsız değişkenler yardımıyla Genelleştirilmiş Lineer Modeller kurulmaktadır ve böylece her bir dokümanın bağımlı değişkende karşılık geldiği kategori (0 veya 1) tahmin edilmeye çalışılmıştır. Bir başka ifadeyle, tüketici yorumunun duygu analizi yapılmıştır. Bu amaçla,

oluşturulan model kullanılarak tüm dokümanlar 0 ve 1 kategorilerine göre sınıflandırılmıştır. Son aşamada, her iki bağımsız değişken grubuna dayalı yapılan sınıflandırmaların doğruluk oranları hesaplanmıştır ve böylece hem modellerin hem de değişken gruplarının başarıları ölçülmüştür. Şekil 4'te GLM'e ait parametre ve katsayı belirleme detayları verilmiştir.

```
<Header copyright="STATISTICA Data Miner, Copyright (c) StatSoft, Inc., www.statsoft.com."/>
<DataDictionary numberOfFields="2">
  <DataField name="Recommended(1:Y/0:N)" optype="categorical">
    <Value value="0" NumericValue="0"/>
    <Value value="1" NumericValue="1"/>
  </DataField>
  <DataField name="REVIEW_REC1_RATE_1" optype="continuous"/>
</DataDictionary>
<GeneralizedLinearModel
  functionName="classification"
  modelName="Generalized linear regression"
  modelType="generallinear"
  targetVariableName="Recommended(1:Y/0:N)">
<Extension name="Distribution" value="binomial"/>
<Extension name="LinkFunction" value="logit"/>
<ParameterList>
  <Parameter name="p1" label="Intercept"/>
  <Parameter name="p2" label="REVIEW_REC1_RATE_1"/>
</ParameterList>
<FactorList>
</FactorList>
<CovariateList>
  <Predictor name="REVIEW_REC1_RATE_1"/>
</CovariateList>
<PPMatrix>
  <PPCell value="1" predictorName="REVIEW_REC1_RATE_1" parameterName="p2"/>
</PPMatrix>
<Extension name="CorrectDummyCode" value="1"/>
<Extension name="IncorrectDummyCode" value="-1"/>
<ParamMatrix>
  <PCell targetCategory="0" parameterName="p1" beta="3.16184705908780e+001"/>
  <PCell targetCategory="0" parameterName="p2" beta="-6.45249167608618e+001"/>
</ParamMatrix>
</GeneralizedLinearModel>
</PMML>
```

Şekil 4. StatSoft Statistica TMİD (Tahmin Modeli İşaretleme Dili) GLM algoritması (StatSoft Statistica TMİD (predictive model markup language) glm algorithm)

#### 4. Sonuçlar (Conclusions)

Kullanılan veride toplam 7480 tüketici yorumu bulunmaktadır. 5 katmanlı çapraz doğrulama testi için, her bir test grubunda (%20) toplam 1496 tüketici yorumu bulunmaktadır. Tüketici sayıları her iki kategori (0 veya 1) için eşit (748 adet) şekilde dağıtılmıştır. Başka bir ifadeyle, veri ayrık ve eşit dağılıma sahip beş parçaya ayrılmıştır. Bu şekilde uygulanan modelin performansı beş kez test edilmiştir. Lineer modeller binomial dağılıma sahip bir veri üzerine uygulanmıştır ve logit olasılık fonksiyonu kullanılmıştır. Her bir katman için iki kez lineer model oluşturulmuştur. Her iki modelde de bağımlı değişken olarak, tüketici cevap

bilgisini içeren 'Recommended (0/1)' değişkeni kullanılmıştır. Birinci modelde bağımsız değişkenler TDA skorları, ikinci modelde ise kompozit değişkenler olarak belirlenmiştir. Lineer Modellerin tamamı  $\alpha = 0,05$  anlamlılık düzeyinde, istatistiksel olarak anlamlı sonuç vermiştir. Bu süreçte modeller iki aşamalı test edilmiştir. Birinci aşamada p-değer ( $\alpha = 0,05$ ) parametresinin anlamlı olması beklenmiştir. İkinci aşamada, anlamlı bulunan modelin 'mutlak doğruluk/mutlak doğru tahmin' performansı ölçülmüştür. Bu ölçüm sonuçları karışıklık matrisleriyle ifade edilmiştir. Tablo 1'de beş katman için, modellerin kategorilere göre yaptığı doğru ve yanlış tahmin sayıları (karışıklık matrisleriyle) verilmektedir. Modellerin doğruluk oranları Şekil-5'te verilmiştir.

**Tablo 1.** Her bir katman için, GLM tahminlerine göre doğru veya yanlış sınıflandırma sayıları (Number of correct or incorrect classifications for each layer based on GLM estimates)

Dokümanların Sınıflandırılması		TDA		POLARİTE	
		Tahmin 0	Tahmin 1	Tahmin 0	Tahmin 1
Katman 1	Gözlem 0	653	95	708	40
	Gözlem 1	131	617	84	664
Katman 2	Gözlem 0	648	100	698	50
	Gözlem 1	133	615	80	668
Katman 3	Gözlem 0	628	120	683	65
	Gözlem 1	127	621	88	660
Katman 4	Gözlem 0	628	120	639	109
	Gözlem 1	144	604	76	672
Katman 5	Gözlem 0	597	151	634	114
	Gözlem 1	149	599	86	662

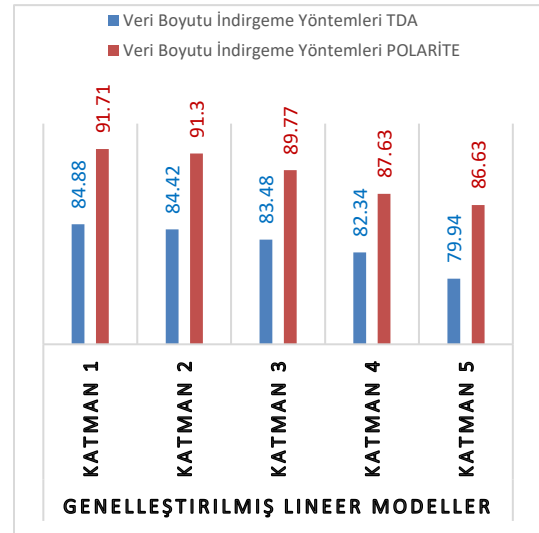
Tablo 2, Tablo 1’de verilen doğru tahmin sayılarının toplam tahmin sayılarına oranı, başka bir ifadeyle, modellerin başarı oranları verilmektedir. Sonuçlardan da açıkça görüldüğü gibi, polariteye dayalı kompozit değişken tüm katmanlarda ortalama %6 daha iyi sonuç vermektedir. Polariteye dayalı değişkenin modellerde sağladığı başarı oranı en düşük %86,63 ve en yüksek %91,71 arasında değişmektedir. Bu durum son yıllarda yapılan benzer polarite çalışmaları dikkate alındığında oldukça dikkat çekici bir başarı sunmaktadır.

**Tablo 2.** Modellerin doğruluk oranları (The accuracy of the models)

Doğruluk Oranları (%)		Veri Boyutu İndirgeme Yöntemleri	
		TDA	POLARİTE
Genelleştirilmiş Lineer Modeller	Katman 1	84,88	91,71
	Katman 2	84,42	91,30
	Katman 3	83,48	89,77
	Katman 4	82,34	87,63
	Katman 5	79,94	86,63

İlgili çalışmaların anlatıldığı kısımda da ifade edildiği üzere, *Zubrinic*’in 2018’de yayınladığı *tüketici*

*Duygu analizi* çalışmasının, ikili (binary) bir veride elde ettiği doğruluk oranı maksimum %84,5 civarında kalmıştır (Zubrinic, 2018). Bunun yanında *Zhao* ve *Xu*’nun detaylarının yukarıda verildiği, 2018’de yayınlanan *fikir tespiti* (opinion detection) çalışmasında, modellerin doğruluk oranları en iyi %70,02 civarında kalmıştır (Zhao, 2018).



**Şekil 5.** Modellerin doğruluk oranları (The accuracy of the models)

## Kaynaklar (References)

- Al-Otaibi, S., Alnassar, A., Alshahrani, A., Al-Mubarak, A., Albugami, S., Almutiri, N., Albugami, A., 2018. Customer Satisfaction Measurement Using Sentiment Analysis, International Journal of Advanced Computer Science and Applications (IJACSA), Vol.9, No.2.
- Arunachalam, N., Sneka S. J., Mathi, G. M., 2017. A Survey On Text Classification Techniques For Sentiment Polarity Detection, Innovations in Power and Advanced Computing Technologies (i-PACT), 1-5. 10.1109/IPACT.2017.8245127.
- Boling, C., Das K., 2015. Reducing Dimensionality of Text Documents Using Latent Semantic Analysis, International Journal of Computer Applications (0975 – 8887), Vol.112, No.5.

- Levy, R., 2012. Probabilistic Models in the Study of Language , ch. 6, pp: 107-108.
- Pajupuu, H., Altrov, R., Pajupuu, J., 2016. Identifying Polarity in Different Text Types, pp 126-138, doi.org/10.7592/FEJF2016.64.polarity.
- Pipino, L. L., Lee, Y. W., Wang, R. Y., 2002. Data Quality Assessment, Communications Of The ACM, Vol.45.
- Rajalakshmi, Narayanan, M., Ramkumar, M., 2015. An Exclusive Study on Unstructured Data Mining with Big Data, International Journal of Applied Engineering Research, Vol.10, No.4, pp.3875-3886.
- Singh, V. K., Piryani, R., Waila, P., Devaraj, M., 2014. Computing Sentiment Polarity of Texts at Document and Aspect Levels, ECTI transactions on computer and information technology, Vol.8, No.1.
- Tian, Y., Stewart, C., 2008. History of E-Commerce, DOI:10.4018/978-1-59904-943-4, ch001.
- Tomar, D. S., Sharma, P., 2016. A Text Polarity Analysis Using Sentiwordnet Based an Algorithm, International Journal of Computer Science and Information Technologies (IJCSIT), pp. 190-193, Vol.7.
- Varghese, N., Verghese, V., Gayathri, P., Jaisankar, N., 2012. A Survey Of Dimensionality Reduction And Classification Methods, International Journal of Computer Science & Engineering Survey (IJCSES) Vol.3, No.3.
- Yom-Tov, G. B., Ashtar, S., Altman, D., Natapov, M., Barkay, N., Westphal, M., Rafaeli, A., 2018. Customer Sentiment in Web-Based Service Interactions: Automated Analyses and New Insights, International World Wide Web Conference Committee (IW3C2), Creative Commons CC BY 4.0 License.
- Yucel, A., 2016. Predictive Text Analytics And Text Classification Algorithms, A dissertation submitted to the Graduate Faculty of Auburn University, pp 19.
- Zhao, S., Xu Z., Liu L., Guo M., Yun, J., 2018. Towards Accurate Deceptive Opinions Detection Based on Word Order-Preserving CNN, Mathematical Problems in Engineering, Article ID 2410206, Vol. 2018
- Zhenxiang, W., Lijie, Z., 2011. Case Study of Online Retailing Fast Fashion Industry, International Journal of e-Education, e-Business, e-Management and e-Learning, Vol.1, No.3.
- Zubrinic, K., Milicevic, M., Sjekavica, T., 2018. Obradovic, I., A Comparison of Machine Learning Algorithms in Opinion Polarity Classification of Customer Reviews, International Journal of Computers, Vol.3.



# Recognition of Sign Language Letters Using Image Processing and Deep Learning Methods

Ali Öztürk<sup>1\*</sup> , Melih Karatekin<sup>2</sup> , İsa Alperen Saylar<sup>3</sup> , Nazım Bahadır Bardakçı<sup>4</sup> 

<sup>1,2,3,4</sup> KTO Karatay University, Faculty of Engineering, Department of Computer Engineering, KONYA/Turkey

aliozturk2002@gmail.com, sefamelihkaratekin@gmail.com, alpisa1907@gmail.com, nazimbahadirbardakci@gmail.com

## Abstract

In order for people to be able to communicate with each other, they must be able to agree mutually. Communication is quite difficult for individuals with hearing problems. Such individuals make their lives much more difficult by isolating themselves from society. The people living with hearing loss can understand the contact person with often lip-reading method, but it is quite difficult for them to express themselves to the people. Since the use of sign language has not become widespread around the world, the number of people who know sign language is very low, except for individuals with hearing disabilities. In this study, it was achieved to dynamically recognize the movements of the sign language finger alphabet via image processing using deep learning methods and to translate it into writing. Accordingly, it is aimed to facilitate communication between people who do not know the sign language in daily life and people with hearing loss. The input given to the system is an image of the hand showing any letter from the alphabet. The image of the hand is interpreted by deep learning methods in the system, and it is compared to one of the letters in the alphabet and an output with the similarity ratio to this letter is displayed on the screen. The system has been tested with a total of 1300 images. The overall accuracy rate of the system was calculated as 88% where true positive rate was 87% and false negative rate was 13%.

**Keywords:** Sign Language, Image Processing, Deep Learning, Image Analysis, Object Detection.

## Görüntü İşleme ve Derin Öğrenme Yöntemleri Kullanılarak İşaret Dili Harflerinin Tanınması

### Öz

İnsanların birbirleriyle iletişim kurabilmeleri için karşılıklı olarak anlaşılabilirliği gerekmektedir. İşitme problemi yaşayan bireyler için iletişim kurmak oldukça zordur. Bu tür bireyler kendilerini toplumdan soyutlayarak yaşamlarını çok daha zor hale getirmektedir. İşitme kaybı yaşayan insanlar iletişimde bulunurken karşısındaki kişiyi genellikle dudak okuma yöntemi ile anlayabilmektedir fakat karşısındaki insanlara kendilerini ifade etmekte oldukça zorlanmaktadır. İşaret dili kullanımının dünya genelinde yeterince yaygınlaşmamış olması sebebiyle, işitme engelli bireyler dışında işaret dilini bilen kişi sayısı oldukça azdır. Bu çalışmada, işaret dili parmak alfabesine ait hareketlerin, derin öğrenme yöntemleri kullanılarak görüntü işleme ile tanınması ve yazıya çevrilmesi dinamik olarak sağlanmıştır. Bu doğrultuda, gündelik yaşamda işaret dilini bilmeyen insanlar ile işitme kaybı olan insanlar arasındaki iletişimi kolaylaştırmak hedeflenmiştir. Sisteme verilen girdi, alfabeden herhangi bir harfi gösteren elin görüntüsüdür. Elin görüntüsü, sistemde derin öğrenme yöntemleriyle yorumlanarak, alfabedeki harflerden biriyle eşleştirilmekte ve bu harfe benzerlik oranı ile birlikte ekrana yazdırılmaktadır. Geliştirilen sistem, 1300 resim ile test edilmiştir. Sistemin doğruluk oranı %88, gerçek pozitif oranı %87 ve yanlış negatif oranı %17 olarak bulunmuştur.

**Anahtar kelimeler:** İşaret Dili, Görüntü İşleme, Derin Öğrenme, Görüntü Analizi, Nesne Tespiti

## 1. Introduction

Creatures use the signals transmitted through various channels when communicating. These communication channels are audible, visual, chemical

etc. Sign language; is a silent, visual language created by the handicapped and speech impaired people by using hand gestures, facial expressions and body language as a whole to communicate with each other. This non-universal language varies from country to

\* Corresponding Author.  
E-mail: aliozturk2002@gmail.com

Received :09 Oct. 2020  
Revision :22 Dec. 2020  
Accepted :22 Dec. 2020

country. In this sense, there are more than 200 sign languages in the world. American, British, German, French, Italian, Indian and Turkish sign languages are some of them. In this article, it is aimed to recognize the Turkish sign language and some English sign language letters (Q, W, X) with image processing using deep learning methods and translate it into writing. Many studies have been done on this subject. As a result of our researches, approximately 4000 image dataset which were specifically created for the Turkish Sign Language alphabet were selected. From this dataset, approximately 1600 photographs were edited and tagged and used in the system training phase. The system was developed with the Python programming language. TensorFlow and OpenCV libraries were used as image processing tools. Image Labeling and Object Detection methods were applied. Thus, the labeling, training and testing of the dataset were achieved. Faster-R-CNN (Convolutional Neural Network) was determined as the deep learning model. In addition, the accuracy rates and performance analysis of the obtained results were performed. As the output of the study, a dynamic and high performance system was developed to facilitate communication between people who do not know sign language and people with hearing impairment.

### *1.1. Literature Survey*

In their paper, Ashish and Aarti (2016) proposed a system hand gesture recognition system where image color analysis and image segmentation methods are used. The proposed system prepared by Sandhya and Ananya (2018) aimed to recognize American Sign Language and convert it into text. The input given to the system was a hand image showing the required alphabet. The histogram of the input image was calculated and checked for similarity to the histograms of previously saved images using the Bhattacharyya Distance Metric. OpenCV was used in the proposed system. The CNN, Machine Learning (ML) methods and depth-sensing technology were used in the study which was performed by Gaikwad et al. (2019). They used a pre-trained GoogleNet architecture. In the study of Tamura and Kawasaki (1988), a 3D dictionary was created from hand shapes. According to this dictionary, 2-dimensional images were interpreted and classified according to their features.

Kumarage et al. (2011) proposed to subdivide the transactions for recognition via parallel processing and mapping the motion data to static data representations. Also, the issue of matching sign language gestures with linear / nonlinear equations was mentioned. In the paper which emerged as a result of the research of Tan et al. (2020), a CNN with spatial pyramid pooling for vision-based hand gesture recognition was introduced. The performance of the proposed method was evaluated on American sign language datasets and

hand gesture dataset. Pramada et al (2013) designed an intelligent system using the concepts of image processing, machine learning and artificial intelligence where sign language alphabet was digitally encoded by filtering. In their research Shivashankara and Srinath (2018), proposed a framework where the 24 static alphabets of the American Sign Language (letter J and letter Z not included) were translated into English text. They achieved an average recognition rate of 98.21%. In their study, Vargas et al (2011) presented an image pattern recognition system using neural network for the identification of sign language. The system had several stored images that showed the specific symbol in a kind of language which was employed to train a multilayer neural network using a backpropagation algorithm. They enabled the definition of the sign language by using image processing with various filtering methods and algorithms. They also stated that different features can be added to improve the accuracy of the system.

In the work of Unutmaz et al. (2019), the Turkish sign language recognition was investigated by hand gestures made in front of Kinect device and converted into writing. Convolutional Artificial Networks (CNN), Support Vector Machines (SVM), k-Nearest Neighbours (kNN) and Decision Trees (DT) were used for recognition and their performances were compared. When the manually extracted video frames were used via data augmentation, the accuracy rate reached to %99. However, when the sequential video frames were used to imitate real-time recognition behaviour, the maximum accuracy rate decreased to %81.

Çelik and Odabaş (2020) suggested a system that automatically detects Turkish sign language for 29 letters in the Turkish alphabet and 10 numbers by hand gestures made in front of the camera. This method has been developed with CNN modeling for sign detection using frames of videos. Different success rates have been achieved by manually processing the frames of videos. It has been reached a 53.8% accuracy rate without preprocessing of hand images using only CNN. By means of manual hand image extraction, %91 accuracy rate has been achieved. When more than one CNNs have been used with LSTM to manually capture the hand movements in the video frames, the accuracy rate has increased to 97%.

## **2. Materials and Methods**

Python 3.7 was preferred as programming language in this project for image processing. Among the deep learning libraries, TensorFlow was chosen as the most suitable one for our system. The training and determination of the dataset were carried out via the TensorFlow library. The video card was one of the main parts of the overall system. The OpenCV libraries on the other hand, enabled the camera to be controlled,

the data to be tagged and the images to be filtered. In addition, it showed which class the detected object belongs to in a rectangle, with a percentage of accuracy.

Any of the drivers prepared by the video card manufacturers for deep learning was required to be able to train the system. These drivers have been developed to make training more efficient, fast and successful. The equipment which was used during training was the Nvidia Ge-Force GTX 1060 graphics card with a capacity of 3 GB, and the CUDA core was 1280.

To apply deep learning methods on this video card, CUDA and cuDNN drivers were downloaded from the video card manufacturer's official page. The most suitable versions of the drivers for the system are CUDA 8.0 and cuDNN 7.1.4. After installing the drivers and assigning their paths to the system variables, a virtual environment was created on the Anaconda IDE. Since the system had a complex structure in itself, it was important to download all libraries completely to the virtual environment.

The algorithm model to be used in training has been determined as Faster R-CNN. As a result of our research, we observed that the most appropriate algorithm model in terms of hardware and object classification time was Faster-R-CNN. Faster R-CNN applies CNN method on to the images. In CNN, a mask filter navigates on the image, the filter converts each mask into a pixel and calculates on the point where it is located, then stores the results in the output matrix. This last reached output matrix is called "feature map". Feature Map makes predictions based on image tagging. Faster RCNN architecture is a combination of region proposal network (RPN) which proposes regions and Fast RCNN detector which uses these proposed regions to give final bounding boxes of the object. An RPN takes feature map as input and outputs a set of rectangular object proposals, each with an objectness score. The objectness measures score between object and background. Faster RCNN uses anchor boxes of 3 aspect ratios and 3 scales. Thus for each pixel in the feature map, there are 9 anchor boxes. The region proposal time with selective search is 2 sec per image, whereas with RPN it's just 10ms (Shaoqing et al., 2017).

### 3. The Proposed System

The main steps of the proposed system is shown in Figure 1.



Figure 1. The main steps of the proposed system

#### 3.1. Preparing The Dataset

In this step, dataset research was carried out for the system to detect the entire communication of the individual, who tries to show a letter from the sign language alphabet by coming across the camera by hand in real-time with high accuracy ratio. The "Turkish Sign Language Dataset" on the Kaggle website has been used in the project for training our system. The dataset contains images of 26 letters that were taken from different angles and made by different individuals. While there were approximately 4000 images, approximately 1600 of them were used for training. Since some images were not of sufficient quality for training, approximately 100 photographs taken by ourselves have been added to the dataset. Some of the sample images in our dataset are given in Figure 2.

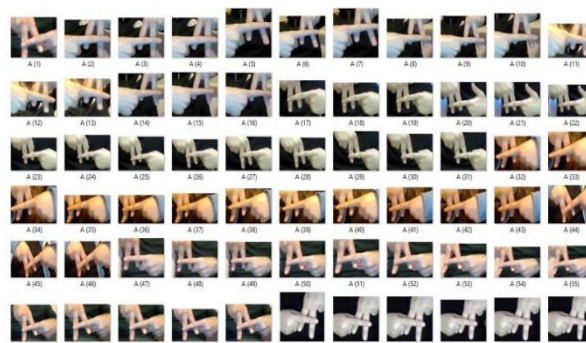


Figure 2. Some images from dataset

Then, the editing or extraction of data that was not sufficient for training and deemed unnecessary in the selected dataset was done. Upon completion of the necessary procedures, approximately 4000 images were reduced to approximately 1600 images. In addition, in the parts that do not seem sufficient, the images we took (approximately 100) were added to the dataset. Thus, the dataset has been prepared for the training. After this stage, the photos were labeled before feeding to the deep learning model for training.

### 3.2. Data Labelling

In order to detect objects in real time in front of the camera, the data must first be introduced to the computer. The photos in the edited dataset were tagged and introduced to the computer to which class it belongs to. This process was done using OpenCV library. After this process, the XML file of each photo was created. However, the computer cannot read XML file during training. Therefore, each XML file has been converted into a single CSV file. This was done using a Python script. Each line in the CSV file contains XML information for a different photo. As a result, the dataset becomes ready for training.

### 3.3. Training

For this step, the libraries required for the training of the system were installed by creating a virtual environment in Anaconda IDE. Among the libraries to be used, the most important for the training of the system is the TensorFlow library. Tensorflow is an open source deep learning library. Being open source allows the problems encountered to be solved easily via customizing the code. One of the biggest advantages of the library; with the APIs it offers, the platform allows us to use one or more CPUs and GPUs independently. In addition, the library's development with Python maximizes its compatibility for our system. Tensorflow contains many configuration files and allows these files to be manipulated for training the system. In this way, the configuration files were prepared in accordance with our system and made ready for training. Tensorflow offers weight files based on the number of data in the system and the training conditions (speed, accuracy rate, etc.). After the weight files were integrated into the system, the number of letters and symbols to be used in the project were also introduced to the system. After this process, the configuration file was edited and the number of output classes was written. This number is 26 in our project.

After then, the training of the system began. The training period lasted approximately 10 hours and the prediction time was less than 2 seconds. Tensorflow displays the loss function after each transaction, informs about the stage of the training and makes it easier for us to make a decision to stop it.

In addition, Tensorflow shows all stages of training on the graph in a web interface running on localhost. Training continues until you achieve the lowest loss rate. When the lowest loss rate is obtained, the training is stopped and the "frozen inference" graphic formed as a result of the training is exported for the operation of the system. The default name of the file is "frozen\_inference\_graph.pb". By means of this file, the performance of the training and the accuracy rates were observed.

### 3.4. Testing

After the training, it is very important to test and check the accuracy of the system. If error ratio is high in real time detection, the system may need to be re-trained and the model may be re-generated. In this step, image quality and resolution are one of the factors that increase the accuracy rate. Two different cameras were used during the test. In addition, the performance of the system was measured by testing hand movements of different people. The effect of complex or empty background during hand movement have been observed on the performance of the system. All 26 letters are statically and dynamically tested. The system has been tested using photos prepared for testing beforehand. Then, real-time tests were carried out using the camera. Figure 3 shows some of the real-time test results.

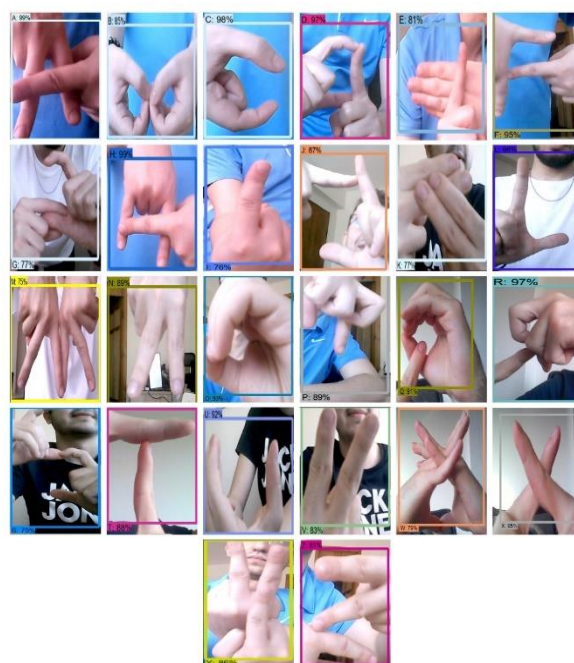


Figure 3. Some of the real-time test results

The statistical analysis was made according to the estimation results produced by the system. There are two cases, the system either shows the correct letter or the wrong letter. While TP (True Positive) shows the result that the model predicts the positive class correctly, FN (False Negative) shows the result where the model incorrectly predicts the negative class. Precision shows how many of the values we predict positively are actually positive. Recall shows how many of the trades we should have predicted positively. Precision and recall values do not give a meaningful comparison when calculated separately. Therefore, it would be more correct to evaluate these two values together. For this, we need to calculate the F-Score value. F-Score is the harmonic mean of precision and recall values.



$$Precision = \frac{TP}{TP + FP}$$

$$Recall = \frac{TP}{TP + FN}$$

$$F - Score = 2 * \frac{Precision * Recall}{Precision + Recall}$$

Table 1 shows the accuracy, the F-Score, true positive and false negative values for each letter that was introduced to the trained system during testing. The results were obtained by showing 50 signs from each letter in front of the camera.

**Table 1.** Test Results Analysis Table

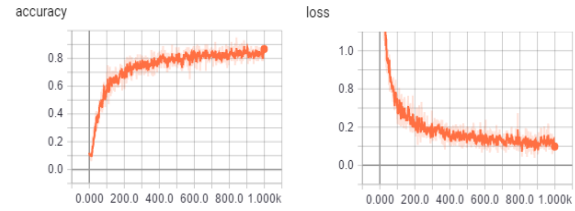
Sign	TP	FN	Accuracy	F-Score
A	98%	2%	98%	98%
B	80%	20%	80%	88%
C	94%	6%	94%	96%
D	80%	20%	80%	88%
E	88%	12%	88%	93%
F	92%	8%	92%	95%
G	74%	26%	74%	85%
H	86%	14%	86%	92%
I	90%	10%	90%	94%
J	80%	20%	80%	88%
K	80%	20%	80%	88%
L	94%	6%	94%	96%
M	76%	24%	76%	86%
N	76%	24%	76%	86%
O	80%	20%	80%	88%
P	88%	12%	88%	93%
Q	74%	26%	74%	85%
R	92%	18%	92%	95%
S	94%	16%	94%	96%
T	90%	10%	90%	94%
U	88%	12%	88%	93%
V	88%	12%	88%	93%
W	86%	14%	86%	92%
X	96%	4%	96%	97%
Y	88%	12%	88%	93%
Z	92%	8%	92%	95%

The system has been tested in real-time with a total of 1300 images. According to the results, the overall accuracy rate of the system was calculated as 88%. The TP rate was 87% and the FN rate was 13%. During the test, it was observed that the delay time of the camera was 28 milliseconds. This indicates that there will be no freezing or noticeable delay when using the system. Thus, users will be able to communicate between them fluently and smoothly.

#### 4. Performance Measurement

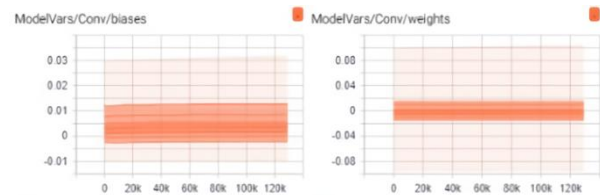
To analyze the performance of the system, the infographics presented by TensorFlow in the local environment were examined after the training phase. Looking at the graphs given in Figure 4, as the iterations increase, accuracy approaches 1 while the

loss function approaches 0. This shows that the more iterations there are, the higher the accuracy rate of training. While the calculated accuracy rate is 88%, the total loss is 0.1.



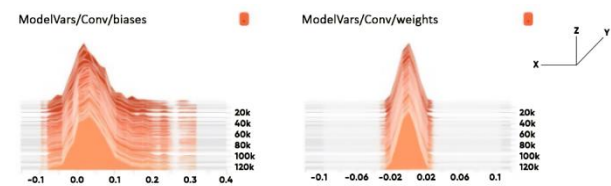
**Figure 4.** Accuracy and Total Loss Graph

The distribution charts of bias and weight are given in Figure 5. The X axis indicates time, the Y axis indicates values. These charts are a top view of histogram charts. It would be more accurate to examine the histogram graphs for interpretation.



**Figure 5.** Biases and Weights Distribution View

Histogram graphs are shown in Figure 6. There are three axes in the histogram. According to the figure, X-axis values (depth), Y-axis time, Z-axis shows the intensity (frequency) of values represented on the Y-axis. Histograms become darker as the depth in the X-axis increases. Taking a higher value on the Z axis means the vector produces a value closer to the specified value.



**Figure 6.** Biases and Weights Histogram

The results in Table 2 were obtained when the methods in the other 2 articles for the recognition of Turkish sign language by image processing were compared.

**Table 2.** Comparison Table with Other Studies

	Dataset	Image Processing	Models	Accuracy
Çelik and Odabaş (2020)	Images from 200 videos	Non-real time	CNN without preprocessing	53%
			CNN with preprocessing	91%
			CNN+LSTM with preprocessing	97%
Unutmaz et al. (2019)	4320 images from videos	Real time	CNN	81%
			SVM	71%
			KNN	61%
		Non-real time	DT	54%
			CNN	93%
			SVM	82%
Our proposed method	1700 images	Real time	KNN	80%
			DT	75%
			Faster-R-CNN	88%

In Table 2, the accuracy values for the methods used in Unutmaz et al. (2019) were given for the session in which the training was performed with the %5 portion of the whole data set. Because, real time performances of those methods were also evaluated using the the same portion value for the training data set. Çelik and Odabaş (2020) used the manually separated hand gesture frames from videos for recognition using CNN model which was a non-real time process. The accuracy of the CNN model increased when they applied preprocessing to extract the hand objects from the frames. However, our method achieved %88 accuracy rate in real time without any preprocessing.

## 5. Conclusions

As a conclusion, we proposed a system that converts 23 letters in the Turkish Sign Language Alphabet and 3 letters (Q, W, X) in English Sign Language into text using image processing and deep learning methods. Our system which performs this process in real time used the Faster R-CNN architecture unlike other systems in the literature. In addition, the development of the system with Python increases its comprehensibility, while the use of open source libraries facilitates its accessibility. While the system takes the hand image that mimics the sign language alphabet as input, it shows the letter that looks like the hand image as output with its accuracy rate. When we look at performance measures such as Accuracy rate, F-Score and Loss value, our system seems to be successful. The developed system will facilitate communication between people who do not speak sign language and individuals with hearing disabilities. Thus, people with hearing impairments

will be prevented from isolating themselves from society and their self-confidence will be increased.

## Acknowledgment

The dataset used for training in this study is available at <https://www.kaggle.com/feronial/turkish-sign-languagefinger-spelling>.

## References

- Ashish S. N., Aarti G. A., 2016. Sign language recognition using image based hand gesture recognition techniques. 2016 Online International Conference on Green Engineering and Technologies (IC-GET), 19 November 2016, Coimbatore, India, pp.1-5.
- Çelik Ö., Odabaş A., 2020. Sign2Text: Konvolüsyonel Sınır Ağları Kullanarak Türk İşaret Dili Tanıma. European Journal of Science and Technology, 19, 923-934.
- Gaikwad, S., Shetty, A., Satam, A., Rathod, M., Shah, P., 2019. Recognition of American Sign Language using image processing and machine learning. International Journal of Computer Science and Mobile Computing, 8(3), 352-357.
- Kumarage, D., Fernando, S., Fernando, P., Madushanka, D., Samarasinghe, R., 2011. Real-time sign language gesture recognition using still-image comparison & motion recognition. 6th International Conference on Industrial and Information Systems, 10 October 2011, Kandy/Sri Lanka, pp.169-174.
- Pramada S., Saylee D., Pranita N., Samiksha N., Vaidya S., 2013. Intelligent sign language recognition using image processing. IOSR Journal of Engineering (IOSRJEN), 3(2), 45-51.
- Sandhya, A., Ananya R., 2018. Recognition of sign language using image processing. International Journal of Business Intelligence and Data Mining, 13(1-3),163-168.
- Shaoqing R., Kaiming H., Ross G., Jian S., 2017. Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks. IEEE Transactions on Pattern Analysis and Machine Intelligence, 39(6), 1137-1149.
- Shivashankara S., Srinath S., 2018. An American sign language recognition system using bounding box and palm features extraction techniques. International Journal of Recent Technology and Engineering (IJRTE), 7(4s), 492-505.

Tamura, S., Kawasaki, S., 1988. Recognition of sign language motion images. *Pattern Recognition*, 21(4), 343-357.

Tan Y.S., Lim K.M., Tee C., Lee C., 2020. Convolutional neural network with spatial pyramid pooling for hand gesture recognition. *Neural Computing and Applications*, Published online: 15 September 2020.

Unutmaz B., Karaca A.C., Güllü M.K., 2019. Turkish Sign Language Recognition Using Kinect Skeleton and Convolutional Neural Network. *27th IEEE Signal Processing and Communication Applications Conference*, April 2019, Sivas, Turkey, pp.1-4.

Vargas L.P., Barba L., Torres C.O., Mattos L., 2011. Sign Language Recognition System using Neural Network for Digital Hardware Implementation. *Journal of Physics: Conference Series*, 274 (2011) 012051, 1-7.



# A Tabu Search and Hybrid Evolutionary Strategies Algorithms for the Integrated Process Planning and Scheduling with Due-date Agreement

Halil İbrahim Demir <sup>1</sup> , Caner Erden <sup>2</sup> , Abdullah Hulusi Kökçam <sup>1\*</sup> , Alper Göksu <sup>1</sup> 

<sup>1</sup>Sakarya University, Faculty of Engineering, Department of Industrial Engineering, Sakarya/Turkey

<sup>2</sup>Sakarya Applied Science University, Faculty of Applied Science, Department of International Trade and Finance, Sakarya/Turkey

hidemir@sakarya.edu.tr, cerden@subu.edu.tr, akokcam@sakarya.edu.tr, agoksu@sakarya.edu.tr

## Abstract

There are three important functions of manufacturing which are process planning, scheduling and due date assignment. Instead of executing these functions separately, combining them together helps making more realistic and applicable plans. In this problem context random search, semi-tabu, random/semi-tabu hybrid, evolutionary strategies, and random/evolutionary strategies hybrid methods are utilized in solution. Different sized job shops are studied for performance tracking. As a result of study differences between individual solutions and integrated solutions was revealed. It is found that integrating these functions are advantageous in terms of total performance measure, and thus customer satisfaction. Random search is better than ordinary solutions. On the other hand, semi-tabu, evolutionary strategies and their hybrids are outperformed random search. Hybrid search methods are found promising.

**Keywords:** Process planning, weighted due-date assignment, weighted scheduling, random search, hybrid evolutionary strategies, hybrid semi-tabu search.

## Teslim Tarihi Anlaşmalı Entegre Süreç Planlama ve Çizelgeleme için Tabu Arama ve Hibrit Evrimsel Stratejiler Algoritmaları

### Öz

İmalatın, süreç planlama, çizelgeleme ve teslim tarihinin belirlenmesi olmak üzere üç önemli fonksiyonu bulunmaktadır. Bu fonksiyonları ayrı ayrı gerçekleştirmek yerine, üçünü birlikte değerlendirmek daha gerçekçi ve uygulanabilir planların hazırlanmasına yardımcı olacaktır. Bu problemin çözümünde rassal arama, kısmi tabu, rassal arama/kısmi tabu hibriti, evrimsel stratejiler ve rassal arama/evrimsel stratejiler hibriti yöntemleri kullanılmıştır. Yöntemlerin performansını değerlendirmek amacıyla farklı büyüklükteki atölyeler üzerinde çalışılmıştır. Çalışma sonucunda ayrı çözümler ile entegre çözümler arasındaki farklar ortaya çıkmıştır. Bu fonksiyonların entegrasyonunun toplam performans açısından ve dolayısıyla müşteri memnuniyeti açısından avantajlı olduğu bulunmuştur. Rassal aramanın sıradan çözümlerden daha iyi olduğu görülmüştür. Öte yandan, kısmi tabu, evrimsel stratejiler ve bunların melezleri rassal aramadan daha iyi performans göstermektedir. Hibrit arama yöntemlerinin gelecek vadettiği görülmüştür.

**Anahtar Kelimeler:** Süreç planlama, ağırlıklı teslim tarihi ataması, ağırlıklı çizelgeleme, rassal arama, hibrit evrimsel stratejiler, hibrit kısmi tabu arama.

## 1. Introduction

Process planning, scheduling and due date assignment are three essential manufacturing functions

in a job shop. Job shop scheduling is one of main production types and problem we studied should be considered seriously. Although classically these three functions are treated separately, lately numerous works are done on Integrated Process Planning and Scheduling

\* Corresponding Author.

E-mail: akokcam@sakarya.edu.tr

Received : 9 July 2020

Revision : 5 January 2021

Accepted : 11 January 2021

(IPPS) and Scheduling with Due-date Agreement (SWDDA). Even though there are numerous works on some level of integration, unfortunately there are only a few works on IPPSDDA (Integrated Process Planning, Scheduling and Due-date Assignment) which fully integrates these three functions.

Since these three functions highly affect each other, it is better to integrate them as much as possible. Output of these functions are inputs to the downstream functions. For example, output of process planning becomes input for scheduling. Poor process plans cause poor scheduling performance. If there is no integration, process planner may always select some desired machines repeatedly and they may not select some undesired machines. This causes unbalanced machine loading and reduce shop floor performance. If due dates are given independently, then it may be unrealistic. We may give far due dates unnecessarily and this increase weighted earliness and due date costs. On the other hand, if we give too close due dates then we may not keep our promise and weighted tardiness cost increases. At the literature due dates are given disregarding the importance of customers, but in this study weight of each customer is considered.

In the literature some works punished tardiness, some punished earliness and tardiness, some punished maximum absolute lateness, and some punished number of tardy jobs etc. But in this research, we penalized sum of weighted tardiness, earliness, and due date related costs. These terms are penalized because we wanted to give realistic due dates and we penalized weighted due dates to prevent unnecessary far due dates especially for important customers. Far due date means loss of customer goodwill, loss of customer or price reduction. Weighted Tardiness is penalized to prevent late delivery. Similarly, to due date related cost, tardiness means customer ill will, loss of customer, loss of good reputation and price reduction. Classically only tardiness is punished but in JIT environment and in reality, earliness is also problem. Stock holding, storage and spoilage costs can be earliness costs. So weighted earliness is also punished in this study.

Only scheduling function is NP-Hard class problem. So integrated problem is even harder to solve. As we mentioned earlier there are not much work on IPPSDDA. For the problems on IPPS and on SWDDA exact solutions are tried for very small problems. But for large problems only some good heuristics are advisable. In this research we applied evolutionary strategies, semi-tabu, random, hybrid evolutionary strategies and hybrid tabu search techniques are applied in the solution of the problem.

We represented problem as chromosomes with  $(n+2)$  genes where  $n$  represent number of the jobs. First two genes are used to represent due date assignment rule and dispatching rule sequentially. Remaining genes are used to represent currently selected route of each job. As we mentioned five search techniques are used and

compared with one another. Hybrid Evolutionary strategies is found best among all others.

Eight shop floors are tested, and characteristics of these shop floors are given at section 3.

We also tested different integration levels and tested the benefit of higher integration levels. Firstly, we tested unintegrated case where process plans are selected randomly, jobs are scheduled in random order and due dates are assigned randomly (externally). Later we integrated Weighted Minimum Slack (WMS) dispatching with process plan selection. After that we tested integration of Weighted Slack (WSLK) due date assignment with process plan selection. Finally, we integrated three functions and we tested integration of process planning with WMS dispatching and WSLK due date assignment. As we mentioned weighted due date assignment with weighted scheduling and process plan selection is not addressed at the literature but in this study, we tried to prove benefit of integration with weighted scheduling and weighted due date assignment.

Process planning is defined as the systematic determination of the methods by which a product is to be manufactured economically and competitively according to Society of Manufacturing Engineers.

Zhang and Mallur (1994) defined production scheduling as a resource allocator, that considers timing data while allocating resources to the tasks.

Pinedo and Chao (1998) defined the job shop-scheduling environment as;  $n$  jobs to be processed and  $m$  machines to process these jobs. Each job follows some predetermined routes, visiting a number of machines. Job shop problems occur primarily in industries where each customer order has specified characteristics and order sizes are moderately small.

Gordon et al. (2002) presented a good literature survey on SWDDA. According to Gordon et al. (2002) the scheduling problems involving due dates are of essential concern. In a conventional production environment, a job is expected to be completed before its due date. In a just-in-time environment, a job is required to be finished precisely at its due date.

If we look at more recent works, we see SWDWA (Scheduling with Due-window Assignment) problem became very popular in place of SWDDA problems. Here due window is tried to be assigned instead of due date. In this problem most suitable window with starting point and length is tried to be determined.

Development in hardware, software and algorithm makes it possible to solve new problems or to solve old problem easier which were hard to solve previously. After recent development in computer it is possible to develop process plans easier. CAPP (Computer Aided Process Planning) became possible and we can prepare process plan faster. Since process planning easier, we can prepare alternative process plans which help to balance shop floor and increase shop floor utilization.

Since we minimize weighted due date, earliness, and tardiness we should better schedule important customers earlier. If we give close due dates for important

customers at the beginning and schedule these customers earlier then there can be substantial improvements in performance measure and reduction in overall weighted cost. This is tried to be observed in this research.

It is important to solve problem in a reasonable amount of time otherwise solution would be practically useless. For this reason, we applied Evolutionary Strategies (ES), Semi-Tabu Search (ST), RS/ES hybrid, RS/ST hybrid and Random Search (RS) metaheuristics to find a good solution in an acceptable amount of time. We also compared search results with initial Ordinary Solutions (OS) solutions and proved searches are very useful. As expected, directed searches (ES, ST) and semi-directed searches (RS/ES, RS/ST) outperformed undirected (RS) search.

After representing problem as chromosome, we gave higher probability for first two genes to be selected for mutation operator. Because changes in these genes greatly affect solution compared to slow effect of a route of a single job. So, we applied dominant genes and found it useful.

We penalized weighted tardiness, earliness, and due dates. This penalty function is found to be realistic and very useful for IPPSDDA problem. These three terms in penalty function are all undesired. Tardiness is punished more according to earliness in terms of fixed and variable cost. These cost terms and penalty function are explained at section 3.

In short as integration level increased solution became better. Searches are found very useful and RS/ES outperformed all others. Best combination is observed where full integration with RS/ES is used. Using weighed due date assignment and weighted dispatching are found useful and dominant genes are used and this was also very helpful. In contrast to literature we penalized three terms which are weighted tardiness, earliness and due dates which is better and more realistic.

## 2. Background and literature survey

Although IPPS and SWDDA are both popular research topics, IPPSDDA are quite novel and only a few researches were made. If we look at the recent decade numerous works are conducted on IPPS. Traditionally three functions are applied sequentially and separately. Before going into detail, it is better to see recent surveys on IPPS. Tan and Khoshnevis (2000) and Phanden et al. (2011) prepared surveys on IPPS. For the SWDDA problem it is better to review survey of Gordon et al. (2002). For the IPPSDDA problem it is better to see Demir and Taskin (2005) and Ceven and Demir (2007). Demir and Taskin (2005) worked on IPPSDDA problem in a Ph.D. Thesis. Later Ceven and Demir (2007) worked on benefit of integrating due date assignment with IPPS problem in an M.S. Thesis.

If more literature is to be listed on IPPS problem we can give the following earlier works on this problem;

Wilhelm and Shin (1985), Khoshnevis and Chen (1991), Zhang and Mallur (1994), Usher and Fernandes (1996), Brandimarte (1999), Weintraub et al. (1999), Morad and Zalzal (1999), Gindy et al. (1999).

If we integrate process plan selection with scheduling, then we should determine alternative process plans and number of alternative process plans wisely. Corti and Portioli-Staudacher (2004) studied alternative process plans availability and their effect on manufacturing system performance.

It is difficult to select best plans if there are multiple process plans. Ming and Mak (2000) studied process plan selection problem by using a hybrid Hopfield network-genetic algorithm. Bhaskaran (1990) studied process plan selection in his study.

Developments in hardware, software and algorithms provide us to solve the problems which could not be solved earlier, or we can solve the problems easier. Recent developments provide CAPP (Computer aided process planning). Usher and Fernandes (1996) and Aldakhlallah and Ramesh (1999) studied integration of process planning with CAPP.

IPPS problem is an NP-hard problem and researchers commonly uses metaheuristics such as genetic or evolutionary algorithms. Morad and Zalzal (1999), Moon et al. (2002), Kim et al. (2003), Drstvenšek and Balič (2003), Moon et al. (2008), Seker et al. (2013), Zhang and Wong (2015) are worked on this problem. Following works are relatively more recent works on IPPS; Ming and Mak (2000), Tan and Khoshnevis (2000), Thomalla (2001), Kim et al. (2003), Usher (2003), Drstvenšek and Balič (2003), Corti and Portioli-Staudacher (2004), Shrestha et al. (2008), Moon et al. (2008), Özgüven et al. (2010), Leung et al. (2010), Phanden et al. (2011), Li et al. (2012), Seker et al. (2013), Zhang and Wong (2015), Petrović et al. (2016), and Zhang et al. (2016).

As in IPPS there are hundreds of works on SWDDA. As it is mentioned earlier before going into detail of SWDDA it is better to see Gordon et al. (2002) as a state of the art survey on scheduling with common due date assignment.

Classically tardiness is tried to be penalized but according to JIT (Just in Time) philosophy we should penalize both earliness and tardiness. Since nobody prefer long due dates in this study all of weighted earliness, tardiness and due date related costs are tried to be minimized.

Due dates can be determined internally or externally. At the latter case we try to catch best performance according to externally dictated due dates. But at the former case we try to assign best due dates for the jobs that serve most to the benefit of the firm. For this reason, SWDDA problem gained importance as research topic and many works are conducted on this problem. If we look at the earlier works on SWDDA problem we can see the following works; Luss and Rosenwein (1993), Yang et al. (1994), Lawrence (1994), Cai et al. (1997),

Kovalyov (1997), Gordon and Kubiak (1998), Cheng and Kovalyov (1999), Gordon and Strusevich (1999).

Some of the recent works on SWDDA can be given as follows; Biskup and Jahnke (2001), Mosheiov (2001), Gordon et al. (2002), Birman and Mosheiov (2004), Lauff and Werner (2004), Baykasoğlu et al. (2008), Gordon and Strusevich (2009), Allaoua and Osmane (2010), Tuong and Soukhal (2010), Li et al. (2011), Vinod and Sridharan (2011), Shabtay (2016), and Koulamas (2017).

We can assign common and separate due dates for every job. If we solve job shop scheduling problem, we may assign separate due dates but in case of assembly and simultaneous delivery cases we may assign common due dates for the jobs to be scheduled. Many works in the literature are on scheduling with common due date assignment (SWCDDA) such as Chen et al. (1997), Kovalyov (1997), Biskup and Jahnke (2001), Mosheiov (2001), Gordon et al. (2002), Gordon and Strusevich (2009), Allaoua and Osmane (2010), Tuong and Soukhal (2010), and Li et al. (2011).

Unlike SWCDDA some works are on SWSDDA (Scheduling with separate due date assignment) such as Gordon and Kubiak (1998), Cheng and Kovalyov (1999), Gordon and Strusevich (1999), Baykasoğlu et al. (2008), Gordon and Strusevich (2009), Li et al. (2011), Vinod and Sridharan (2011). In this study every job gets its own due date.

Some of these studies have single, some double, and some have multiple machines. As an example to SMSWDDA (Single machine scheduling with due date assignment) following works can be given: Cai et al. (1997), Kovalyov (1997), Gordon and Strusevich (1999), Gordon et al. (2002), Gordon and Strusevich (2009), Allaoua and Osmane (2010), Tuong and Soukhal (2010), Li et al. (2011).

Birman and Mosheiov (2004) studied two machine flow shop scheduling with due date determination (TMFSWDDA).

Following works are on PMSWDDA (Parallel machine scheduling with due date assignment); Cheng and Kovalyov (1999), Gordon et al. (2002), and Tuong and Soukhal (2010).

Following works are on MMSWDDA (Multi machine scheduling with due date assignment); Luss and Rosenwein (1993), Lawrence (1994), and Lauff and Werner (2004).

Some works are on JSSWDDA (Job shop scheduling with due date assignment) such as Yang et al. (1994), Baykasoğlu et al. (2008), and Vinod and Sridharan (2011). In this study jobs are tried to be assigned

separate due dates and every shop floor tested as a case of job shop with different sizes.

One of the current study areas is the dynamic scheduling problem. There are a limited number of studies in the literature on the dynamic integrated process planning, scheduling, and due date assignment (DIPPSDDA). Demir and Erden (2020) tried to optimize the dynamic environment with ant colony algorithm in their work. DIPPSDDA, which was previously improved with the genetic algorithm and some meta-heuristic algorithms (Erden et al., 2019), was a new field of study.

### 3. Problem definition

As we mentioned earlier IPPS and SWDDA problems are studied extensively. In this research we studied IPPSDDA problem. We have three functions to be integrated which are process planning, scheduling, and due date assignment. Step by step we integrated these functions with each other and tried to see benefit of integration level. Problem is represented as chromosomes which have  $(n+2)$  genes where  $n$  is the number of jobs. First two genes are used to represent due date assignment and dispatching rule genes. There are two different due date assignments and four different dispatching rules. Mainly, we used WSLK weighted due date assignment rule and RDM (Random) due date assignment rule. At dispatching gene, we used WMS and SIRO (Service in Random Order) rules, so we have two dispatching rules to choose from.

We studied eight different shop floors with varying size. There are five different routes to choose from for each job in smaller shop floors. At the smallest shop floor, we have 5 machines, 25 jobs and 5 routes. There are 10 operations in each route. Processing time of each operation practically changes in between 1 and 30 minutes according to formula  $[(12 + z * 6)]$  where  $z$  is the standard normal numbers in which  $\sigma = 6$  and  $\mu = 12$ .

At the largest shop floor, we have 40 machines, 200 jobs, 3 routes and there are 10 operations in each route. Processing times are same as in other shop floors. We took 3 alternative routes in larger shop floors to find a solution in a reasonable amount of time. Characteristics of each shop floors are listed at Table 1.

**Table 1.** Shop floors

Shop Floor	1	2	3	4	5	6	7	8
# of machines	5	10	15	20	25	30	35	40
# of Jobs	25	50	75	100	125	150	175	200
# of Routes	5	5	5	5	3	3	3	3
Processing Times	[(12 + z * 6)]							
# of op. per job	10							

We started from unintegrated version of the problem where process plan selection is made independently, due dates are determined randomly (externally) and jobs are dispatched randomly. Later WMS rule is integrated with process plan selection, but due dates are still determined randomly. After that we integrated WSLK due date assignment with process plan selection, but jobs are scheduled according to SIRO rule.

At the end we integrated all of the three functions. Process plan selection is performed with WMS dispatching and WSLK due date assignment. We found this case as the best. General flow diagram is given in Figure 1.

We assumed a working day as one shift which is 8 hours or 480 minutes. As a performance measure we tried to minimize weighted tardiness, earliness and due date related costs. We penalized these terms proportional with the weights of the customers and proportional to the tardiness, earliness and due dates

multiplied with different constants. For tardiness and earliness, we also used fixed cost if there is tardiness or earliness. Penalty function for each term is given below where PD is penalty for due-date (Equation 1), PE is penalty for earliness (Equation 2), PT is penalty for tardiness (Equation 3), Penalty of a job is Penalty(j) (Equation 4) and Total penalty (Equation 5).

$$PD = weight_j * 8 * (DueDate / 480) \quad (1)$$

$$PE = weight(j) * (5 + 4 * (E / 480)) \quad (2)$$

$$PT = weight(j) * (10 + 12 * (T / 480)) \quad (3)$$

$$Pen(j) = PD + PE + PT \quad (4)$$

$$Total\ Penalty = jPen(j) \quad (5)$$

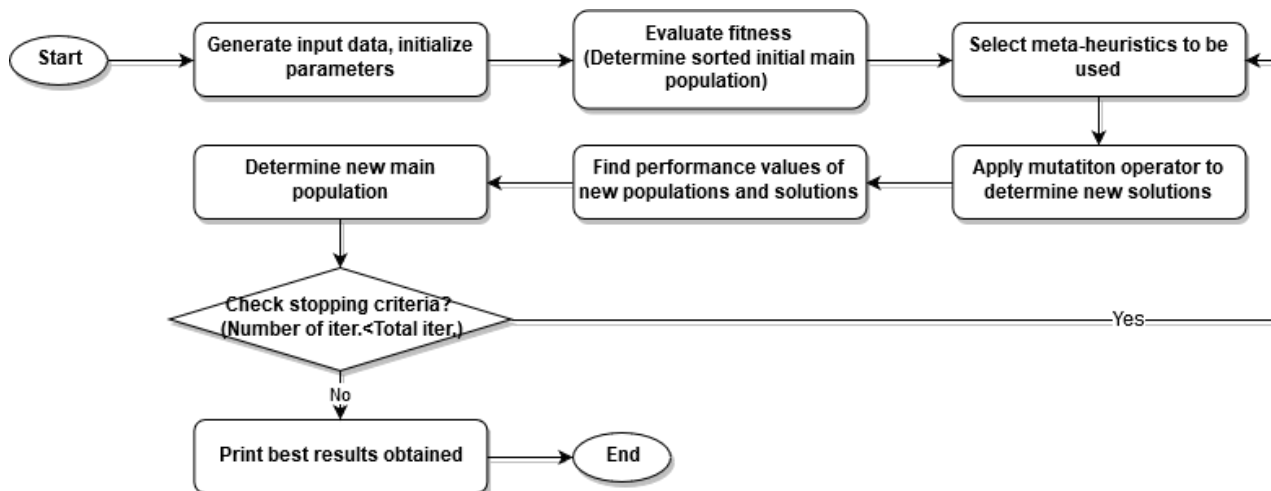


Figure 1. General flow diagram

#### 4. Solution techniques

We used initially randomly produced chromosome as the ordinary solution and compared this result with the results of evolutionary strategies, semi-tabu search, hybrid searches and random search results.

Scheduling problem alone belongs to NP-Hard problem and integrated problem is even harder to solve that is why some good heuristics are required to solve this problem in a reasonable amount of time. We applied pure, hybrid and random searches because of the characteristics of the problem. Each solution type is explained below.

**Ordinary Solution (OS):** OS is the initially randomly produced chromosome which represent any random chromosome possible at the beginning and as expected it is the poorest solution compared to the pure, hybrid and random search metaheuristics.

**Random Search (RS):** Here only RS is applied, and always brand-new solutions are produced randomly. To

be fair with other search techniques same number of chromosomes are produced in each iteration as in other pure and hybrid metaheuristics. We applied 200, 150, 100 and 50 random iterations for eight shop floors in doubles, respectively. At each iteration we produced 10 new random chromosomes. Later we took best 10 chromosomes from previous main population and newly produced 10 chromosomes. Random search is found very useful compared to ordinary solutions and marginal improvements are very high at the very beginning of the search, but marginal improvement reduces sharply as iteration goes on. That is why random search is poor compared to the other metaheuristics. Hybrid searches are powerful with the combination of high marginal improvement of RS at the beginning and good search characteristics of the directed searches later on.

**Evolutionary Strategies (ES):** During 1960s Rechenberg (1965), and Schwefel (1981) two students of Technical University of Berlin, Germany, developed ES to solve optimization problem. Main difference



between ES and Genetic Algorithm (GA) is the operators used. While ES uses only mutation operator, GA uses both mutation and crossover operators. We

produce ten new chromosomes by applying mutation operator in each iteration. ES flow diagram is given in Figure 2.

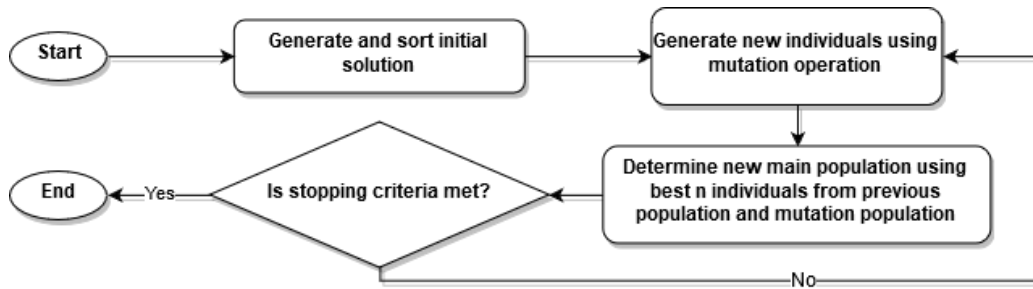


Figure 2. Evolutionary strategies flow diagram

**Semi-Tabu Search (ST):** For a shop floor which consists of 200 jobs, there are 202 genes and totally  $4 \times 2 \times 3200 = 2.1249119 \times 10^6$  combinations are possible. If we list every gene in tabu list, then problem becomes too complex to consider. Instead, we list only dominant genes in tabu list and apply mutation operator for remaining genes as given in Figure 3. Unlike ES we work on a single chromosome in every iteration, so we applied 10 times more iterations in this metaheuristic to be fair in comparison with ES.

- Terms used in Figure 3 is defined as follows:
- $S_n$  :  $n^{\text{th}}$  iteration solution
  - $S_{\text{candidate}}$  : candidate solution
  - $S_{\text{opt}}$  : optimum solution
  - $N$  : total iteration number (stopping criterion)
  - $n$  : current iteration number
  - $m$  : current trial number in tabu list
  - $M$  : Number of allowed trials in tabu list

**Hybrid Evolutionary Strategies (RS/ES):** This is a hybrid metaheuristic and initially at the first 5% of the iterations RS is applied and ES is applied to remaining 95% of the iterations. If a random number is generated between 0 and 1000 then expected value of the minimum is 500. If we generate two random numbers and take the minimum, then we get expected value of 330. If we generate three random numbers and take the minimum of these the expected value, we get is 250. Now if we look at marginal improvements, we get 500, 170 and 80, respectively. So, it is obvious that initial random iterations are very useful but marginal improvements reduces sharply. So here we started with RS and later we continued with ES. Since marginal benefits reduce sharply RS rate should be low and we applied 5% RS iterations.

**Hybrid Semi-Tabu Search (RS/ST):** Here again initially 5% RS is applied and later 95% ST search is applied. We used dominant genes in ES, ST, RS/ES and RS/ST while applying mutation operator. First two genes are dominant and that is why they had more probability to be selected for mutation. Using dominant genes improved efficiency of solution technique. While running program we recorded CPU times required. These times are listed at the Table 4 given at section 6.

As mentioned earlier, problem is represented as a chromosome with  $n+2$  genes where  $n$  is the number of jobs. First two genes are used to represent due date assignment and dispatching rules, respectively. Remaining genes represent active selected route of each job. A sample chromosome is given at Figure 4.

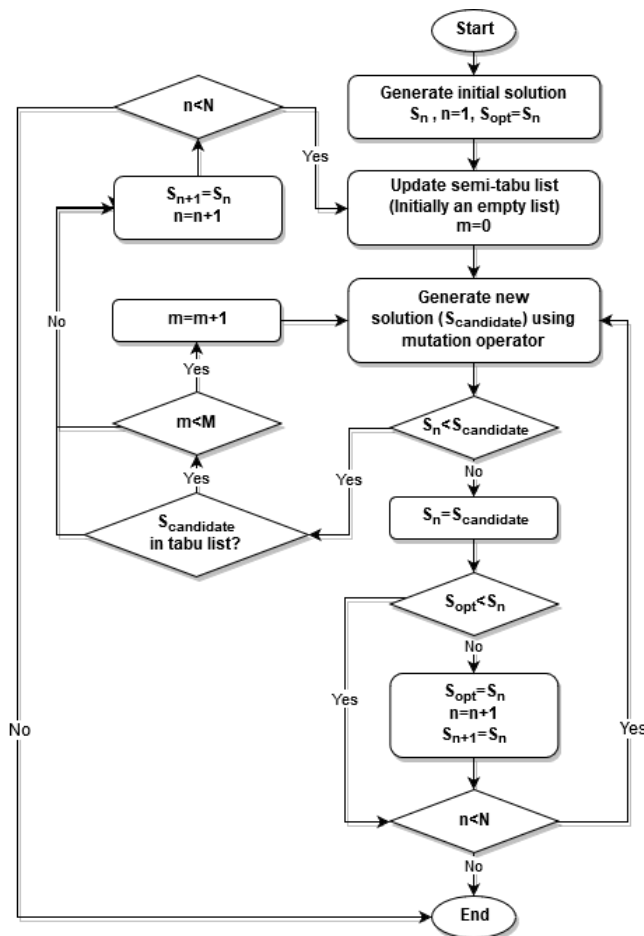


Figure 3. Semi-tabu search flow diagram

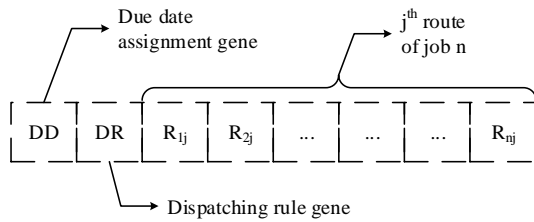


Figure 4. Sample chromosome

## 5. Experimentation

Mainly two types of due date assignment rules are used. With different constants first gene takes one of four different values. Mainly WSLK and RDM due date assignment rules are used. At the first rule, which is WSLK, some constant added to total processing time of each job according to weight of that job. At the RDM due date assignment, due dates are determined randomly as explained at Appendix 1. Rules are listed at Table 2. At the second gene we used two dispatching rules which are WMS and SIRO. These rules are listed at Table 3 and explained at Appendix.

Table 2. Due-date assignment rules

Method	Constant $q_x$	Rule no
WSLK	$q_x = q_1, q_2, q_3$	1,2,3
RDM		4

Table 3. Dispatching rules

Method	Rule No
WMS	1
SIRO	2

We used a desktop computer with 3.1 GHz Intel i5-2400 processor and 4 GB ram with 64-bit Windows 10 operating system to run the program. We used Borland C++ 5.02 compiler. CPU Times are listed at Table 4. Screenshot of the program is given in Figure 5 and Figure 6.

```
F:\2-1-AHK-2-2-WMS-WSLK-REVISE\1-25-5-5>1-4-WMS-WSLK-es-mix 2048
Random number generator seed= 2048
Population (10 chromosomes) :
PERF DD DR J0---->Rest of the jobs
Row 0 : 294.91 3 6 0 0 1 4 2 3 2 0 0 2 4 2 4 3 2 3 4 2 3 2 1 0 1 0 3
Row 1 : 301.77 3 6 0 2 0 1 0 0 2 0 4 2 3 3 1 1 2 1 3 3 1 0 0 1 1 2 4
Row 2 : 303.43 3 6 1 3 2 2 2 2 0 0 0 3 0 4 3 1 2 1 2 4 1 4 2 0 0 0 0
Row 3 : 304.75 3 6 3 3 4 2 3 3 1 4 3 1 2 0 1 0 3 3 0 3 2 3 0 1 1 0 0
Row 4 : 311.14 4 6 1 0 1 4 4 2 4 4 3 0 2 1 0 1 4 0 1 2 0 4 3 4 4 2 2
Row 5 : 330.80 4 6 4 3 3 1 1 4 3 4 4 2 1 2 0 4 0 3 2 3 2 4 3 1 0 2 0
Row 6 : 341.78 5 6 1 2 2 0 4 1 4 2 2 0 4 3 3 2 3 1 2 0 4 4 0 4 1 2 2
Row 7 : 344.35 5 6 0 2 2 3 1 0 2 0 0 2 0 1 1 4 2 0 3 4 2 1 0 2 1 3 0
Row 8 : 355.04 5 6 3 3 3 2 3 2 1 1 4 3 1 0 0 2 3 4 4 2 1 3 4 4 2 0 2
Row 9 : 362.65 5 6 3 2 2 1 3 3 4 3 0 3 1 4 1 1 3 2 0 4 1 2 3 0 3 3 4
```

Figure 5. Screenshot of the program start

We tested eight shop floors and certain number of iterations are applied for each shop floor. To be fair, same number of iterations are applied for both evolutionary search and random search. Since at the ES and RS, we produce 10 chromosomes and in ST we produce only one new chromosome thus we executed 10 times more iterations compared to the ES and RS. 200 iterations are applied for the smallest shop floors which

are SF1 (Shop Floor 1) and SF2. For SF3 and SF4 we applied 150 iterations. For SF5 and SF6 we used 100 iterations and finally for the largest two shop floors which are SF7 and SF8 we applied 50 iterations.

```
Iteration 200
PERF DD DR J0---->Rest of the jobs
Row 0 : 265.10 3 6 0 1 3 1 0 3 3 4 0 1 0 0 1 3 4 0 2 1 3 4 4 4 3 2 3
Row 1 : 266.53 3 6 0 1 1 1 0 0 4 1 0 4 0 0 1 3 1 1 2 2 3 0 1 4 2 2 3
Row 2 : 266.54 3 6 0 1 3 1 0 3 3 0 0 1 0 4 1 3 1 0 2 1 3 4 4 4 3 2 3
Row 3 : 266.57 3 6 2 1 1 1 0 2 4 1 0 4 0 2 1 2 0 4 0 0 3 0 1 4 1 4 3
Row 4 : 266.99 3 6 0 1 3 1 0 3 3 4 0 1 0 0 1 1 4 0 2 1 3 3 4 4 3 4 3
Row 5 : 267.36 3 6 2 1 2 0 0 2 4 1 0 0 0 3 1 3 2 0 2 0 3 0 1 4 2 4 2
Row 6 : 267.63 3 6 2 1 1 0 0 2 4 1 0 4 0 0 1 0 0 0 2 1 3 2 1 4 2 4 2
Row 7 : 267.68 3 6 2 1 1 1 0 2 4 1 0 4 0 2 1 2 0 0 2 0 3 0 1 4 2 4 3
Row 8 : 267.74 3 6 2 1 2 0 0 2 4 1 0 4 0 3 1 2 0 0 2 0 3 0 1 4 2 4 2
Row 9 : 267.84 3 6 2 1 1 3 0 2 4 1 0 0 0 3 1 2 3 0 2 1 3 0 1 4 2 4 2
```

Best: 265.10 Average: 267.00 Worst: 267.84  
200 iterations are applied before stopping.

Results obtained in last iteration

```
Best performance : 265.10
Best due-date assignment rule : 3
Best dispatching rule : 6

Best route for job 0 : 0
Best route for job 1 : 1
Best route for job 2 : 3
Best route for job 3 : 1
Best route for job 4 : 0
Best route for job 5 : 3
Best route for job 6 : 3
Best route for job 7 : 4
Best route for job 8 : 0
Best route for job 9 : 1
Best route for job 10 : 0
Best route for job 11 : 0
Best route for job 12 : 1
Best route for job 13 : 3
Best route for job 14 : 4
Best route for job 15 : 0
Best route for job 16 : 2
Best route for job 17 : 1
Best route for job 18 : 3
Best route for job 19 : 4
Best route for job 20 : 4
Best route for job 21 : 4
Best route for job 22 : 3
Best route for job 23 : 2
Best route for job 24 : 3
```

Cpu time : 18.88 sec  
Program execution time : 19 sec

Figure 6. Screenshot of the program results

If we look at the approximate CPU times for every shop floor from the tables mentioned above; for the small shop floor, it took approximately between 16 and 88 seconds CPU time. For the small-medium shop floor, it took approximately between 173 and 322 seconds. For the medium-large shop floor, it took approximately between 257 and 376 seconds and for the largest shop floor, it took between 250 and 346 seconds. CPU time of the largest shop floor is lower than the medium-large shop floor due to its iteration number.

SIRO-RDM (OS, RS, ES, ST, RS/ES, RS/ST), WMS-RDM (OS, RS, ES, ST, RS/ES, RS/ST), SIRO-WSLK (OS, RS, ES, ST, RS/ES, RS/ST), and WMS-WSLK (OS, RS, ES, ST, RS/ES, RS/ST) are the twenty-four solutions, we compared in this study. Initially, we tested SIRO-RDM combinations. Here three functions are unintegrated. Later we tested WMS-RDM combinations in which WMS scheduling is integrated with process plan selection and due-dates are determined randomly. After that, we tested SIRO-WSLK combinations. Now WSLK due-date assignment is integrated with process planning but now jobs are scheduled in random order. Later we tested all combinations at the fully integrated level which are

WMS-WSLK combinations. According to results, full integration with RS/ES found as the best solution. Searches are found always better compared to ordinary solutions and RS. Obtained results are explained in the conclusion section.

We tested eight shop floors for twenty-four different solutions. Shop floor characteristics and twenty-four types of solutions are explained in the previous sections of the study. At every shop floor, operation times are determined randomly according to the formula  $[(12 + z * 6)]$  where z is the standard normal numbers.

In first small shop floor (SF-1) there are 25 jobs and 5 machines. For this shop floor, we applied 200 iterations for ES and RS, 2000 iterations for ST which took 16 seconds on average. Smallest shop floor results are given in **Hata! Başvuru kaynağı bulunamadı.**

In second small shop floor (SF-2) there are 50 jobs and 10 machines. For this shop floor, we applied 200 iterations for ES and RS, 2000 iterations for ST which took 88 seconds on average. Second smallest shop floor results are given in **Hata! Başvuru kaynağı bulunamadı.**

In first small-medium shop floor (SF-3) there are 75 jobs 15 machines. For this shop floor, we applied 150 iterations for ES and RS, 1500 iterations for ST which took 173 seconds on average. First small-medium shop floor results are given in **Hata! Başvuru kaynağı bulunamadı.**

In second small-medium shop floor (SF-4) there are 100 jobs 20 machines. For this shop floor, we applied 150 iterations for ES and RS, 1500 iterations for ST which took 321 seconds on average. Second small-medium shop floor results are given in **Hata! Başvuru kaynağı bulunamadı.**

In first medium-large shop floor (SF-5) there are 125 jobs 25 machines. For this shop floor, we applied 100 iterations for ES and RS, 1000 iterations for ST which took 256 seconds on average. First medium-large shop floor results are given in Figure 11.

In second medium-large shop floor (SF-6) there are 150 jobs 30 machines. For this shop floor, we applied 100 iterations for ES and RS, 1000 iterations for ST which took 376 seconds on average. Second medium-large shop floor results are given in Figure 12.

In first largest shop floor (SF-7) we have 175 jobs to be scheduled and 35 machines on this shop floor. We applied 50 iterations for ES and RS, 500 iterations for ST which took 250 seconds on average. First largest shop floor results are given in Figure 13.

For the largest shop floor (SF-8) we have 200 jobs to be scheduled and 40 machines on this shop floor. We applied 50 iterations for ES and RS, 500 iterations for ST which took 346 seconds on average. Second largest shop floor results are given in Figure 14.

According to results, we found similar conclusions. Higher integration gave better results and full integration was the best. Searches are found superior to ordinary and RS solutions. Full integration with RS/ES gave the best results. Comparisons of twenty-four

solution combinations for all the shop floors are given in Table 4. Best solution found in a level is indicated with bold text, for each shop floor. Best of all levels (24 solutions) are indicated with green bold text, for each shop floor.

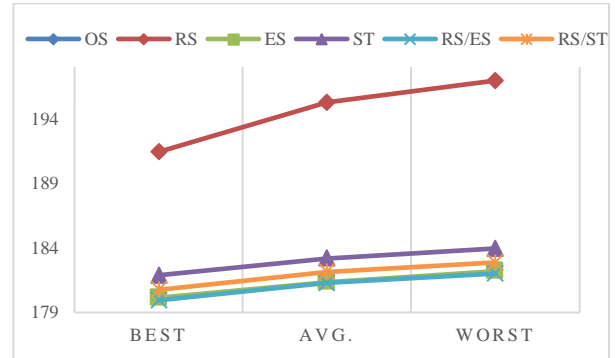


Figure 7. Results of Shop Floor 1 (25x5x5)



Figure 8. Results of Shop Floor 2 (50x10x5)



Figure 9. Results of Shop Floor 3 (75x15x5)

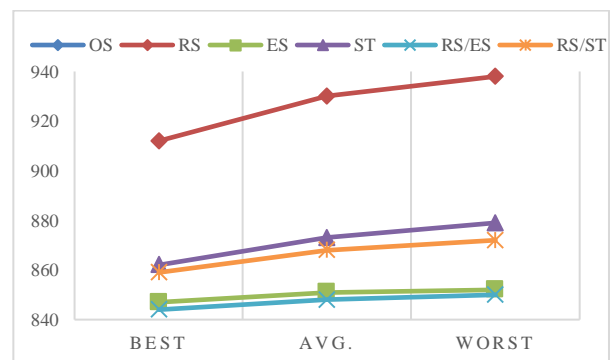


Figure 10. Results of Shop Floor 4 (100x20x5)

**Table 4** Comparison of twenty-four solution combinations for all of the shop floors

Level	Approach	Shop Floor 1				Shop Floor 2				Shop Floor 3				Shop Floor 4						
		Best	Avg.	Worst	CPU	Best	Avg.	Worst	CPU	Best	Avg.	Worst	CPU	Best	Avg.	Worst	CPU			
SIRO-RDM	OS	319	319	319	-	OS	646	646	646	-	OS	983	983	983	-	OS	1309	1309	1309	-
	RS	265	272	276	17	RS	586	598	606	88	RS	853	895	908	179	RS	1269	1278	1285	331
	ES	246	252	254	15	ES	546	552	556	84	ES	833	842	846	176	ES	1197	1204	1210	311
	ST	252	255	256	14	ST	554	562	565	79	ST	<b>829</b>	841	848	162	ST	1234	1246	1254	299
	RS/ES	<b>240</b>	248	251	15	RS/ES	<b>545</b>	548	551	84	RS/ES	831	836	842	174	RS/ES	<b>1207</b>	1216	1222	312
RS/ST	267	268	270	15	RS/ST	557	570	574	82	RS/ST	847	862	868	167	RS/ST	1219	1230	1238	297	
WMS-RDM	OS	270	270	270	-	OS	564	564	564	-	OS	817	817	817	-	OS	1184	1184	1184	-
	RS	230	236	239	15	RS	520	528	532	88	RS	770	782	787	182	RS	1121	1139	1150	333
	ES	201	205	208	14	ES	<b>459</b>	465	469	83	ES	<b>700</b>	706	713	172	ES	<b>1058</b>	1069	1075	306
	ST	208	213	216	13	ST	494	501	505	79	ST	738	745	748	162	ST	1092	1099	1104	286
	RS/ES	<b>195</b>	201	203	16	RS/ES	467	475	481	82	RS/ES	729	734	735	171	RS/ES	1061	1068	1073	306
RS/ST	196	205	210	13	RS/ST	485	489	493	80	RS/ST	731	739	744	163	RS/ST	1088	1096	1101	294	
SIRO-WSLK	OS	314	314	314	-	OS	666	666	666	-	OS	997	997	997	-	OS	1372	1372	1372	-
	RS	263	272	279	16	RS	571	588	596	94	RS	861	883	893	176	RS	1204	1222	1232	343
	ES	<b>248</b>	255	257	16	ES	530	548	551	95	ES	812	824	831	170	ES	<b>1130</b>	1153	1160	331
	ST	<b>248</b>	257	263	15	ST	544	557	563	87	ST	854	861	867	161	ST	1186	1200	1206	313
	RS/ES	249	256	259	16	RS/ES	<b>529</b>	538	545	86	RS/ES	<b>804</b>	817	824	174	RS/ES	1138	1167	1178	333
RS/ST	261	265	268	15	RS/ST	541	549	553	86	RS/ST	833	847	850	162	RS/ST	1187	1195	1199	318	
WMS-WSLK	OS	270	270	270	-	OS	533	533	533	-	OS	783	783	783	-	OS	1142	1142	1142	-
	RS	191	195	197	17	RS	393	399	403	97	RS	593	600	606	191	RS	912	930	938	368
	ES	<b>180</b>	181	182	17	ES	372	374	375	99	ES	566	568	570	185	ES	847	851	852	347
	ST	182	183	184	17	ST	381	382	384	94	ST	576	579	581	176	ST	862	873	879	333
	RS/ES	<b>180</b>	181	182	17	RS/ES	<b>371</b>	372	374	98	RS/ES	<b>564</b>	566	568	185	RS/ES	<b>844</b>	848	850	345
RS/ST	181	182	183	17	RS/ST	378	380	381	90	RS/ST	577	580	582	177	RS/ST	859	868	872	332	
SIRO-RDM	Shop Floor 5				Shop Floor 6				Shop Floor 7				Shop Floor 8							
	OS	1831	1831	1831	-	OS	2110	2110	2110	-	OS	2154	2154	2154	-	OS	2783	2783	2783	-
	RS	1590	1632	1644	260	RS	1914	1930	1944	403	RS	2108	2154	2171	256	RS	2659	2707	2719	357
	ES	1556	1569	1577	251	ES	<b>1840</b>	1858	1867	386	ES	2067	2092	2101	248	ES	<b>2585</b>	2600	2606	341
	ST	1594	1608	1616	230	ST	1863	1875	1883	335	ST	2091	2103	2114	220	ST	2640	2654	2661	308
RS/ES	<b>1538</b>	1558	1566	249	RS/ES	1860	1877	1884	391	RS/ES	<b>2060</b>	2085	2092	249	RS/ES	2624	2628	2635	344	
RS/ST	1555	1574	1581	226	RS/ST	1863	1881	1887	338	RS/ST	2112	2121	2128	223	RS/ST	2594	2613	2626	304	
WMS-RDM	OS	1558	1558	1558	-	OS	1836	1836	1836	-	OS	1947	1947	1947	-	OS	2470	2470	2470	-
	RS	1402	1429	1441	268	RS	1708	1716	1723	393	RS	1915	1926	1938	262	RS	2412	2423	2431	368
	ES	1372	1377	1381	249	ES	<b>1633</b>	1636	1639	365	ES	1886	1896	1902	256	ES	2335	2346	2351	348
	ST	1374	1380	1386	233	ST	1662	1669	1675	342	ST	<b>1858</b>	1885	1892	224	ST	2344	2356	2362	315
	RS/ES	1371	1376	1378	254	RS/ES	1647	1652	1656	369	RS/ES	1867	1872	1877	249	RS/ES	<b>2332</b>	2343	2349	348
RS/ST	<b>1366</b>	1375	1379	231	RS/ST	1653	1663	1668	338	RS/ST	1900	1904	1907	227	RS/ST	2354	2362	2371	316	
SIRO-WSLK	OS	1921	1921	1921	-	OS	2196	2196	2196	-	OS	2252	2252	2252	-	OS	2840	2840	2840	-
	RS	1577	1603	1615	270	RS	1843	1851	1857	398	RS	2025	2066	2080	264	RS	2560	2589	2607	369
	ES	<b>1531</b>	1544	1552	264	ES	1797	1809	1814	390	ES	1984	1999	2011	260	ES	<b>2441</b>	2461	2473	353
	ST	1546	1566	1576	241	ST	1829	1849	1860	350	ST	2031	2045	2059	231	ST	2481	2499	2512	318
	RS/ES	<b>1531</b>	1540	1547	265	RS/ES	<b>1792</b>	1803	1812	387	RS/ES	1988	2007	2015	261	RS/ES	2477	2486	2491	358
RS/ST	1556	1578	1588	239	RS/ST	1845	1854	1860	346	RS/ST	<b>1982</b>	2009	2026	231	RS/ST	2473	2519	2538	322	
WMS-WSLK	OS	1538	1538	1538	-	OS	1801	1801	1801	-	OS	1791	1791	1791	-	OS	2348	2348	2348	-
	RS	1036	1083	1107	301	RS	1329	1359	1370	422	RS	1486	1497	1503	284	RS	1862	1911	1928	395
	ES	994	1001	1004	279	ES	1282	1286	1288	407	ES	<b>1457</b>	1462	1465	279	ES	1801	1812	1818	384
	ST	999	1017	1025	269	ST	1209	1216	1221	375	ST	1458	1467	1472	251	ST	1819	1835	1844	346
	RS/ES	<b>990</b>	998	1002	280	RS/ES	1274	1279	1282	406	RS/ES	<b>1457</b>	1462	1464	278	RS/ES	<b>1786</b>	1800	1807	381
RS/ST	1006	1014	1021	269	RS/ST	<b>1206</b>	1215	1221	375	RS/ST	1460	1468	1473	251	RS/ST	1802	1823	1835	344	



Figure 11. Results of Shop Floor 5 (125x25x3)

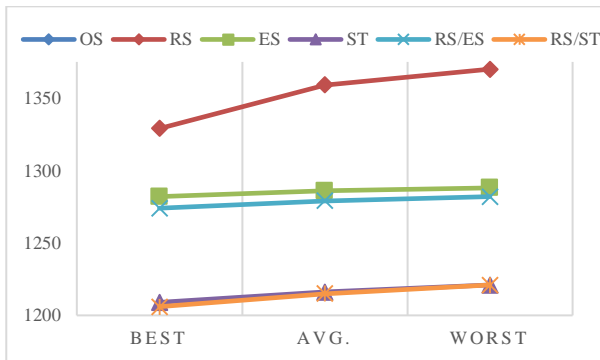


Figure 12. Results of Shop Floor 6 (150x30x3)

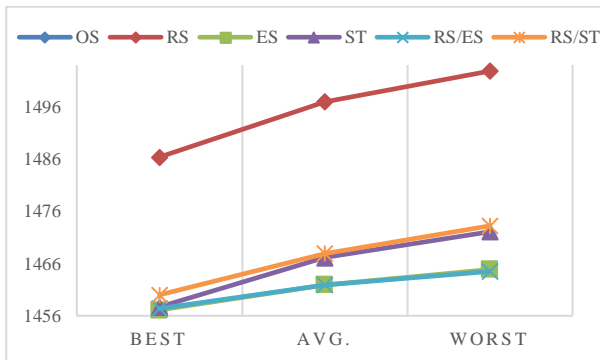


Figure 13. Results of Shop Floor 7 (175x35x3)

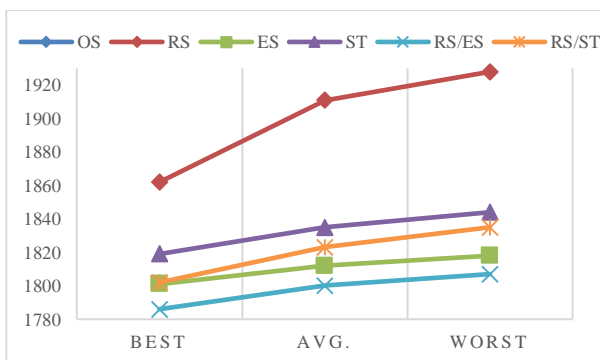


Figure 14. Results of Shop Floor 8 (200x40x3)

For each SIRO-RDM combination, a single experiment was performed, and RS/ES 4 (50%), ES 3 (37.5%), and ST 1 (12.5%) times gave the best results. For each WMS-RDM combination, a single experiment was performed and ES 4 (50%), RS/ES 2 (25%), RS/ST

1 (12.5%), and ST 1 (12.5%) times gave the best results. For each SIRO-WSLK combination, a single experiment was performed, and ES 4 (50%), RS/ES 3 (37.5%), and RS/ST 1 (12.5%) times gave the best results. For the WMS-WSLK combination, 5 experiments were conducted on each shop floor and totally 40 experiments were performed on 8 shop floors. RS/ES 19 (47.5%), ES 11 (27.5%), RS/ST 6 (15%), and ST 4 (10%) times gave the best results in these experiments.

When the best value of all combinations was considered, RS/ES 28 (43.75%), ES 22 (34.38%), RS/ST 8 (12.5%), and ST 6 (9.38%) times gave the best results in a total of 64 experiments. When the average value of all combinations was considered, RS/ES 34 (53.13%), ES 20 (31.25%), RS/ST 7 (10.94%), and ST 3 (4.69%) times gave the best results in a total of 64 experiments.

## 7. Conclusion

Production process takes place upon three functions, which are production planning, scheduling, and due-date assignment. Conventionally these three functions are executed separately in practice. On the other hand, these functions affect each other significantly as they are tightly connected with each other. These functions should be considered simultaneously to prepare more accurate production plans, schedules, and due-date assignments.

To be more realistic on due-date assignments weights are given to customers related to their relative importance for a company, and a penalty function is applied to optimize due-dates, in this study. As there will be times that all customers could not be satisfied at the same time. There will be a decision to be made in which customers will be delivered early, and which will be delivered late. Not only being late in production is a problem but also being early. As there will be stock holding costs etc.

Integration of three functions (IPPSDDA problem), which are mentioned above, are discussed in this study. Weighted scheduling and weighted due-date assignment are integrated with process plan selection. WMS is used as a dispatching rule and WSLK used as a due-date assignment rule. Studies made over IPPS, SWDDA, and IPPSDDA are surveyed and briefly given to comprehend the problem scope.

To present this idea and to explain the point of view clearly a problem set is generated. Shop floors with distinct characteristics in terms of machines, jobs, routes, and processing times are generated in order to evaluate the effectiveness and efficiency of the integration and the algorithms.

Evolutionary Strategies, Semi-Tabu Search and their hybrids with Random Search are used and compared with Random search, ordinary solution, and each other for all integration levels. Algorithms are provided with flow diagrams to better understand them.

Starting with unintegrated problem (SIRO-RDM) integration level is increased step by step to the fully integrated problem (WMS-WSLK) and the solution performance is observed with the above-mentioned algorithms. The best performance is obtained in the fully integrated level in most of the shop floors with RS/ES algorithm. In some of them, ES performs better than its hybrid.

## References

- Aldakhilallah, K.A., Ramesh, R., 1999. Computer-integrated process planning and scheduling (CIPPS): intelligent support for product design, process planning and control. *International journal of production research* 37, 481–500.
- Allaoua, H., Osmane, I., 2010. Variable Parameters Lengths Genetic Algorithm for Minimizing Earliness-Tardiness Penalties of Single Machine Scheduling With a Common Due Date. *Electronic Notes in Discrete Mathematics, ISCO 2010 - International Symposium on Combinatorial Optimization* 36, 471–478. <https://doi.org/10.1016/j.endm.2010.05.060>
- Baykasoğlu, A., Göçken, M., Unutmaz, Z.D., 2008. New approaches to due date assignment in job shops. *European Journal of Operational Research* 187, 31–45. <https://doi.org/10.1016/j.ejor.2007.02.020>
- Bhaskaran, K., 1990. Process plan selection. *The International Journal Of Production Research* 28, 1527–1539.
- Birman, M., Mosheiov, G., 2004. A note on a due-date assignment on a two-machine flow-shop. *Computers & Operations Research* 31, 473–480. [https://doi.org/10.1016/S0305-0548\(02\)00225-3](https://doi.org/10.1016/S0305-0548(02)00225-3)
- Biskup, D., Jahnke, H., 2001. Common due date assignment for scheduling on a single machine with jointly reducible processing times. *International Journal of Production Economics* 69, 317–322. [https://doi.org/10.1016/S0925-5273\(00\)00040-2](https://doi.org/10.1016/S0925-5273(00)00040-2)
- Brandimarte, P., 1999. Exploiting process plan flexibility in production scheduling: A multi-objective approach. *European Journal of Operational Research* 114, 59–71. [https://doi.org/10.1016/S0377-2217\(98\)00029-0](https://doi.org/10.1016/S0377-2217(98)00029-0)
- Cai, X., Lum, V.Y.S., Chan, J.M.T., 1997. Scheduling about a common due date with kob-dependent asymmetric earliness and tardiness penalties. *European journal of operational research* 98, 154–168.
- Ceven, E., Demir, H.I., 2007. Benefits of Integrating Due-Date Assignment with Process Planning and Scheduling (Master of Science Thesis). Sakarya University.
- Chen, Z.-L., 1997. Scheduling with batch setup times and earliness-tardiness penalties. *European Journal of Operational Research* 96, 518–537.
- Cheng, T.C.E., Kovalyov, M.Y., 1999. Complexity of parallel machine scheduling with processing-plus-wait due dates to minimize maximum absolute lateness. *European Journal of Operational Research* 114, 403–410. [https://doi.org/10.1016/S0377-2217\(98\)00111-8](https://doi.org/10.1016/S0377-2217(98)00111-8)
- Corti, D., Portioli-Staudacher, A., 2004. A concurrent engineering approach to selective implementation of alternative processes. *Robotics and Computer-Integrated Manufacturing* 20, 265–280. <https://doi.org/10.1016/j.rcim.2003.10.012>
- Demir, H.I., Erden, C., 2020. Dynamic integrated process planning, scheduling and due-date assignment using ant colony optimization. *Computers & Industrial Engineering* 149, 106799. <https://doi.org/10.1016/j.cie.2020.106799>
- Demir, H.I., Taskin, H., 2005. Integrated Process Planning, Scheduling and Due-Date Assignment (PhD Thesis). Sakarya University.
- Drstvenšek, I., Balič, J., 2003. Responding ability in dynamic production circumstances. *Journal of materials processing technology* 133, 71–78.
- Erden, C., Demir, H.I., Kökçam, A.H., 2019. Solving Integrated Process Planning, Dynamic Scheduling, and Due Date Assignment Using Metaheuristic Algorithms. *Mathematical Problems in Engineering* 2019.
- Gindy, N.N., Saad, S.M., Yue, Y., 1999. Manufacturing responsiveness through integrated process planning and scheduling. *International Journal of Production Research* 37, 2399–2418.
- Gordon, V., Kubiak, W., 1998. Single machine scheduling with release and due date assignment to minimize the weighted number of late jobs. *Information Processing Letters* 68, 153–159. [https://doi.org/10.1016/S0020-0190\(98\)00153-7](https://doi.org/10.1016/S0020-0190(98)00153-7)
- Gordon, V., Proth, J.-M., Chu, C., 2002. A survey of the state-of-the-art of common due date

- assignment and scheduling research. *European Journal of Operational Research* 139, 1–25. [https://doi.org/10.1016/S0377-2217\(01\)00181-3](https://doi.org/10.1016/S0377-2217(01)00181-3)
- Gordon, V.S., Strusevich, V.A., 2009. Single machine scheduling and due date assignment with positionally dependent processing times. *European Journal of Operational Research* 198, 57–62. <https://doi.org/10.1016/j.ejor.2008.07.044>
- Gordon, V.S., Strusevich, V.A., 1999. Earliness penalties on a single machine subject to precedence constraints: SLK due date assignment. *Computers & Operations Research* 26, 157–177.
- Khoshnevis, B., Chen, Q.M., 1991. Integration of process planning and scheduling functions. *J Intell Manuf* 2, 165–175. <https://doi.org/10.1007/BF01471363>
- Kim, Y.K., Park, K., Ko, J., 2003. A symbiotic evolutionary algorithm for the integration of process planning and job shop scheduling. *Computers & Operations Research* 30, 1151–1171. [https://doi.org/10.1016/S0305-0548\(02\)00063-1](https://doi.org/10.1016/S0305-0548(02)00063-1)
- Koulamas, C., 2017. Common due date assignment with generalized earliness/tardiness penalties. *Computers & Industrial Engineering* 109, 79–83. <https://doi.org/10.1016/j.cie.2017.04.040>
- Kovalyov, M.Y., 1997. Batch scheduling and common due date assignment problem: an NP-hard case. *Discrete applied mathematics* 80, 251–254.
- Lauff, V., Werner, F., 2004. Scheduling with common due date, earliness and tardiness penalties for multimachine problems: A survey. *Mathematical and Computer Modelling* 40, 637–655. <https://doi.org/10.1016/j.mcm.2003.05.019>
- Lawrence, S.R., 1994. Negotiating due-dates between customers and producers. *International Journal of Production Economics* 37, 127–138.
- Leung, C.W., Wong, T.N., Mak, K.L., Fung, R.Y.K., 2010. Integrated process planning and scheduling by an agent-based ant colony optimization. *Computers & Industrial Engineering* 59, 166–180. <https://doi.org/10.1016/j.cie.2009.09.003>
- Li, S., Ng, C.T., Yuan, J., 2011. Scheduling deteriorating jobs with CON/SLK due date assignment on a single machine. *International Journal of Production Economics* 131, 747–751. <https://doi.org/10.1016/j.ijpe.2011.02.029>
- Li, X., Gao, L., Li, W., 2012. Application of game theory based hybrid algorithm for multi-objective integrated process planning and scheduling. *Expert Systems with Applications* 39, 288–297. <https://doi.org/10.1016/j.eswa.2011.07.019>
- Luss, H., Rosenwein, M.B., 1993. A due date assignment algorithm for multiproduct manufacturing facilities. *European Journal of Operational Research* 65, 187–198.
- Ming, X.G., Mak, K.L., 2000. A hybrid Hopfield network-genetic algorithm approach to optimal process plan selection. *International Journal of Production Research* 38, 1823–1839. <https://doi.org/10.1080/002075400188618>
- Moon, C., Kim, J., Hur, S., 2002. Integrated process planning and scheduling with minimizing total tardiness in multi-plants supply chain. *Computers & Industrial Engineering* 43, 331–349. [https://doi.org/10.1016/S0360-8352\(02\)00078-5](https://doi.org/10.1016/S0360-8352(02)00078-5)
- Moon, C., Lee, Y.H., Jeong, C.S., Yun, Y., 2008. Integrated process planning and scheduling in a supply chain. *Computers & Industrial Engineering* 54, 1048–1061. <https://doi.org/10.1016/j.cie.2007.06.018>
- Morad, N., Zalzalá, A., 1999. Genetic algorithms in integrated process planning and scheduling. *Journal of Intelligent Manufacturing* 10, 169–179. <https://doi.org/10.1023/A:1008976720878>
- Mosheiov, G., 2001. A common due-date assignment problem on parallel identical machines. *Computers & Operations Research* 28, 719–732. [https://doi.org/10.1016/S0305-0548\(99\)00127-6](https://doi.org/10.1016/S0305-0548(99)00127-6)
- Özgülven, C., Özbakır, L., Yavuz, Y., 2010. Mathematical models for job-shop scheduling problems with routing and process plan flexibility. *Applied Mathematical Modelling* 34, 1539–1548. <https://doi.org/10.1016/j.apm.2009.09.002>
- Petrović, M., Vuković, N., Mitić, M., Miljković, Z., 2016. Integration of process planning and scheduling using chaotic particle swarm optimization algorithm. *Expert Systems with Applications* 64, 569–588.
- Phanden, R.K., Jain, A., Verma, R., 2011. Integration of process planning and scheduling: a state-of-the-art review. *International Journal of Computer Integrated Manufacturing* 24, 517–534.

<https://doi.org/10.1080/0951192X.2011.562543>

- Pinedo, M., Chao, X., 1998. Operations Scheduling with Applications in Manufacturing and Services. McGraw-Hill Higher Education.
- Rechenberg, I., 1965. Cybernetic solution path of an experimental problem. Ministry of Aviation, Royal Aircraft Establishment, Library Translation.
- Schwefel, H.-P., 1981. Numerical optimization of computer models. John Wiley & Sons, Inc.
- Seker, A., Erol, S., Botsali, R., 2013. A neuro-fuzzy model for a new hybrid integrated Process Planning and Scheduling system. Expert Systems with Applications 40, 5341–5351. <https://doi.org/10.1016/j.eswa.2013.03.043>
- Shabtay, D., 2016. Optimal restricted due date assignment in scheduling. European Journal of Operational Research 252, 79–89. <https://doi.org/10.1016/j.ejor.2015.12.043>
- Shrestha, R., Takemoto, T., Ichinose, K., Sugimura, N., 2008. A study on integration of process planning and scheduling system for holonic manufacturing with modification of process plans. International Journal of Manufacturing Technology and Management 14, 359–378.
- Tan, W., Khoshnevis, B., 2000. Integration of process planning and scheduling—a review. Journal of Intelligent Manufacturing 11, 51–63. <https://doi.org/10.1023/A:1008952024606>
- Thomalla, C.S., 2001. Job shop scheduling with alternative process plans. International Journal of Production Economics 74, 125–134. [https://doi.org/10.1016/S0925-5273\(01\)00119-0](https://doi.org/10.1016/S0925-5273(01)00119-0)
- Tuong, N.H., Soukhal, A., 2010. Due dates assignment and JIT scheduling with equal-size jobs. European Journal of Operational Research 205, 280–289. <https://doi.org/10.1016/j.ejor.2010.01.016>
- Usher, J.M., 2003. Evaluating the impact of alternative plans on manufacturing performance. Computers & Industrial Engineering 45, 585–596. [https://doi.org/10.1016/S0360-8352\(03\)00076-7](https://doi.org/10.1016/S0360-8352(03)00076-7)
- Usher, J.M., Fernandes, K.J., 1996. Dynamic process planning—the static phase. Journal of Materials Processing Technology 61, 53–58.
- Vinod, V., Sridharan, R., 2011. Simulation modeling and analysis of due-date assignment methods and scheduling decision rules in a dynamic job

shop production system. International Journal of Production Economics 129, 127–146. <https://doi.org/10.1016/j.ijpe.2010.08.017>

- Weintraub, A., Cormier, D., Hodgson, T., King, R., Willson, J., Zozom, A., 1999. Scheduling with alternatives: a link between process planning and scheduling. IIE Transactions 31, 1093–1102. <https://doi.org/10.1023/A:1007683710427>
- Wilhelm, W.E., Shin, H.-M., 1985. Effectiveness of alternate operations in a flexible manufacturing system. International Journal of Production Research 23, 65–79. <https://doi.org/10.1080/00207548508904691>
- Yang, T., He, Z., Cho, K.K., 1994. An effective heuristic method for generalized job shop scheduling with due dates. Computers & industrial engineering 26, 647–660.
- Zhang, H.-C., Mallur, S., 1994. An integrated model of process planning and production scheduling. International Journal of Computer Integrated Manufacturing 7, 356–364. <https://doi.org/10.1080/09511929408944623>
- Zhang, L., Wong, T.N., 2015. An object-coding genetic algorithm for integrated process planning and scheduling. European Journal of Operational Research 244, 434–444. <https://doi.org/10.1016/j.ejor.2015.01.032>
- Zhang, Z., Tang, R., Peng, T., Tao, L., Jia, S., 2016. A method for minimizing the energy consumption of machining system: integration of process planning and scheduling. Journal of Cleaner Production 137, 1647–1662.

## Appendix. Due-date assignment rules

$$\text{Due date} = TPT + q_x \times k$$

where

$$q_x = q_1, q_2 \text{ or } q_3$$

$$q_1 = 0.5 \times P_{avg},$$

$$q_2 = P_{avg},$$

$$q_3 = 1.5 \times P_{avg}$$

k is inversely determined according to the customer weights.

RDM (Random due assignment)

$$\text{Due} = N \sim (3 \times P_{avg}, (P_{avg})^2)$$

TPT: total processing time

$P_{avg}$ : mean processing time of all job waiting





# Classifier Selection in Resource Limited Hardware: Decision Analysis and Resolution Approach

Atilla Özgür <sup>1</sup> 

<sup>1</sup>Jacobs University Logistics and Mathematics, Bremen/Germany  
a.oezguer@jacobs-university.de

## Abstract

Digitalization, Industry 4.0 and Internet of things (IoT) have become more popular in the recent years. Most of these systems depend on micro-controllers and sensors. These micro-controllers and sensors are mostly cheap, low RAM and low CPU systems; thus, they are resource constrained environments. In this study, a supervised learning classifier comparison technique suitable for resource constrained environments is proposed. This technique, Decision Analysis and Resolution (DAR), is originated in the domain of Software Engineering. First, DAR is explained using an example of car buying scenario. Then 11 off-the-shelf classifiers are compared using DAR for low RAM and less powerful CPU environments in an intrusion detection scenario. This scenario simulated on well-known KDD99 intrusion detection dataset. All the experiments are realized using python scikit-learn package. According to the experiments, Decision Tree classifier is the most suitable to implement for resource constrained environments with a small lead. Results for the other three classifiers (Bagging, Multi Layer Perceptron, Random Forest) are also very similar. To aid the reproducibility of the experiments, the whole source code of the study is provided in the popular open source repository <https://github.com/ati-ozgur/classifier-comparison-using-DAR>.

**Keywords:** Classifier selection, Decision analysis and resolution, Machine learning, Performance metrics, Resource limited environment

## Kısıtlı Kaynaklı Donanımlarda Sınıflandırıcı Seçimi: Karar Analizi ve Çözüm Yaklaşımı

### Öz

Dijitalleşme, Endüstri 4.0 ve Nesnelerin İnterneti (IoT) son yıllarda daha popüler hale gelmiştir. Bu sistemlerin çoğu mikro denetleyicilere ve sensörlere bağlıdır. Bu mikro denetleyiciler ve sensörler çoğunlukla ucuz, düşük RAM ve düşük CPU sistemleridir; bu nedenle, kaynak kısıtlı ortamlardır. Bu çalışmada, kaynak kısıtlı ortamlara uygun, denetimli bir öğrenme sınıflandırıcı karşılaştırma tekniği önerilmiştir. Bu teknik, Karar Analizi ve Çözümü (DAR), Yazılım Mühendisliği alanında ortaya çıkmıştır. İlk olarak DAR, örnek bir araba satın alma senaryosu ile açıklanmıştır. Ardından, 11 hazır sınıflandırıcı, bir saldırı tespit senaryosunda düşük RAM ve düşük CPU ortamları için DAR kullanılarak karşılaştırılmıştır. Bu senaryo, iyi bilinen KDD99 saldırı tespit veri setinde gerçekleştirilmiştir. Tüm deneyler python scikit-learn paketi kullanılarak gerçekleştirilmiştir. Deneylere göre, Karar Ağacı sınıflandırıcısı, diğer sınıflandırıcılara göre küçük bir fark ile kaynak kısıtlı ortamlara uygulanmak için en uygun sınıflandırıcıdır. Diğer üç sınıflandırıcı (Boosting, Çok Katmanlı Algılayıcı, Rastgele Orman) sonuçları da çok benzerdir. Deneylerin tekrarlanabilirliğine yardımcı olmak için, tüm kaynak kod popüler açık kaynak kod deposu <https://github.com/ati-ozgur/classifier-comparison-using-DAR>'da verilmiştir.

**Anahtar Kelimeler:** Kaynak kısıtlı ortam sınıflandırıcı seçimi, Karar analizi ve çözümü, Makine öğrenmesi, Performans metrikleri

\* Corresponding Author.  
E-mail: a.oezguer@jacobs-university.de

Received : 20 June 2020  
Revision : 28 January 2021  
Accepted : 08 February 2021

## 1. Introduction

Industry 4.0 is first termed by Germany Government to increase digitalization and automation in manufacturing sector (Lasi et al. 2014, Uygun & Ilie 2018). Internet of things (IoT), sensors and mobile systems are very important topics in Industry 4.0 and digitalization (Ileri & Furat 2020). These systems are depended on computers. Even though some computers they work on are powerful, most of the computers they work on are resource constrained environments such as micro-controllers and sensors (Karahan & Hökelek 2020). This can be seen from increasing market size of micro controllers (Research 2019), which was 18.6 billion USD in 2018. Most of the market still belongs to 8- and 16-bit micro-controllers (Research 2019). Most prevalent reasons for this phenomenon are cost and size. Resource-limited hardware products, for example 8-16-bit micro-controllers, are cheaper and smaller. Algorithms and systems that work on in these resource constrained environments are becoming more valuable.

Supervised learning also known as classification is a widely used technique for variety of problems, such as face recognition (Yavuz et al. 2016), bioinformatics (Yılmaz 2020), medical (Saleh & Hussein 2019), and intrusion detection (Özgür & Erdem 2012). Widely different classification algorithms are proposed in the literature (Taşçı 2019, Özgür et al. 2018, Özgür & Erdem 2018). Proposed algorithms mostly assume that necessary computing capabilities exist for algorithms to work and rarely address low computing requirements. But computing capabilities of micro-controllers are lower than general-purpose computers. Even though some algorithms are suitable for working with limited resources, how to choose among such algorithms in a limited hardware situation is rarely addressed.

In contrast, choosing between alternative solutions is a widely researched topic in other domains. For example, Basheleishvili et al. (Basheleishvili et al. 2019) proposed fuzzy-logic based model for selection of university staff. Similarly, Çınar and Uygun (Çınar & Özer Uygun 2019) proposed another fuzzy based model for supplier selection. On the other hand, Faydalı and Erkan (Faydalı & Erkan 2020) proposed VIKOR model for selection of factory machines.

This article proposes a decision theoretic approach for a practitioner to make an informed decision between classification algorithms in resource-limited hardware, such as a micro-controller with low RAM and low CPU power. This approach, called Decision Analysis and Resolution (DAR) (Team 2006), is originated in the domain of software engineering.

Remainder of the paper is as follows. Section 2 introduces different classifier comparison metrics and explains why using only accuracy is not a good choice. Section 3 gives basic introduction to DAR with car example. Section 4 introduces a classification example using a micro-controller in the intrusion detection

domain with a well-known dataset. Finally, section 6 concludes the paper.

## 2. Classifier Comparison Metrics

Most of the time, classifier evaluation is made with single metric: accuracy. For example, Özgür and Erdem reviewed 149 articles in intrusion detection domain that was published in SCI-index journals (Özgür & Erdem 2016). Among these reviewed articles, accuracy was used by 130 articles, making the accuracy metric the most used metric. On the other hand, second most used metric, False positive (False alarm) was used only 70 times. But using only accuracy to compare classifiers is simplistic at best and may be plain wrong in some cases.

Netflix (Amatriain & Basilico 2012) did not use their competition winner algorithm that increased their previous accuracy more than 10%. They had considered engineering efforts vs improved accuracy and had chosen another algorithm for their systems. Engineering effort is one example about considering other metrics for classifier evaluation among many others.

According to different domains and purposes, different base metrics may be more important. For example, according to Axelsson's seminal paper (Axelsson 1999), one of the most important performance metric for an intrusion detection system is false positive rate (false alarm rate). Axelsson proposed that all alarms should be investigated, and this investigation is a costly endeavor.

For screening diseases, ideally a classifier should be highly sensitive (high true positive rate) and should miss very few persons with the disease (Wilson & Jungner 1968). Missing a sick person will be more costly for the society in most situations. If disease is a contagious one, missing one could lead to more patients. If disease is expensive to treat in further levels, such as cancer, it is again helpful to detect it very early. For the wrong results (false positive), more qualified doctors or more expensive tests could check the results again; therefore, false positive results will be corrected in the second checks.

For low resource-environments, computer resource related metrics are more important. In such environments, resources like RAM and CPU may become more important parameters than accuracy. Therefore, low model size for low RAM and low training/testing time for better usage of CPU becomes important.

Table 1 shows other performance metrics that are used in different situations. Using different performance metrics may lead to choosing different classifiers. Instead of trying to choose a classifier based on a single metric, incorporating different criteria to decision process would be more reliable. Here, we propose using a decision technique from software engineering domain, Decision Analysis and Resolution (Team 2006) for this purpose.

**Table 1.** Classifier Performance Metrics

Metric
Detection Rate
False Negative
False Positive
Training Memory
Training Time
Testing Time
F-1 Measure
Model Size in RAM
Model Size in Storage
Others

### 3. Decision Analysis and Resolution (DAR)

Decision Analysis and Resolution (DAR) technique is originated in software engineering domain (Team 2006). It is a technique to make more informed decisions, for example buying versus building a new tool, subcontracting versus building a new software in house. An example helps to understand this technique better: we start with a common example of buying a car. Suppose that a company needs to buy a car and wants to make an informed decision. There are a lot of metrics which can influence the decision. Some of them are provided below.

- Cost
- Secondhand Worth
- Fuel Consumption
- Baggage Capacity

More metrics can be added to this list but to make the example simpler, count of metrics is intentionally restricted.

A decision between two brands (Brand F and Brand H) should be made. Table 2 shows DAR comparison between these two brands. A DAR process starts with 100 points and distribute these points to metrics, (weight column). In Table 2, Initial Cost is given 40 points; Secondhand Worth is given 20 points; Fuel Usage is given 30 points; and Baggage Capacity is given 10 points for a total of 100 points.

Initial Cost, Secondhand Worth, Fuel Usage and Baggage Capacity could not be compared because to the different range of values of these variables. Initial Cost and Secondhand Worth has values in dollars, while Fuel Usage has liters/100 km and Baggage capacity is  $cm^2$ . To be able to use mathematical operations meaningfully (multiplication and addition), we need to normalize these values to same range. Min-max normalization,

Equation 1 and Equation 2 could be used for this purpose. Equation 1 is used when higher values are better, and Equation 2 is used when higher values are worse. Since higher values for second hand worth and baggage size are better, these features are normalized using Equation 1. Similarly, since higher values for cost and fuel consumption are worse, these features are normalized using Equation 2.

$$x_{norm} = \frac{x_i - \min(x)}{\max(x) - \min(x)} \quad (1)$$

$$x_{norm} = 1 - \frac{x_i - \min(x)}{\max(x) - \min(x)} \quad (2)$$

Summarized version of DAR car buying scenario can be seen in Table 2. The Min-Max column values are found using other brands. For example, 10 different cars from similar segments are used to find minimum and maximum values for these features. F and H columns are real values for these cars, while normalized F and normalized H values are min-max normalized (Equation 1 and Equation 2) values. Since we normalize values using Equation 1 and Equation 2), total column shows how good this car it is. Excel version of the Table 2 is provided in the source code repository.

After calculation, Brand H is found to be a better choice according to given weights, even though initial cost of Brand H is higher. Here, Brand H having better values in the metrics of fuel consumption and baggage capacity are deciding factors for this decision, even though its initial cost is more expensive. DAR is a valuable technique to use when making a decision which has more than two criteria.

#### 3.1. How to decide weights

Decision weights change from user to user and domain to domain. Those who can decide these weights called differently according to domain, for example users, evaluators, stakeholders (Shukla & Auriol 2013), decision makers (Faydalı & Erkan 2020) and experts (Çınar & Özer Uygun 2019). Thus, weights chosen by these experts are often very subjective (Phillips & Polen 2002). Even though, weights chosen by single user is subjective, a number methods to find common ground between different users are proposed such as Rank ordinal method (Danielson & Ekenberg 2017) and Utility Rank Order Weighting (UROW) (Shukla & Auriol 2013).

**Table 2.** Decision Analysis and Resolution - Car Buying Scenario

	Weights	Min	Max	Brand F	Brand H	Normalized F	Normalized H	Weight*Normalized F	Weight*Normalized H
Initial Cost	60	75000	100000	80000	95000	0,6	0,2	24	8
2nd-Hand Worth	20	20000	40000	30000	37000	0,75	0,85	15	17
Fuel Usage	30	4	6	5	4,4	0,5	0,8	15	24
Baggage Capacity	10	350	500	400	500	0,34	1	3,34	10
							<b>Total</b>	<b>57,34</b>	<b>59</b>

#### 4. Decision Analysis and Resolution (DAR): Intrusion Detection Domain

Intrusion detection is a widely researched subject in literature (Özgür & Erdem 2012, Özgür et al. 2018, Özgür & Erdem 2018, Sahingoz 2019). In our example problem, an intrusion detection classifier should be decided for a resource constrained environment, such as a micro-controller with a low RAM and less powerful CPU environment. An example for this situation is given by Karahan & Kaya 2020. Different metrics can be used for this purpose, see Table 1.

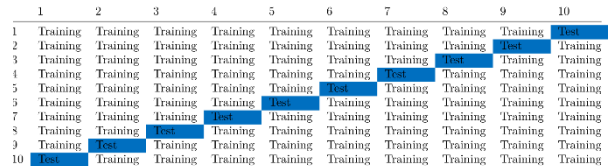
KDD99 is the most used data set in the intrusion detection domain, see (Özgür & Erdem 2016). According to Özgür and Erdem (Özgür & Erdem 2016), KDD99 dataset is used by 149 articles indexed in SCI-index between 2010-2015. KDD99 has 41 features to 23 classes. KDD99 is a suite of different datasets, one of these datasets is 10% of KDD99. KDD99 10% is widely used since training and testing time issues with full dataset. In our experiments, KDD99 10% (494021 instances) have been used due to training and testing time problems with full dataset. Table 3 shows the 10 fold cross validation results for different classifier comparison metrics using KDD99 10% dataset. As can be seen from Table 3, making a comparison between different classifiers using all these metrics is not easy.

**Table 3.** All Metrics for KDD99 dataset

Classifier name	Mean training time (sc)	Mean testing time(sc)	Mean accuracy score	Mean precision score	Mean recall score	Mean F1 score	Mean model size (k bytes)
AdaBoost	52.4111	1.8633	0.7866	0.7866	0.7866	0.7866	63810
Bagging	15.0923	0.2439	0.9997	0.9997	0.9997	0.9997	1027874
Decision Tree (CART)	2.3330	0.0084	0.9995	0.9995	0.9995	0.9995	126409
K Neighbors	0.0486	441.6670	0.9986	0.9986	0.9986	0.9986	149392857
Logistic Regression	154.8023	0.0195	0.9574	0.9574	0.9574	0.9574	8642
Multi Layer Perceptron	298.5620	0.2062	0.9963	0.9963	0.9963	0.9963	214557
Naive Bayes	0.4724	0.3285	0.9484	0.9484	0.9484	0.9484	16251
One Rule	0.0269	0.0547	0.4098	0.4098	0.4098	0.4098	836
Random Forest	19.5079	0.7065	0.9998	0.9998	0.9998	0.9998	17025372
Support Vector Machines	14453.56	1306.280	0.6352	0.6352	0.6352	0.6352	179564859
Zero Rule	0.0222	0.0004	0.5684	0.5684	0.5684	0.5684	839

Since our working environment is a resource constrained environment, following metrics are chosen for DAR comparison. These are model-size for low-memory constraint, training and testing time for less powerful CPU constraint and accuracy for overall performance. For these metrics 100 points are distributed. Accuracy is given 40 points, model size is given 30 points, and training time is given 10 points, and testing time is given 20 points. These metrics are chosen by the authors subjectively according to their knowledge in intrusion detection domain.

Using the same approach in section 3, Table 4 DAR Results is obtained. All the results are obtained using 10-fold cross validation. In 10-fold cross validation, dataset divided into 10 folds. Among these 10 folds, 9 of them are used for training, while remaining 1-fold is used for testing, see Figure 1. This means that for every cross validation 444619 instances are used for training and 49402 instances are used for testing. This procedure is repeated 10 times and results are averaged. Full metric results for every run can be seen in github repository.



**Figure 1.** Cross Validation 10-fold

Experiments are realized using python (3.8) and scikit learn toolkit (0.23.2) (Pedregosa et al. 2011). Experiment computer has an operating system of Linux Mint 20 Ulyana with 32G of RAM. Its CPU is an Intel(R) Core(TM) i77700K CPU @ 4.20GHz with 8 cores. All classifiers are used with default parameters. Software code for the experiments is provided in <https://github.com/ati-ozgur/classifier-comparison-using-DAR>.

In the results, best possible total score is 100 points of 100 points. Total column is found using the DAR process outlined in the section 3. Here it is weighted normalized summation of accuracy, model size, training and testing time. Accuracy is normalized using Equation 1 since high values are good, while other three metrics are normalized using Equation 2 since high values are worse. Then normalized values of these metrics are multiplied with weights and summed to get the total column in Table 4.

**Table 4.** DAR Results for Intrusion Detection in KDD99 dataset time.

Classifier Name	Accuracy (%)	Model Size(Mb)	Training Time (sec)	Testing Time (sec)	Total
Weights	40	30	10	20	100
Decision Tree (CART)	99.95	123.51	3.21	0.01	99.97
Bagging	99.97	1008.04	10.53	0.20	99.89
Multi Layer Perceptron	99.62	210.02	1054.49	0.46	99.42
Random Forest	99.98	16642.80	129.21	3.45	98.29
Logistic Regression	95.98	8.54	504.06	0.03	97.14
Naive Bayes	94.84	15.91	0.21	0.17	96.52
AdaBoost	78.66	64.39	24.32	0.94	85.55
Zero Rule	56.84	0.88	0.01	0.01	70.79
One Rule	40.90	0.88	0.01	0.02	60.00
K Neighbors	99.86	302790.12	2858.42	1517.63	57.78
Support Vector Machines	63.52	175356.42	34064.04	2686.32	27.94

According to results, best classifier with very small lead is the Decision Tree classifier. Other three

classifiers —Bagging, Multi Layer Perceptron, Random Forest — get very similar scores in DAR table. Firstly, the DAR process effectively removes outliers from consideration. Support Vector Machines and K Neighbors classifiers are outliers in model size and testing time categories and accordingly their total scores becomes lower. Secondly, if we have used only accuracy as metric, K Neighbors with accuracy of 99.86% will be among the best classifiers. But its total score is below 60 points since its other 3 metrics are very bad compared to other classifiers.

In short, DAR approach helps to choose among different classifiers according to given constraints.

#### 4. Discussion

Novelty of the proposed method is its simplicity. Compared to previously proposed decision methods (Basheleishvili et al. 2019), (Çınar & Özer Uygun 2019), (Faydalı & Erkan 2020), DAR is easier to understand and implement. An example Excel file for car example and python code for the IDS example is provided in the software source code site. The main model complexity is  $O(n)$  since only 1 loop exists in the solution. All other calculations are simple arithmetic operations, again showing simplicity of the method compared to alternatives.

Even though, DAR is a simple method, it is helpful to implement it for comparing alternatives. In the machine learning domain using only accuracy to rank classifiers is rampant. But, in real world implementations, other considerations should be considered.

Limitation of the current study is that weights are subjective and taken from only one expert. In a real-world scenario, other methods such as given in section 3.1 should be implemented to use information given by multiple experts.

#### 5. Conclusions

A method to choose among different classifiers named Decision Analysis and Resolution (DAR) is proposed. This method is originated in software engineering domain. According to given constraints, this method ranks the classifiers. A resource constrained environment has been chosen for demonstration purposes. For this environment, accuracy, model size, training time and testing time have been chosen for comparison metrics. Using scikit-learn toolbox, 11 supervised learning classifiers have been applied to well-known intrusion detection dataset KDD99. According to our results, Decision Tree is most suitable classifier for this resource constrained environment, even though Random Forest is the best accuracy classifier.

A natural progression of this work is to test the current code on real micro controllers like Raspberry Pi or Arduino. Using different datasets on the resource

constrained environments may be another application. For example, deep learning systems are very popular for image processing tasks like face detection and car plate detection. But these deep learning systems are also resource hungry. Comparing them would be an interesting application.

#### References

- Amatriain, X. & Basilico, J. (2012), 'Netflix recommendations: Beyond the 5 stars (part 1)', <https://netflixtechblog.com/netflix-recommendations-beyond-the-5-stars-part-1-55838468f429>.
- Axelsson, S. (1999), The base-rate fallacy and its implications for the difficulty of intrusion detection, in 'In Proceedings of the 6th ACM Conference on Computer and Communications Security', pp. 1–7.
- Basheleishvili, I., Bardavelidze, A. & Tsiramua, S. (2019), 'The development of a model for decision support system of assessment and selection of university academic staff', *Journal of Intelligent Systems: Theory and Applications* 2(2), 18–23.
- Çınar, A. & Özer Uygun (2019), 'Selecting green supplier using intuitionistic fuzzy AHP', *Journal of Intelligent Systems: Theory and Applications* 2(2), 24–31.
- Danielson, M. & Ekenberg, L. (2017), Trade-offs for ordinal ranking methods in multi-criteria decisions, in 'Lecture Notes in Business Information Processing', Springer International Publishing, pp. 16–27.
- Faydalı, R. & Erkan, E. F. (2020), 'A fuzzy VIKOR method for machine selection', *Journal of Intelligent Systems: Theory and Applications* 3(1), 7–12.
- Ileri, Y. & Furat, M. (2020), 'A roadmap for digitalization of industrial processes', *European Journal of Science and Technology* pp. 349 – 357.
- Karahan, O. & Hökelek, H. (2020), 'Mobile robot position controlling system based on IoT through Raspberry Pi', *Journal of Intelligent Systems: Theory and Applications* 3, 25 – 30.
- Karahan, O. & Kaya, B. (2020), 'Raspberry Pi firewall and intrusion detection system', *Journal of Intelligent Systems: Theory and Applications* 3, 21 – 24.
- Lasi, H., Fettke, P., Kemper, H.-G., Feld, T. & Hoffmann, M. (2014), 'Industry 4.0', *Business & Information Systems Engineering* 6(4), 239–242.
- Özgür, A. & Erdem, H. (2012), 'Saldırı tespit sistemlerinde kullanılan kolay erişilen makine öğrenme algoritmalarının karşılaştırılması', *Bilişim Teknolojileri Dergisi* 5, 41–48.
- Özgür, A. & Erdem, H. (2016), 'A review of KDD99 dataset usage in intrusion detection and machine learning between 2010 and 2015', *PeerJ Preprints* .
- Özgür, A. & Erdem, H. (2018), 'Feature selection and multiple classifier fusion using genetic algorithms in intrusion detection systems', *Journal of the Faculty of Engineering and Architecture of Gazi University* 33, 0 – 0.
- Özgür, A., Nar, F. & Erdem, H. (2018), 'Sparsity-driven weighted ensemble classifier', *International Journal of Computational Intelligence Systems* 11, 962–978.
- Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A.,

- Cournapeau, D., Brucher, M., Perrot, M. & Édouard Duchesnay (2011), 'Scikit-learn: Machine learning in python', *Journal of Machine Learning Research* 12(85), 2825–2830.
- Phillips, B. C. & Polen, S. M. (2002), 'Add decision analysis to your cots selection process', *Software Engineering Technology*.
- Research, G. V. (2019), 'Microcontroller market size, share & trends analysis report', <https://www.grandviewresearch.com/industry-analysis/microcontroller-market>. Last Accessed August 2019.
- Sahingoz, O. K. (2019), 'A clustering approach for intrusion detection with big data processing on parallel computing platform', *Balkan Journal of Electrical and Computer Engineering* 7, 286 – 293.
- Saleh, N. & Hussein, N. (2019), 'Artificial intelligence in corneal topography', *Journal of Intelligent Systems: Theory and Applications* 2(1), 1–6.
- Shukla, V. & Auriol, G. (2013), Methodology for determining stakeholders' criteria weights in systems engineering, in 'Proceedings of the Posters Workshop at CSDM'.
- Taşcı, E. (2019), 'A meta-ensemble classifier approach: Random rotation forest', *Balkan Journal of Electrical and Computer Engineering* 7, 182 – 187.
- Team, C. P. (2006), Cmmi for development, version 1.2, Technical report.  
URL: <http://btd.gazi.edu.tr/dergi/sayi/volume5-2-5.pdf>
- Uygun, Y. & Ilie, M. (2018), *Autonomous Manufacturing-related Procurement in the Era of Industry 4.0*, Springer Fachmedien Wiesbaden, pp. 81–97.
- Wilson, J. & Jungner, G. (1968), Principles and practice of screening for disease, Technical report, World Health Organization.
- Yavuz, H. S., Çevikalp, H. & Edizkan, R. (2016), 'A comprehensive comparison of features and embedding methods for face recognition', *Turkish Journal of Electrical Engineering and Computer Science* 24, 313 – 340.
- Yılmaz, A. (2020), 'Assessment of mutation susceptibility in DNA sequences with word vectors', *Journal of Intelligent Systems: Theory and Applications* 3(1), 1–6.



# Küme Birleşimli Sırt Çantası Probleminin Adaptif Yapay Arı Kolonisi Algoritması ile Çözümü

Rafet DURGUT<sup>1\*</sup>, İlim Betül YAVUZ<sup>2</sup>, Mehmet Emin AYDIN<sup>3</sup>

<sup>1</sup>Karabük Üniversitesi, Bilgisayar Mühendisliği Bölümü, Karabük, Türkiye

<sup>2</sup>Karabük Üniversitesi, Bilgisayar Mühendisliği Bölümü, Karabük, Türkiye

<sup>3</sup>UWE Bristol, Dept. of Computer Science and Creative Technologies, Bristol, İngiltere

rafetdurgut@karabuk.edu.tr, 1828126016 @karabuk.edu.tr, Mehmet.aydin@uwe.ac.uk

## Öz

Meta-sezgisel ve sürü zekâsı algoritmaları, NP-Zor optimizasyon problemlerine yaklaşık çözümler sunmak için uzun süredir kullanılmaktadır. Özellikle kombinatoriyal ve ikili problemler söz konusu olduğunda, algoritmalar içerisine gömülü komşu çözüm üretmek için kullanılan operatör fonksiyonları, aramanın çeşitliliğine sınırlamalar getirirken algoritmaların başarısında önemli bir rol oynar. Bu tür sınırlamalardan kaçmak ve çeşitliliği iyileştirmek için, birden fazla operatörün tek bir operatör yerine bir seçim şeması yoluyla kullanılması tercih edilir. Daha önce farklı sürü zekâsı ve meta-sezgisel algoritmalarla çeşitli kombinatoriyal problemleri çözmek için bir dizi operatör seçim şeması kullanılması daha yüksek etkinlik elde etmek için kullanılmıştır. Bu makalede, küme birleşimli sırt çantası problemleri, ilk kez, alternatif operatör seçim şemaları aracılığıyla seçilen birden fazla operatör içeren ikili bir yapay arı kolonisi algoritması ile çözülmüştür. Önerilen yöntem için farklı kredi atama yaklaşımları, farklı kayan pencere boyutları ve parametre konfigürasyonları test edilmiştir. Seçim şemalarının özellikleri kapsamlı olarak 30 kıyaslama problemi üzerinde incelenmiştir. Bu problem kümeleri için en iyi performans gösteren algoritma konfigürasyonu önerilmiştir. Çalışma, başarılı bir seçim şemasına sahip adaptif ikili yapay arı kolonisi algoritmasını sunmaktadır.

**Anahtar kelimeler:** Yapay Arı Kolonisi, Adaptif Operatör Seçimi, Küme Birleşimli Sırt Çantası Problemi

## Solving Set Union Knapsack Problems with Adaptive Binary Artificial Bee Colony

### Abstract

Metaheuristic and swarm intelligence algorithms have been utilised to solve optimization problems with NP-Hard nature providing approximate solutions for a long time. Especially in the case of combinatorial and binary problems, operator functions embedded in the algorithms to generate neighboring solutions play a crucial role in the success of the algorithms while each operator imposes limitations upon the diversity of the search. In order to escape of such limitations and improve the diversity, multiple operators are preferred to use through a selection scheme instead of a single operator. However, the nature of selection scheme whereby operators are opted out also matters for higher efficiency, where a number of operator selection schemes have been used to solve various combinatorial problems with different swarm intelligence and metaheuristic algorithms before. In this paper, set union knapsack problems are, first time, solved with a binary artificial bee colony algorithm embedded with multiple operators selected through alternative operator selection schemes. Different credit assignment approaches, different sliding window lengths and parameter configurations are tested for the proposed method. The characteristics of the selection schemes are studied and the best performing one is suggested using a comprehensive experimentation over 30 benchmark problems. The study concludes a particular variant of binary artificial bee colony algorithm with a successful selection scheme.

**Keywords:** Artificial Bee Colony, Adaptive Operator Selection, Set Union Knapsack Problem.

\* Sorumlu yazar. Rafet Durgut  
E-posta adresi: rafetdurgut@karabuk.edu.tr

Alındı : 05 Ocak 2021  
Revizyon : 08 Şubat 2021  
Kabul : 15 Şubat 2021

## 1. Giriş (Introduction)

Sırt çantası problemi eldeki imkanlardan en yüksek faydayı sağlamak amacıyla çözülmesi gereken ve doğrusal zamanda çözülmesi mümkün olmayan bir optimizasyon problemidir. Günlük yaşamda, kaynak planlama (Bitran vd., 1981), kriptografi (Odlyzko vd., 1990), lojistik (Klamroth vd., 2000), üretim sistemleri (Bretthauter vd., 2002), gibi birçok alanda uygulama alanına sahiptir.

Sırt çantası uygulanacak probleminin yapısına göre 0/1 Sırt çantası (Moradi vd., 2021), çok boyutlu sırt çantası (García vd., 2020) veya küme birleşimli sırt çantası (Wei ve Hao, 2021) gibi farklı tipleri bulunmaktadır; Küme birleşimli sırt çantası (KBSC) problemi 0-1 sırt çantası probleminin özel bir halidir. Bu problemde mevcut elemanlar buldukları kümeye göre ele alınmaktadır. KBSC probleminin şimdiye kadar çeşitli uygulama ve alanlarda değerli oluşu bilinmektedir (Goldschmidt vd., 1994).

Problemde  $n$  tane eleman için  $U = \{1, 2, \dots, n\}$  elemanlar kümesini ve  $m$  tane nesne için  $S = \{1, 2, \dots, m\}$  nesne kümesini ifade etmektedir. Öyle ki  $S$  kümesindeki her nesne  $i \in S$  ( $i = 1, 2, \dots, m$ )  $U_i \subseteq U$  karşılık gelmektedir ve bir fayda değerine ( $p_i > 0$ ) sahiptir.

$$\max P(A) = \sum_{i \in A} p_i \quad (1)$$

$$W(A) = \sum_{u_j \in U_{i \in A}} u_i w_j \leq C, A \subseteq S \quad (2)$$

Amaç fonksiyonu Eşitlik (1)'de belirtilmiş olup, bu fonksiyonun amacı maksimum faydayı sağlayacak en iyi  $A \subseteq S$  alt kümesini ( $S^*$ ) belirlemektir. Eşitlik (2)'deki  $w_j$  değişkeni elemanın ağırlığını,  $C$  ise sırt çantasının kapasite sınırını ifade etmektedir.

Goldschmidt vd. (1994) KBSC probleminin kesin çözümü için hipergraf tabanlı dinamik programlama algoritmasının kullanılmasını önermiştir. Fakat bu algoritma düşük boyutlu problemler için hızlı çözüm önermesine karşın, problem boyutu arttıkça doğrusal zamanda bir çözüm sunamamaktadır. Benzer şekilde, Arulsevan (2014) aç gözlü bir yakınsama algoritması önermiştir.

Metasezgisel algoritmalar, doğrusal zamanda çözülemeyen problemlere uygulanabilecek alternatif ve güçlü bir seçimdir. Doğadan esinlenilerek geliştirilen bu algoritmalar içerisinde; Genetik Algoritma (Holland, 1975, Lin vd., 2020), Parçacık Sürüsü Optimizasyonu (Russel ve Eberhart, 1995, Bansal, 2019) veya Yapay Arı Kolonisi (Karaboğa ve Baştürk, 2008, Xiang vd., 2021) gibi oldukça başarılı ve popüler yöntemler bulunmaktadır. Popülasyon tabanlı olan bu yaklaşımlarda her bir çözüm ilgili bireye atanır ve

komşu çözümlerden faydalanılarak arama uzayındaki en iyi çözümü sunan noktaya ulaşılmaya çalışılır.

Yapay arı kolonisi algoritması Karaboğa (2005) tarafından bal arılarının yiyecek bulma davranışından esinlenilerek sürekli optimizasyon problemlerine çözüm bulmak için geliştirilmiştir (Wang vd., 2020). Yöntem kullandığı eşitlikler gereği ikili optimizasyon problemlerine doğrudan uygulanamamaktadır. Bu sebeple, ikili optimizasyon problemlerine uygulanabilmesi için çeşitli düzenlemeler gerekmektedir (Lin vd., 2020). Bu düzenlemeler komşu çözüm üretilmesi için yapılmaktadır. Kiran vd. (2013) tarafından sunulan mantıksal özel veya kapısı kullanan binABC algoritması, Kasha vd. (2012) tarafından sunulan disABC algoritması gibi çeşitli komşu çözüm üretme mekanizmaları bulunmaktadır. Durgut (2020) ise, binABC algoritmasının yakınsama hızını arttırmak için bazı düzenlemeler sunmuştur.

Yapay arı kolonisi yaklaşımı ve benzeri sürü zekâsına dayalı metasezgisel optimizasyon algoritmaları, arama sürecinde ulaştıkları çözümleri geliştirerek (sömürü fazı) daha iyi çözümlere ulaşmaya çalışmakta veya arama bölgesinin farklı noktalarına ulaşarak (keşif fazı) yeni başarılı çözümlere ulaşmaya çalışmaktadır. Bahsi geçen iki faz (sömürü, keşif) arasında denge sağlanması oldukça zor olmakla beraber, bu denge yerel minimuma takılma ya da hızlı yakınsamama probleminin oluşmaması için önemlidir (Arani vd., 2013). Bu tip problemlerin oluşma durumunu en aza indirmek, yerel minimuma takılmadan en iyi çözüme hızlı bir şekilde ulaşabilmek amacıyla adaptif operatör seçimi yaklaşımı önerilmektedir (Fialho, 2010).

Bu yaklaşımda komşu çözüm üretimi aşamasında bir operatör yerine birden çok operatör içeren operatör havuzu kullanılması öne çıkarılmıştır. Temel fikir operatörlerin birbirlerini tamamlayıcı davranış gösterebilmeleri beklentisidir. Burada tamamlayıcılık bir kriterle dayandırılarak tarif edilmektedir. Başarılı operatörlerin seçim şansının artırılması sayesinde operatörlerin birbirlerini tamamlayıcı davranabilecekleri ve böylece en iyi çözüme hızlı şekilde ulaşılması mümkün olabilecektir. Bu amaçla bu çalışma yapılmış ve tatmin edici sonuçlara ulaşılmıştır.

Bu çalışmada, küme birleşimli sırt çantası problemine kaliteli çözümler sunulabilmesi için adaptif yapay arı kolonisi yaklaşımı önerilmiştir. Literatür incelendiğinde, KBSC probleminin çözümü için kullanılmış olan Metasezgisel algoritmaların tüm süreci tek bir operatör ile yürüttüğü görülmüştür (He, 2018, Özsoydan vd., 2019). Önerilen yaklaşım ise, üç farklı komşuluk operatörü ve üç farklı operatör seçim yöntemi için çalıştırılmış ve en iyi konfigürasyon elde edilmiştir. Elde edilen sonuçlar literatürdeki diğer başarılı yöntemler ile karşılaştırılmış ve var olan yöntemlerden daha kaliteli çözümler üretilmiştir. Sonuç olarak bu çalışmanın katkısı, KBSC probleminin çözümü için literatürdeki çalışmalardan daha kaliteli çözümler üretmesi olarak görülmektedir.



## 2. YAPAY ARI KOLONİSİ (Artificial Bee Colony)

Sürü zekasına dayalı bir optimizasyon algoritması olan Yapay Arı Kolonisi (YAK) Algoritması, bal arılarının yiyecek arama/bulma davranışlarından esinlenilerek Karaboğa (2005) tarafından geliştirilmiştir.

Yapay Arı Kolonisi algoritmasının esinlendiği doğada arılar besin arama davranışlarında işçi, gözcü ve kâşif olmak üzere farklı görevlere sahiptirler. Arıların besin arama davranışlarının geçtiği aşamalar, başka bir deyişle YAK çalışma adımları Şekil 1’de verilmiştir.

### Algorithm 1 Yapay Arı Kolonisi

- 1: Başlangıç popülasyonunu oluştur.
- 2: Durdurma kriteri sağlanana kadar
- 3: Eşitlik 4’ü kullanarak komşu çözüm üret.
- 4: Seçili çözüm ile aday çözüm arasında aç gözlü seçim işlemi yap.
- 5: Her besin kaynağı için olasılık değerini hesapla.
- 6: Olasılık değerine göre besin kaynağı seç.
- 7: Eşitlik 4’ü kullanarak komşu çözüm üret.
- 8: Seçili çözüm ile aday çözüm arasında aç gözlü seçim işlemi yap.
- 9: Tüklenen besin kaynağı var ise yeni geçerli çözüm ile değiştir.
- 10: Şimdiye kadar bulunan en iyi çözümü sakla

### Şekil 1. YAK çalışma adımları (ABC Working Steps)

Başlangıç olarak yiyecek kaynaklarının üretilmesi için Kâşif arılar yiyecek kaynaklarını rastgele aramaya başlarlar. Yöntem içerisinde ise, rastgele geçerli çözüm atanması Eşitlik (3)’e göre yapılmaktadır.

$$x_{ij} = x_{ij}^{min} + rand(0,1)(x_{ij}^{max} - x_{ij}^{min}) \quad (3)$$

Eşitlik (3)’te,  $i$ . besin kaynağının  $j$ . boyutuna (parametresine) ait rastgele değer ( $x_{ij}$ ) verilen sınırlar ( $x_{ij}^{max} - x_{ij}^{min}$ ) aralığında rastgele olarak belirlenir.

Bütün besinlerin tüm boyutlarına denklemden elde edilmiş değerler atanarak ilk çözümler (kaynaklar) oluşturulur.

İşçi arı fazında, arılar mevcut yiyecek kaynağının komşuluğunda yeni aday çözüm oluşturulur. Yeni oluşturulan çözümün kalitesi eskisinden daha iyi ise aç gözlü seçim uygulanarak bu çözüm ile güncellenir. Böylelikle işçi arı fazında her arı üzerinde çalıştığı besin kaynağını geliştirmektedir.

$$V_{ij} = x_{ij} + \phi_{ij}(x_{ij} - x_{kj}) \quad (4)$$

Eşitlik (4)’te her bir arı komşu bir arıdan aldığı bilgi dahilinde bu arının komşuluğunda denklem ile yeni bir çözüm üretir ve bu sayede besin kaynağını geliştirir.  $V_{ij}$  değeri,  $i$ . arının  $j$ . boyutu için üretilmiş yeni değerdir. Bu değer, mevcut kaynağın komşuluğunda rastgele başka bir kaynağın yine aynı boyut bilgisine göre ( $x_{kj}$ ) güncellenmektedir.  $\phi$  değeri hangi ölçüde ilerleneceğini ifade eden katsayıdır ve  $[-1,1]$  aralığında rastgele üretilmektedir.

$$F_i = \begin{cases} -\frac{1}{1+fi}, & fi \geq 0 \\ 1 + abs(fi), & fi < 0 \end{cases} \quad (5)$$

Bulunan kaynağın kalitesi, uygunluk fonksiyonu ( $fi$ ) sonucunda hesaplanır. Kalite değerinin sıfırdan büyük ve eşit ya da küçük olmasına göre bu kaynağa Eşitlik (5)’te bulunan denklemde hesaplanan uygunluk değeri atanmış olur.

Gözcü arı fazında, arılar tüm görevli arıların elde ettikleri kaynak bilgilerini toplarlar. Bu bilgiler ışığında gözcü arının en iyi besin kaynağını seçebilmesi için kaynaktaki nektar miktarı ile orantılı olarak Eşitlik (6)’da belirtildiği gibi bir olasılık değeri belirlenir. Bu sayede gözcü arı aşamasında geliştirilmesi hedeflenen besin kaynakları belirlenmiş olur.

$$p_i = \frac{F_i}{\sum_{j=1}^{SN} F_j} \quad (6)$$

Kaynakların kalitesine göre bulunmuş olan uygunluk değerlerinin ardından bu kaynakların hangisinin geliştirileceği tüm kaynaklar baz alınarak hesaplanan olasılık değerine bağlıdır. Her bir çözümün olasılık değeri Eşitlik 6’da elde edilen değere göre belirlenmektedir. Bu olasılık değerlerine göre çözümler üzerinde güncelleme test edilir. Kâşif arı fazında, limit değerine ulaşan güncellemeye rağmen iyileşmeyen ilk çözüm popülasyondan atılarak, yerine rastgele ve geçerli bir çözüm dahil edilir. Bu şekilde algoritma ilk iterasyonunu tamamlayıp tekrar işçi arı aşamasına dönmektedir. Algoritma durdurma kriteri sağlanana kadar devam etmektedir.

## 3. İKİLİ YAPAY ARI KOLONİSİ (Binary Artificial Bee Colony)

Yapay arı kolonisi algoritması sürekli optimizasyon problemleri çözme amacıyla geliştirildiği için, ikili optimizasyon problemlerine doğrudan uygulanamamaktadır. Uygulanabilmesi için bazı düzenlemeler yapılmalıdır. Bu düzenlemeler iki sınıfa ayrılabilir; bunlardan biri probleme uygulama aşamasında sürekli karar değişkenlerinin ikili uzaya yerleştirilmesi (Kiran vd., 2015), diğeri ise karar değişkenlerinin ikili formda (Kiran vd., 2013) olmasıdır. İlk sınıftaki düzenleme kullanıldığında yöntemin sürekli uzayda aramaya devam eder, fakat amaç fonksiyonu içerisinde sürekli uzaydan ikili uzaya haritalama fonksiyonu gereklidir. İkinci sınıftaki düzenlemeler de ise, genellikle mantıksal karşılaştırma ve ifadeler kullanılmaktadır.

Bu çalışmada literatürde sıklıkla kullanılan binABC (Kirav vd., 2013), disABC (Kashan vd., 2013) ve yeni önerilmiş olan ibinABC (Durgut, 2020) operatörleri adaptif bir mekanizma içerisinde birlikte kullanılmaktadır.

### 3.1. BinABC Algoritması (BinABC Algorithm).

Çözüm uzayı ikili yapıda olan optimizasyon problemleri için geliştirilmiş XOR tabanlı arı kolonisi algoritmasıdır. Bu algoritma Kiran vd. tarafından "Kapasite tesis yerleştirme problemi" çözümü için geliştirilmiştir (Kirav vd., 2013). YAK algoritmasının ikili yapıya uygun hale gelmesi için Eşitlik (3) ve Eşitlik (4)'ün uygun formata dönüştürülmesi gerekmektedir.

$$X_{i,j} = \begin{cases} 0, & r_{i,j} < p \\ 1, & r_{i,j} \geq p \end{cases} \quad (7)$$

Diğer yaklaşımlarda olduğu gibi bu yöntemde de başlangıçta kullanılacak arı popülasyonu için üretilen rastgele çözümler, Eşitlik (3)'ün Eşitlik (7) olarak güncellenmesiyle oluşturulmaktadır.

$$V_i^j = X_i^j \oplus [\varphi(X_i^j \oplus X_k^j)] \quad (8)$$

İlk arı popülasyonu oluşturulduktan sonra, işçi ve gözcü arı pozisyonlarının hesaplandığı aday çözümün ifade eden Eşitlik (4), Eşitlik (8)'deki gibi güncellenir. Eşitlik (8)'de  $V_i^j$ , i. aday çözümün j. boyutunu temsil eder.  $X_i^j$  i. işçinin j. boyutu,  $X_k^j$  k. işçinin j. boyutu,  $\oplus$  XOR mantıksal operatörü,  $\varphi$  ise %50 olasılıkla mantıksal değil kapısını ifade eder. Eğer  $\varphi$  0,5'ten küçükse sonuç  $(X_i^j \oplus X_k^j)$  ifadesinin terslenmiş halidir, büyük ve eşitse sonuç terslenmemiştir.

### 3.2. DisABC Algoritması (DisABC Algorithm)

ABC nin diğer ikili versiyonu olan DisABC algoritması Kashan vd. tarafından önerilmiştir (Kashan vd., 2012). Bu algortmada YAK algoritmasının yapısal durumu baz alınarak işçi arı ( $X_i$ ) diğer bir deyişle seçili çözüm ve komşu arı ( $X_k$ ) arasındaki farklılığın ölçülmesi önemlidir. Bu noktada, Eşitlik (9)'da Jaccard benzerlik katsayısı bulunarak 1'den çıkarılır, böylelikle farklılık bulunur.

$$\text{Dissimilarity}(X_i, X_k) = 1 - \text{Similarity}(X_i, X_k) \quad (9)$$

$$\text{Dissimilarity}(X_i, X_k) = 1 - \frac{m_{11}}{m_{01} + m_{10} + m_{11}} \quad (10)$$

Eşitlik (10)'da,  $m_{01}$  değeri  $X_{i,j}=0$  ve  $X_{k,j}=1$  olduğu durumdaki bitlerin sayısını,  $m_{10}$  değeri  $X_{i,j}=1$  ve  $X_{k,j}=0$  olduğu durumdaki bitlerin sayısını,  $m_{11}$  değeri  $X_{i,j}=1$

ve  $X_{k,j}=1$  olduğu durumdaki bitlerin sayısını temsil eder.

Eşitlik (11) için, Aday çözüm:  $V_i$ , pozitif ölçek faktörü:  $\varphi$

$$\text{Dissimilarity}(V_i, X_i) \approx \varphi \times \text{Dissimilarity}(X_i, X_k) \quad (11)$$

Eşitlik (12) için  $n_1$  ve  $n_0$  katsayıları sırasıyla  $X_i$  binary (ikili) vektördeki 1 ve 0' ların sayısını temsil eder. Model sonucunda aday çözüm ve seçili çözüm arasındaki farklılık en aza indirgenmeye çalışılmaktadır. Seçili çözümde 1 olan hanelerden (bitlerden) seçilir, aday çözüme aktarılır ve diğer bitler sıfırlanır.

$$M_{11} + M_{01} = n_1, \quad M_{10} \leq n_0,$$

$M_{01}, M_{10}, M_{11} \geq 0$  ve tam sayı olmak üzere;

$$\min \left[ \left( 1 - \frac{M_{11}}{M_{01} + M_{10} + M_{11}} \right) - \varphi \left( 1 - \frac{m_{11}}{m_{01} + m_{10} + m_{11}} \right) \right] \quad (12)$$

### 3.3. IbinABC Algoritması (IbinABC Algorithm)

Durgut (2020), binABC operatörünü iyileştirerek ibinABC algoritmasını önermiştir. ibinABC algoritması iki yeni düzenleme içermektedir. Bunlardan ilki komşuluk operatörünün birden fazla bit üzerinde uygulanmasıdır. Operatörün uygulanacağı bit sayısı Eşitlik (13)'e göre belirlenmektedir.

$$d_t = \text{rast}[0, a] + 0.1D * e^{-(t/t_{max})} + 1 \quad (13)$$

Eşitlik 10'da a yöntem parametresi olup, ölçeklendirme için kullanılmaktadır. D, problem boyutunu, t o anki iterasyon değerini,  $t_{max}$  ise maksimum iterasyon sayısını ifade eder.

Bir diğer düzenleme ise  $\varphi$  değişkenin seçili ( $X_i$ ) ve komsu ( $X_k$ ) çözümün uygunluk değerine göre belirlenmesidir. binABC algoritmasında  $\varphi$  değeri rastgele olarak belirlenmektedir. Böyle bir durumda, iki çözümün uygunluk değerlerinin bir önemi yoktur. ibinABC'de ise eğer komsu çözüm daha iyi ise komşuluk operatöründe komsu çözümün ilgili hanesi (biti) işleme daha fazla etkili olacak şekilde belirlenmektedir. Aksi durumda ise,  $\varphi$  iterasyon değerine göre Eşitlik (14) 'teki gibi belirlenmektedir.

$$\varphi = \begin{cases} \varphi_{max} - \frac{(\varphi_{max} - \varphi_{min})}{\varphi_{max}} t, & f(X_i) < f(X_k) \\ 0, & \text{aksi halde} \end{cases} \quad (14)$$

#### 4. Adaptif İkili Yapay Arı Kolonisi (Adaptive Binary Artificial Bee Colony)

İkili yapay arı kolonisi içerisine adaptif bir yapı eklemek için iki temel problemi yanıtlamak gerekmektedir. Bunlardan ilki operatörlerin başarısının nasıl değerlendirileceği ve atanacak kredinin nasıl belirleneceğidir. İkinci problem ise, operatörlerin mevcut kredilere göre nasıl seçileceğidir. Her iki problemde kendi içerisinde farklı sorunlar barındırmaktadır.

Bir operatöre kredi değeri atanması için ilk olarak operatörün önceki uygulamalardaki başarısı dikkate alınmalıdır. Üretilen komşu çözümlerin amaç değerine göre veya iyileşme durumuna göre ödül atanmalıdır. Bu çalışmada ödül değeri Eşitlik (15)'e göre belirlenmektedir.

$$\text{ödül} = \frac{PD}{GB} (f(x') - f(x)) \quad (15)$$

Burada, PD problem boyutunu, GB o ana kadar elde edilmiş en iyi çözümün amaç fonksiyon değerini,  $x'$  aday çözümü,  $x$  ise mevcut çözümü ifade etmektedir. Her bir komşu çözüm üretilmesi sonucunda ödül değeri hesaplanmakta ve eğer pozitif bir değer elde edilmiş ise dikkate alınmaktadır. Her iterasyon sonrasında operatörlerin toplam ödül değerleri belirlenmektedir. Ardından bu ödül değerlerine göre operatörlere kredi değeri ataması yapılmaktadır. Bu noktada, o iterasyondaki elde edilen ödül değerleri kullanılabilir gibi belirli sayıda iterasyondan elde edilecek ödüller de dikkate alınabilir. Önceki çalışmalarımızda (Durgut ve Aydın, 2020) görüldüğü üzere anlık ödül kullanımının yerine ortalama ve en yüksek ödül değerlerinin kredi atamasında kullanılması daha faydalı sonuçlar üretmektedir.

Her bir operatöre ait kredi değerleri belirlendikten sonra operatör seçimi için kullanılabilir farklı yaklaşımlar mevcuttur. Bunlardan ilki Probability Matching (PM) yaklaşımıdır. Bu yaklaşımda operatörlerin seçilme olasılığı kredi değerlerine göre dağıtılır. Eşitlik (16) kullanılarak her bir olasılığın o iterasyondaki seçilme olasılığı belirlenir.

$$p_{i,t} = p_{min} + (1 - (K - 1) * p_{min}) \frac{Kred_{i,t}}{\sum_{j=1}^K Kred_{j,t}} \quad (16)$$

Eşitlik (16)'daki  $K$  değişkeni kullanılacak operatör sayısını,  $p_{min}$  en kötü operatöre atanacak minimum olasılık değerini,  $Kred_{i,t}$  ise  $i$ . operatörün kredi değerini ifade etmektedir. Bir diğer operatör seçimi yöntemi olan

Adaptive Pursuit (AP) ise “kazanan tümünü alır” prensibini kullanır. En iyi operatör en yüksek olasılık değeri ile güncellenirken diğer operatörler minimum olasılığa sahip olur. Yöntemin olasılık atama eşitliği Eşitlik (17)'de verilmiştir.

$$p_{i,t+1} = \begin{cases} p_{i,t} + \beta(p_{max} - p_{i,t}), & \text{eğer } i_t = i_t^* \\ p_{i,t} + \beta(p_{min} - p_{i,t}), & \text{aksi durumda} \end{cases} \quad (17)$$

Eşitlik (17)'de  $\beta$  uyarılma katsayısı,  $i_t^*$  ise en yüksek krediye sahip operatördür. Bu iki yaklaşımda operatörlerin kullanım sayıları ile ilgilenmemektedir. Bu sebeple az kullanılan operatörlere öncelik tanımamaktadır. Upper Confidence Bound (UCB) yaklaşımında ise az fırsat bulan operatörlere daha fazla fırsat sunmak amacıyla kullanılmaktadır. Yöntem Eşitlik (18)'e göre operatörlere seçim olasılığı atamaktadır.

$$p_{i,t+1} = \begin{cases} 1 - (K - 1)p_{min}, & \text{eğer } i_t = i_t^* \\ p_{min}, & \text{aksi durumda} \end{cases} \quad (18)$$

$i_t^*$  ise Eşitlik (19)'a göre belirlenmektedir. Burada  $C$  ayarlama katsayısı  $n$  ise seçilme sayılarıdır.

$$i_t^* = \underset{i=1..K}{\operatorname{argmax}} \{kred_{i,t} + C \sqrt{\frac{2 \log \sum_{j=1}^K n_{j,t}}{n_{i,t}}}\} \quad (19)$$

Operatörlere olasılık değeri atanmasından sonra rulet tekerine seçim yöntemi uygulanarak, operatör seçim işlemi gerçekleştirilmektedir. Çalışmada kullanılan algoritmanın çalışma adımları Şekil 2'de verilmiştir.

#### Algorithm 2 Adaptif Yapay Arı Kolonisi

- 1: Başlangıç popülasyonunu oluştur.
- 2: Durdurma kriteri sağlanana kadar
- 3: Kullanılacak operatörleri seç ve Toplam ödülleri sıfırla.
- 4: Operatörü kullanarak komşu çözüm üret.
- 5: Seçili çözüm ile aday çözüm arasında aç gözlü seçim işlemi yap.
- 6: Ödül hesapla. Eğer ödül değeri pozitif ise toplam ödüle ekle.
- 7: Her besin kaynağı için olasılık değerini hesapla.
- 8: Olasılık değerine göre besin kaynağı seç.
- 9: Kullanılacak operatörleri seç ve Toplam ödülleri sıfırla.
- 10: Operatörü kullanarak komşu çözüm üret.
- 11: Seçili çözüm ile aday çözüm arasında aç gözlü seçim işlemi yap.
- 12: Ödül hesapla. Eğer ödül değeri pozitif ise toplam ödüle ekle.
- 13: Tüklenen besin kaynağı var ise yeni geçerli çözüm ile değiştir.
- 14: Operatör kredilerini güncelle.
- 15: Şimdiye kadar bulunan en iyi çözümü sakla

Şekil 2. Adaptif YAK çalışma adımları (Adaptive ABC working steps)

#### 5. Deneysel Çalışmalar (Experimental Studies)

KBSC problemi kapasite kısıtlı optimizasyon problemi olduğu için geçersiz çözümler mevcuttur. YAK içerisinde geçersiz çözümlere izin verilmediğinden dolayı, çözümler üzerinde onarma işlemi yapılması gerekmektedir. He vd. (2018) önerdikleri çalışmada aç gözlü bir onarma yaklaşımı

uygulamaktadır. Bu yaklaşımda elemanlar küme içerisindeki frekanslarına ve fayda değerlerine göre sıralanır. Ardından kapasite sınırını aşan en değersiz elemanlar çantadan çıkarılır. Eğer mevcut kapasitede hala eleman eklenebilecek yer mevcut ise, en yüksek fayda sağlayacak elemanlar çanta içerisine eklenir. Daha detaylı bilgi için (He vd., 2018), (Özsoydan, Bakyasoglu, 2019) ve (Özsoydan, 2019) çalışmaları incelenebilir.

Yöntemlerin başarılarının karşılaştırılabilmesi için He vd. (2018) 30 problem örneği içeren veri kümesi oluşturmuşlardır. Tüm problem örnekleri  $m_n_x_y$  formatına göre oluşturulmuştur.  $m$  ve  $n$  nesne ve eleman sayılarını ifade ederken  $x$  ve  $y$  elemanların yoğunluğunu ve toplam ağırlığın çanta kapasitesine oranını ifade etmektedir. Veri kümesini, nesne ve eleman sayısına göre üç farklı gruba ayrılmış olup sayıları 100'den 500'e kadar artmaktadır. Kullanılan örnek veri kümesi Tablo 1'de verilmiştir.

**Tablo 1.** Kullanılan problem örnekleri. (The problem instances)

Grup 1 ( $m > n$ )		Grup 2 ( $m = n$ )		Grup 3 ( $m < n$ )	
Örnek No	Örnek Adı	Örnek No	Örnek Adı	Örnek No	Örnek Adı
G1_1	sukp 100_85_0.1_0.75	G2_1	sukp 100_100_0.1_0.75	G3_1	sukp 85_100_0.1_0.75
G1_2	sukp 100_85_0.15_0.85	G2_2	sukp 100_100_0.15_0.85	G3_2	sukp 85_100_0.15_0.85
G1_3	sukp 200_185_0.1_0.75	G2_3	sukp 200_200_0.1_0.75	G3_3	sukp 185_200_0.1_0.75
G1_4	sukp 200_185_0.15_0.85	G2_4	sukp 200_200_0.15_0.85	G3_4	sukp 185_200_0.15_0.85
G1_5	sukp 300_285_0.1_0.75	G2_5	sukp 300_300_0.1_0.75	G3_5	sukp 285_300_0.1_0.75
G1_6	sukp 300_285_0.15_0.85	G2_6	sukp 300_300_0.15_0.85	G3_6	sukp 285_300_0.15_0.85
G1_7	sukp 400_385_0.1_0.75	G2_7	sukp 400_400_0.1_0.75	G3_7	sukp 385_400_0.1_0.75
G1_8	sukp 400_385_0.15_0.85	G2_8	sukp 400_400_0.15_0.85	G3_8	sukp 385_400_0.15_0.85
G1_9	sukp 500_485_0.1_0.75	G2_9	sukp 500_500_0.1_0.75	G3_9	sukp 485_500_0.1_0.75
G1_10	sukp 500_485_0.15_0.85	G2_10	sukp 500_500_0.15_0.85	G3_10	sukp 485_500_0.15_0.85

Yöntem üç farklı adaptif operatör seçimi (Probability Matching (PM), Adaptive Pursuit (AP) ve Upper Confidence Bound (UCB)) ile test edilmiştir. Ödül mekanizması olarak pencere boyunca ortalama ve en yüksek ödül miktarları karşılaştırılmıştır. Pencere boyutu olarak ise küçük (5) ve orta (25) değerleri test edilmiş olup, bu kavram yöntem içerisinde son 5 ve 25 iterasyon boyunca elde edilen ödül değerlerini temsil etmektedir. Minimum seçilme olasılığı ( $P_{min}$ ) ise 0.1 ve 0.2 değerleri için test edilmiştir. Yöntem parametresi olan  $Alpha$  için ise 0.1, 0.5 ve 0.9 değerleri kullanılmıştır.

Parametre ayarlama için literatürdeki çalışmalarda olduğu gibi G2\_6 problem örneği kullanılmıştır. Bu fazda 30 ve diğer çalıştırmalarda için algoritma 100 farklı kez çalıştırılmış ve sonuçlar tablo halinde görselleştirilmiştir. Popülasyon boyutu 20 olarak seçilmiştir. Maksimum iterasyon sayısının belirlenmesi için ise önceki çalışmalar referans alınarak eleman veya nesne sayısına göre büyük olan belirlenmiştir.

Tablo 2'de üç yöntem üzerinden parametre ayarlama için elde edilen sonuçlar görülmektedir. AP için en iyi konfigürasyonların; ortalama ödül, pencere uzunluğu için 5, minimum olasılık değeri için 0.2 ve  $Alpha$  katsayısı için 0.9 değerleri olduğu görülmektedir. PM için ise AP'den farklı olarak pencere uzunluğunun 5 olduğu görülmektedir. UCB ise diğer yöntemlerden farklı olarak daha küçük  $Alpha$  değeri için en başarılı sonuçlarını üretmiştir. Bu üç yöntem arasında en iyi

ortalama değer PM ile üretildiği için sonraki deneylerde bu yöntem kullanılmıştır.

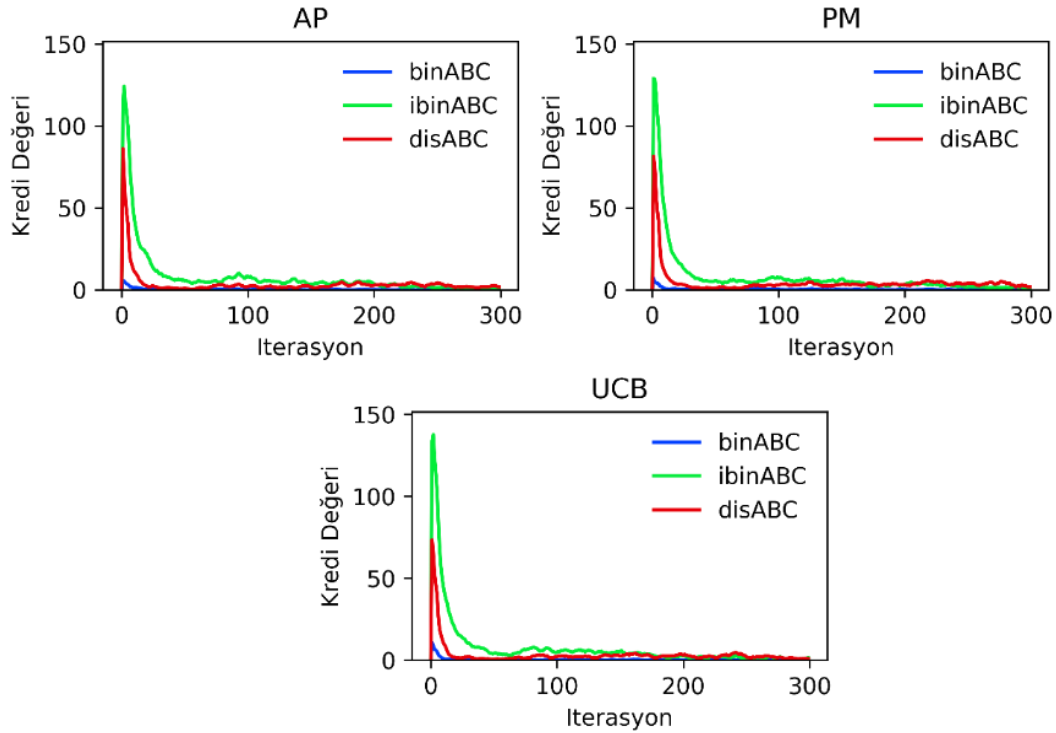
Şekil 3'te yöntemlerin (PM için elde edilmiş en iyi konfigürasyon için) zamana bağlı olarak verilmiş kredilerin ortalama değerleri ve Şekil 4'te ödül değerleri verilmiştir. Başlangıçta daha çok başarılı çözüm üretilebildiği için, ödül değerleri ve buna bağlı olarak kredi değerleri de yüksektir. Fakat yeni üretilen çözümlerin iyileşme oranları ve sayıları azalmaya başladığında ödül değeri ve dolayısıyla kredi değerleri de azalmaktadır. Bu noktada operatör seçiminin farklılıkları oldukça azdır.

Şekil 5 ve Şekil 6'da ise, sırasıyla operatör kullanım sayıları ve başarılı çözüm üretme sayıları verilmiştir. Kullanım grafikleri de 30 çalıştırmanın ortalaması olarak verilmiştir. Her üç yaklaşımda da başlangıç iterasyonlarında ibinABC daha fazla kullanılmaktadır. Farklılık, seçilme sayılarıdır. AP yaklaşımı kredisine göre bu operatöre diğer yaklaşımlarda daha çok fırsat vermiştir. UCB algoritmasında geçişler daha yumuşak olmasına karşın, en sert geçişler AP yaklaşımındadır. PM yaklaşımında 100. İterasyondan sonra disABC daha çok öne çıkmaya başlamıştır.

**Tablo 2.** Parametre ayarlama için elde edilen sonuçlar (The results of parameter tuning)

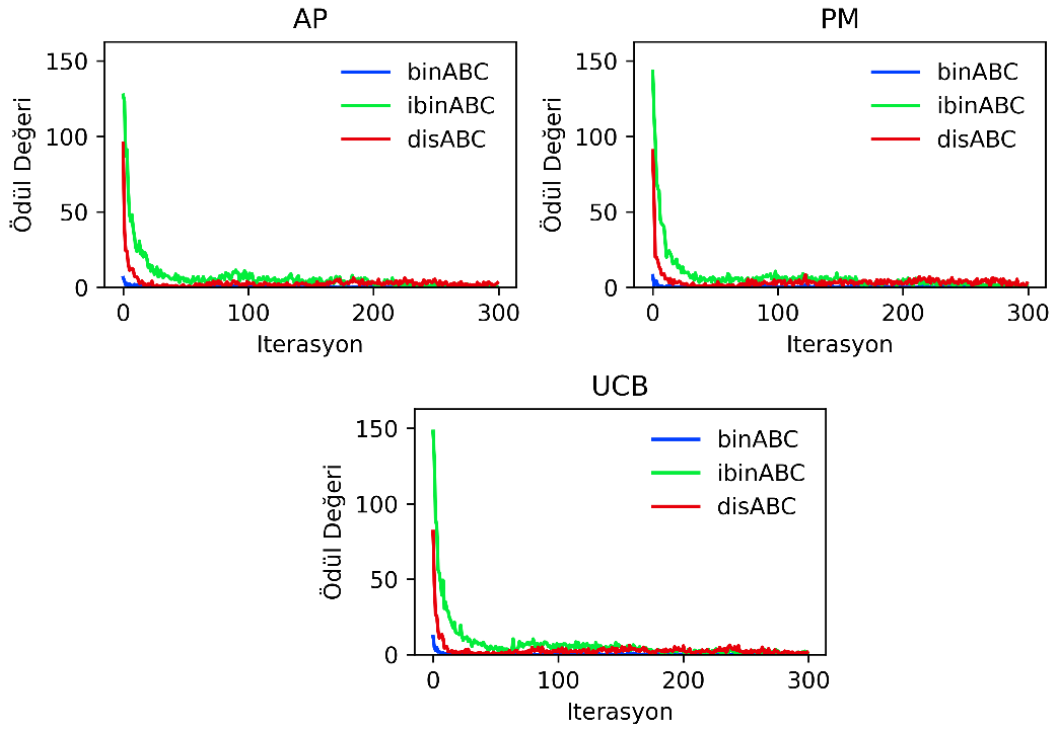
Ödül	Pencere	$P_{min}$	Alpha	AP		PM		UCB	
				En iyi	Ortalama	En iyi	Ortalama	En iyi	Ortalama
Ortalama	5	0.1	0,1	10998	10658,63	11113	10719,8	10948	10650,2
			0,5	11012	10676,30	10961	10692,13	10892	10686,57
			0,9	10735	10638,30	11023	10693,8	11054	10693,07
			0,1	11039	10670,40	10889	10700,8	<b>11410</b>	<b>10697,23</b>
			0,5	11113	10667,50	11000	10697,07	11081	10712,23
			0,9	10793	10651,83	<b>11425</b>	<b>10713,03</b>	11057	10697,7
	25	0.2	0,1	11178	10709,37	11093	10679,13	10793	10667,37
			0,5	11106	10704,50	11081	10670,27	11113	10657,9
			0,9	11025	10672,17	11051	10652,43	11106	10663,37
			0,1	11093	10662,50	11305	10687,93	10889	10677,67
			0,5	11082	10682,70	10851	10667,5	11025	10692,53
			0,9	<b>11425</b>	<b>10708,63</b>	11139	10700,73	10804	10669,37
En Yüksek	5	0.1	0,1	11007	10657,13	11222	10704,13	11039	10684,9
			0,5	11046	10683,53	11064	10724,37	11113	10689,3
			0,9	10949	10690,63	11178	10692,93	10835	10666,57
			0,1	11093	10722,43	11093	10684,7	11093	10656,2
			0,5	11081	10691,70	11139	10666,73	10931	10660,97
			0,9	10949	10634,17	11057	10671,23	10953	10657,57
	25	0.2	0,1	11078	10683,90	11132	10691,23	11251	10693,33
			0,5	11425	10691,93	11251	10697,4	11113	10708,83
			0,9	10990	10707,43	11410	10723,53	10953	10688,03
			0,1	10968	10681,27	10860	10654,13	11093	10715,7
			0,5	11093	10695,57	10990	10700,33	11178	10724,33
			0,9	10966	10662,50	11037	10704,23	11132	10713,1

**Kredi Grafiği**



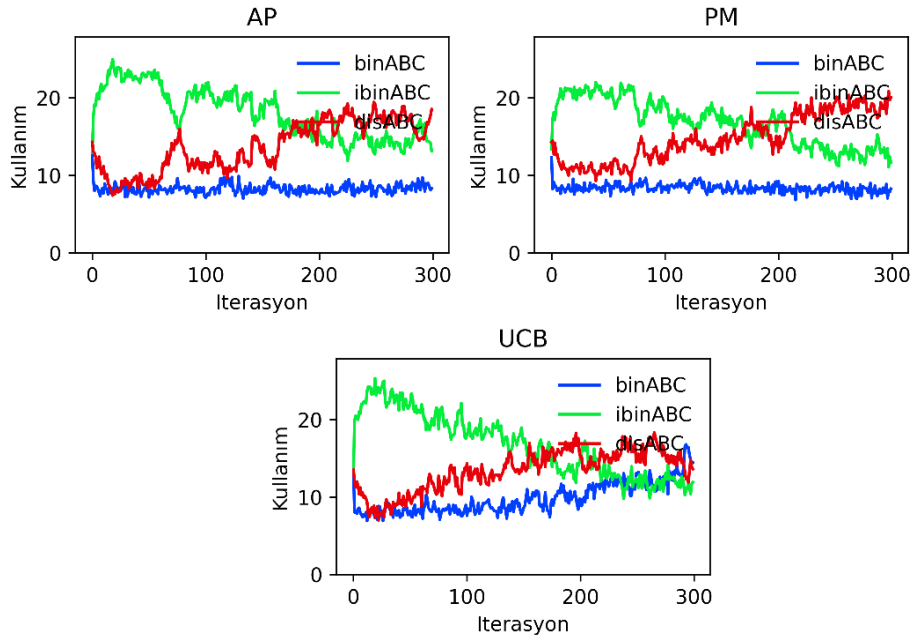
**Şekil 3.** Operatör seçim yöntemlerinin iterasyona bağlı kredi değerleri (The credit values of operator selection methods through iterations)

## Ödül Grafiği



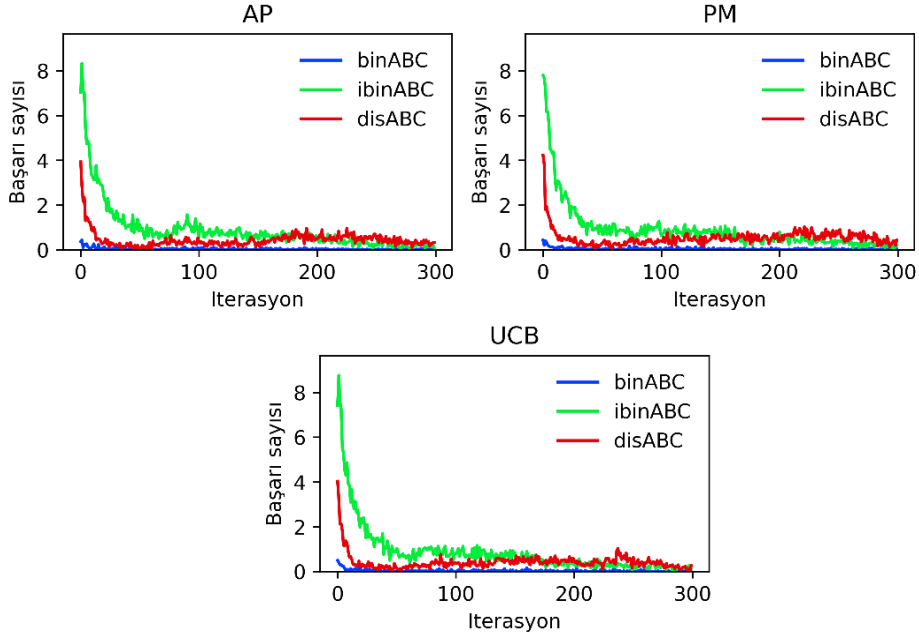
Şekil 4. Operatör seçim yöntemlerinin iterasyona bağlı ödül değerleri (The reward values of operator selection methods through iterations)

## Operatör Kullanım Sayıları



Şekil 5. Operatör kullanım sayılarının iterasyona göre değişimi (The usage counters of operator selection methods through iterations)

### Operatör Başarılı Güncelleme Sayıları



Şekil 6. Operatör başarı sayılarının iterasyona göre değişimi (The succes counters of operator selection methods through iterations)

Önerilen yaklaşım ile elde edilmiş sonuçlar literatürde var olan diğer yöntemler ile karşılaştırmalı olarak Ek Tablo' da verilmiş olup, karşılaştırılan yöntemlere ait değerler (He, 2018) ve (Özsoydan vd., 2019) çalışmalarından doğrudan alınmıştır. Bu tabloda 100 farklı çalıştırma sonucunda elde edilmiş en iyi sonuç, ortalama, standart sapma değerleri verilmiştir. Geliştirilen yaklaşım ile karşılaştırılan diğer yöntemler; A-SUKP, BABC, ABCBin, BinDE, GPSO\*, GPSO algoritmalarıdır. A-SUKP, KBŞÇ problemi için önerilmiş yakınsama algoritmasıdır. GA, Genetik Algoritma, BABC ve ABCBin İkili YAK algoritmalarıdır. BinDE ikili diferansiyel evrim algoritması, GPSO\* ve GPSO ise ağırlıklı ikili parçacık sürüşü algoritmasıdır.

Elde edilen sonuçların anlamlılığını kolaylaştırmak adına, Tablo 3'te ilk problem grubu için algoritmaların başarı sıralamaları verilmiştir. Önerilen adaptif

operatör seçimi yaklaşımı hem diğer YAK düzenlemelerinden hem de güncel literatürdeki çalışmalar içerisinde ortalama başarı sırasına göre birinci sıradadır. Burada 10 farklı örnek için 7 kez en iyi çözümleri sunmuş 3 kez ise ikinci sırada kalmıştır. En yakın rakibi GPSO olmuştur.

Tablo 4'te ise, ikinci grup problem örnekleri için algoritmaların performansları karşılaştırılmıştır. Önerilen çalışma bu problem örnekleri içerisinde en

başarılı algoritma olmuştur. Burada 10 farklı örnek için 6 kez en iyi çözümleri sunmuş, 4 kez ise ikinci sırada kalmıştır. En yakın rakibi GPSO olmuştur.

Tablo 5'te ise, üçüncü grup problem örnekleri için algoritmaların performansları karşılaştırılmıştır. Önerilen çalışma bu problem örnekleri içerisinde en başarılı algoritma olmuştur. 10 farklı örnek için 7 kez en iyi çözümleri sunmuş 3 kez ise ikinci sırada kalmıştır. En yakın rakibi GPSO olmuştur.

Tablo 3. İlk grup problem örnekleri (G1\_1 – G1\_10) için algoritmaların performansları (The performance of algorithms on first group problem instances)

P. No	A-SUKP	GA	BABC	ABCBin	binDE	GPSO*	GPSO	PMABC
G1_1	7	3	5	8	6	4	2	1
G1_2	8	7	2	5	4	6	3	1
G1_3	8	6	4	7	5	3	1	2
G1_4	8	5	4	7	6	3	2	1
G1_5	8	6	3	7	5	4	2	1
G1_6	8	3	5	7	6	4	2	1
G1_7	7	4	5	8	6	3	1	2
G1_8	7	5	4	8	6	3	2	1
G1_9	8	6	3	7	4	5	1	2
G1_10	8	7	4	3	6	5	2	1
Ortalama:	7,7	5,2	3,9	6,7	5,4	4	1,8	1,3

**Tablo 4.** İkinci grup problem örnekleri (G2\_1 – G2\_10) için algoritmaların performansları (The performance of algorithms on second group problem instances)

P. No	A-SUKP	GA	BABC	ABCBin	binDE	GPSO*	GPSO	PMABC
G2_1	8	5	4	7	6	3	2	1
G2_2	8	7	2	6	4	5	3	1
G2_3	8	6	3	7	5	4	2	1
G2_4	8	3	5	7	6	4	2	1
G2_5	7	6	3	8	5	4	1	2
G2_6	7	4	5	8	6	3	2	1
G2_7	8	5	4	7	6	3	1	2
G2_8	8	6	3	7	5	4	1	2
G2_9	8	6	3	7	4	5	1	2
G2_10	8	6	3	7	4	5	2	1
Ortalama:	7,8	5,4	3,5	7,1	5,1	4	1,7	1,4

**Tablo 5.** Üçüncü grup problem örnekleri (G3\_1 – G3\_10) için algoritmaların performansları (The performance of algorithms on third group problem instances)

P. No	A-SUKP	GA	BABC	ABCBin	binDE	GPSO*	GPSO	PMABC
G3_1	8	5	3	7	4	6	2	1
G3_2	8	7	1	5	3	6	4	2
G3_3	8	6	3	7	4	5	1	2
G3_4	8	6	3	7	4	5	2	1
G3_5	8	5	4	7	6	3	2	1
G3_6	8	6	3	7	5	4	2	1
G3_7	8	6	3	7	5	4	1	2
G3_8	8	4	5	7	6	3	2	1
G3_9	8	3	5	7	6	4	2	1
G3_10	8	6	3	7	4	5	2	1
Ortalama:	8	5,4	3,3	6,8	4,7	4,5	2	1,3

## 6. Sonuçlar (Conclusions)

Temel metasezgisel optimizasyon algoritmalarının başarısı kullandıkları komşu çözüm üretme operatörleri ile doğrudan ilişkilidir. Bu operatörün yerel minimumdan kurtulabilme ve hızlı yakınsama kabiliyetlerine aynı anda sahip olması oldukça zordur. Bu kabiliyetleri kazandırabilmek için yöntemlerin arama uzayının koşullarına göre adaptif olması gerekmektedir. Bu adaptasyon sayesinde zor problemlerin çözümü hızlandırılabilir. Bu çalışmada yapay arı koloni algoritmasına adaptif operatör seçimi eklenerek, ilk kez, küme birleşimli sırt çantası problemleri üzerindeki etkisi ortaya konulmuştur. Ele alınan problem için, üç farklı operatör seçim yöntemi ile yapay arı kolonisinin birleştirilmesi ve detaylı parametre analizi yapılmış olup, en iyi konfigürasyon belirlenmiştir. En iyi konfigürasyonda son 5 iterasyon boyunca elde edilmiş ödüllerin ortalama değerleri kullanılması halinde en iyi çözümler elde edilmiştir. Ardından, 30 farklı problem örneği için üretilmiş çözümler literatürdeki çalışmalar ile karşılaştırılmıştır. Karşılaştırma sonucunda önerilen yöntemin literatürdeki çalışmalardan daha gürbüz çözümler ürettiği görülmüştür.

Sonraki çalışmalarda, Operatör sayısının artırılarak KBŞÇ problemi üzerindeki etkisi incelenebilir. Bu sayede operatörlerin problem üzerindeki etkinliği ortaya konulabilir.

## Kaynaklar (References)

- Arani, B.O., Mirzabeygi, P., Panahi, M.S., 2013, "An improved PSO algorithm with a territorial diversity-preserving scheme and enhanced exploration-exploitation balance.", *Swarm and Evolutionary Computation*, 11,1-15.
- Arulselvan, A., 2014, "A note on the set union knapsack problem.", *Discrete Applied Mathematics*, 169, 214-218.
- Bansal, J.C., 2019, "Particle swarm optimization. In *Evolutionary and swarm intelligence algorithms*", Springer, Cham, 11-23.
- Bitran, G.R., Hax, A.C., 1981, "Disaggregation and resource allocation using convex knapsack problems with bounded variables.", *Management Science*, 27(4), 431-441.
- Brethauer, K.M., Shetty, B., 2002, "The nonlinear knapsack problem-algorithms and applications." *European Journal of Operational Research* 138(3), 459-472.
- Durgut, R., Aydın, M.E., 2020, "Adaptive binary artificial bee colony algorithm.", *Applied Soft Computing*, 107054.
- Durgut, R., 2020, "Improved binary artificial bee colony algorithm.", *arXiv preprint, arXiv, 2003(11641)*.
- Eberhart, R., Kennedy, J., 1995, "A new optimizer using particle swarm theory.", *MHS'95. Proceedings of the Sixth International Symposium on Micro Machine and Human Science*. Ieee.
- Fialho, A., 2010, "Analyzing bandit-based adaptive operator selection mechanisms.", *Annals of Mathematics and Artificial Intelligence*, 60(1-2), 25-64.
- García, J., Lalla-Ruiz, E., Voss, S. and Droguett, E. L., 2020, "Enhancing a machine learning binarization framework by perturbation operators: analysis on the multidimensional knapsack problem", *International Journal of Machine Learning and Cybernetics*, 1-20.



- Goldschmidt, O., Nehme, D., Yu., G., 1994, "Note: On the set-union knapsack problem." *Naval Research Logistics (NRL)*, 41(6),833-842.
- He, Y., 2018, "A novel binary artificial bee colony algorithm for the set-union knapsack problem.", *Future Generation Computer Systems*, 78, 77-86.
- Holland, J., 1975, "Adaptation in natural and artificial systems: an introductory analysis with application to biology.", *Control and artificial intelligence*.
- Karaboga, D., Basturk.B., 2008, "On the performance of artificial bee colony (ABC) algorithm.", *Applied soft computing*, 8(1),687-697.
- Karaboga D.,2005 "An Idea Based on Bee Swarm for Numerical Optimization", Technical Report-TR06.
- Kashan, M.H., Nahavandi, N. and Kashan, A. H., 2012, "DisABC: A new artificial bee colony algorithm for binary optimization.", *Applied Soft Computing*, 12(1),342-352.
- Kiran, M.S., Gündüz, M., 2003, "XOR-based artificial bee colony algorithm for binary optimization.", *Turkish Journal of Electrical Engineering & Computer Sciences*, 21(2), 2307-2328.
- Klamroth, K., Margaret, M.W., 2000, "Dynamic programming approaches to the multiple criteria knapsack problem.", *Naval Research Logistics (NRL)*, 47(1), 57-76.
- Kiran, M.S., 2015, "The continuous artificial bee colony algorithm for binary optimization.", *Applied Soft Computing*, 33, 15-23.
- Lin, G., Xu, H., Chen, X., Guan, J., 2020, "An effective binary artificial bee colony algorithm for maximum set k-covering problem.", *Expert Systems with Applications*, 161, 113717.
- Mirjalili, S., 2019, "Genetic algorithm. In *Evolutionary algorithms and neural networks* ", Springer, Cham, 43-55.
- Moradi, N., Kayvanfar, V., Rafiee, M., 2021, "An efficient population-based simulated annealing algorithm for 0-1 knapsack problem", *Engineering with Computers*, 1-20.
- Odlyzko, A.M., 1990, "The rise and fall of knapsack cryptosystems.", *Cryptology and computational number theory*, 42, 75-88.
- Ozsoydan, F.B., Baykasoglu, A., 2019, "A swarm intelligence-based algorithm for the set-union knapsack problem.", *Future Generation Computer Systems*, 93, 560-569.
- Ozsoydan, F. B., 2019, "Artificial search agents with cognitive intelligence for binary optimization problems.", *Computers & Industrial Engineering*, 136, 18-30.
- Wei, Z., Hao, J. K., 2021, "Kernel based tabu search for the Set-union Knapsack Problem", *Expert Systems with Applications*, 165, 113802.
- Wang, H., Wang, W., Xiao, S., Cui, Z., Xu, M., Zhou, X., 2020, "Improving artificial bee colony algorithm using a new neighborhood selection mechanism.", *Information Sciences*, 527, 227-240.
- Xiang, W.L., Li, Y.Z., He, R.C., An, M.Q., 2021, "Artificial bee colony algorithm with a pure crossover operation for binary optimization", *Computers & Industrial Engineering*, 152, 107011.

## Ek (Appendix)

**Ek Tablo:** Problem örnekleri üzerinde yöntemlerden elde edilen sonuçlar (Results obtained from methods on problem instances).

P. No	A-SUKP			GA			BABC			ABCBin			binDE			GPSO*			GPSO			PMABC		
	En iyi	Ort	Std	En iyi	Ort	Std	En iyi	Ort	Std	En iyi	Ort	Std	En iyi	Ort	Std	En iyi	Ort	Std	En iyi	Ort	Std	En iyi	Ort	Std
G1_1	12459	12459	0	13044	12956	130.66	13251	13029	92.63	13044	12819	153.06	13044	12991	75.95	13167	12937	190	13283	13051	37.41	13251	13056,31	44
G1_2	11119	11119	0	12066	11546	214.94	12238	12155	53.29	12238	12049	96.11	12274	12124	67.61	12210	11778	277	12274	12085	95.38	12274	12137,04	60
G1_3	11292	11292	0	13064	12493	320.03	13241	13064	99.57	12946	11862	324.65	13241	12941	205.7	13302	12766	305	13405	13287	93.18	13405	13266,85	110
G1_4	12262	12262	0	13671	12803	291.66	13829	13359	234.99	13671	12537	289.53	13671	13110	269.69	13993	12949	326	14044	13493	328.72	14215	13640,36	223
G1_5	8941	8941	0	10553	9981	142.97	10428	9995	154.03	9751	9339	158.15	10420	9899	153.18	10600	10090	236	11335	10670	227.85	11411	10702,97	163
G1_6	9432	9432	0	11016	10350	215.13	12012	10903	449.45	10913	9958	276.9	11661	10499	403.95	11935	10750	525	12245	11607	477.8	12245	11711,15	293
G1_7	9076	9076	0	10083	9642	168.94	10766	10065	241.45	9674	9188	167.08	10576	9681	275.05	10698	9947	295	11484	10916	367.75	11244	10733,08	229
G1_8	8514	8514	0	9831	9327	192.2	9649	9136	151.9	8978	8540	161.83	9649	9021	150.99	10168	9417	360	10710	9865	315.38	10328	10071,65	120
G1_9	9864	9864	0	11031	10568	123.15	10784	10452	114.35	10340	9910	120.82	10586	10364	93.39	11258	10566	260	11722	11185	322.98	11546	11195,34	142
G1_10	8299	8299	0	9472	8693	180.12	9090	8858	94.55	8789	8364	114.1	9191	8784	131.05	9759	8779	300	10022	9300	277.62	10194	9361,02	182
G2_1	13634	13634	0	14044	13806	144.91	13860	13735	70.76	13860	13547	199.11	13814	13676	119.53	13963	13740	120	14044	13855	96.23	14044	13920,20	91
G2_2	11325	11325	0	13145	12235	388.66	13508	13352	155.14	13498	13103	343.46	13407	13212	287.45	13498	12937	418	13508	13347	194.34	13508	13434,01	66
G2_3	10328	10328	0	11656	10889	237.85	11846	11194	249.58	11191	10424	197.88	11535	10969	302.52	11972	11233	369	12522	11899	391.83	12350	11890,76	211
G2_4	9784	9784	0	11792	10828	334.43	11521	10945	255.14	11287	10346	273.47	11469	10717	341.08	12167	11027	21	12317	11585	275.32	12317	11691,87	187
G2_5	10208	10208	0	12055	11755	144.45	12186	11946	127.8	11494	10922	182.63	12304	11865	160.42	12736	11934	294	12695	12411	225.8	12713	12583,80	127
G2_6	9183	9183	0	10666	10099	337.42	10382	9860	177.02	9633	9187	147.78	10382	9710	208.48	10724	9907	399	11425	10568	327.48	11093	10688,38	118
G2_7	9751	9751	0	10570	10112	157.89	10626	10101	196.99	10160	9549	141.27	10462	9976	185.57	11048	10400	282	11531	10959	274.9	11310	10861,11	140
G2_8	8497	8497	0	9235	8794	169.52	9541	9033	194.18	9033	8366	153.4	9388	8768	212.24	10264	9195	312	10927	9845	358.91	10725	9880,29	283
G2_9	9615	9615	0	10460	10185	114.19	10755	10328	91.61	10071	9738	111.64	10546	10228	103.32	10647	10205	190	10888	10681	125.36	10885	10637,93	82
G2_10	7883	7883	0	9496	8883	158.21	9318	9181	84.91	9262	9618	141.32	9312	9096	145.45	9839	9107	258	10194	9704	252.84	10176	9819,00	160
G3_1	10231	10231	0	11454	11093	171	11664	11183	184	11206	10880	164	11352	11075	119	11710	11237	169	12045	11487	138	12020	11590,33	171
G3_2	10483	10483	0	12124	11326	417	12369	12082	194	12006	11485	248	12369	11876	337	12369	11684	354	12369	11994	437	12369	12156,68	184
G3_3	11508	11508	0	12841	12237	198	13047	12523	201	12308	11668	177	13024	12278	234	13298	12514	356	13696	13204	367	13609	13307,07	136
G3_4	8621	8621	0	10920	10352	208	10602	10151	153	10376	9684	185	10547	10085	161	10856	10208	264	11298	10801	206	11298	10817,49	140
G3_5	9961	9961	0	10994	10640	127	11158	10776	117	10269	9957	141	11152	10661	150	11310	10762	199	11568	11318	183	11538	11217,25	191
G3_6	9618	9618	0	11093	10190	250	10528	9898	187	10051	9424	197	10528	9832	233	11226	10309	389	11517	10899	300	11590	11042,90	217
G3_7	8672	8672	0	9799	9433	164	10085	9538	185	9235	8905	112	9883	9315	192	8971	9552	234	10483	10013	202	10397	9955,62	105
G3_8	8064	8064	0	9173	8704	154	9456	9090	157	8932	8407	149	9352	8847	211	9389	8881	283	10338	9525	286	9865	9439,48	141
G3_9	9559	9559	0	10311	9993	118	10823	10483	228	10537	9615	151	10728	10159	198	10595	10145	200	11094	10688	168	11018	10603,61	126
G3_10	8157	8157	0	9329	8849	142	9333	9086	116	8799	8348	123	99218	8920	169	9807	8917	267	10104	9383	241	9686	9389,84	125



# Makine Öğrenmesi Algoritmaları Kullanarak Erken Dönemde Diyabet Hastalığı Riskinin Araştırılması

Gürkan BİLGİN<sup>1\*</sup> 

<sup>1</sup>Burdur Mehmet Akif Üniversitesi, Elektrik Elektronik Mühendisliği Bölümü, Burdur, Türkiye

## Öz

Diyabet insanoğlunun yaşam kalitesini önemli derecede etkileyen, dünyada ve Türkiye’de görülme sıklığı giderek artan önemli bir hastalıktır. Özellikle sinir sistemi, böbrek, kalp, gözler, uzuvlar ve kan damarlarının tahribatına yol açmakta ve önemli kayıplara sebebiyet verebilmektedir. Bu sebeple diyabetin önlenmesi veya vereceği tahribatın en aza indirilebilmesi için erken tanısı ve takibi büyük önem kazanmaktadır. Makine öğrenme algoritmaları ile elde edilen sınıflandırma teknikleri, hastalığın risk tahmin modeli için araştırmacılar tarafından önemli olarak kabul görmüştür. Çalışmada, diyabete yakalanma olasılığını tahmin etmek için, 520 denekten alınan bilgiler ile oluşturulmuş olan bir veri tabanı kullanılmıştır. Çalışmada, makine öğrenmesi metotları olarak Çok Katmanlı Algılayıcı Yapay Sinir Ağları (ÇKAYSA), Destek Vektör Makinaları (DVM), Karar Ağaçları (KA), Topluluk Öğrenme Algoritmaları (TÖA), Doğrusal Ayrımcı Analizi (DAA), k-NN Metotları kullanılmıştır. Bu metotlar arasında en yüksek doğruluğu k-NN algoritması sağlamış ve bu algoritma ile %99,81 doğruluk elde edilmiştir. En yüksek doğruluk değeri sağlayan algoritmanın çalışma kapsamında geliştirilmiş olan bir bilgisayar kullanıcı arayüzü içerisine dâhil edilmesiyle bir diyabet erken tanı kiti geliştirilmiştir.

**Anahtar kelimeler:** Diyabet riski, Erken dönem, Destek Vektör Makinesi, Topluluk Öğrenme Algoritmaları, Karar ağaçları.

## Investigation of The Risk of Diabetes in Early Period using Machine Learning Algorithms

### Abstract

Diabetes significantly affecting the quality of human life, the world and the incidence of a disease in Turkey is increasingly important. In particular, it causes damage to the nervous system, kidney, heart, eyes, limbs and blood vessels and can cause significant losses. For this reason, early diagnosis and follow-up is of great importance in order to prevent diabetes or to minimize the damage it will cause. Classification techniques obtained by machine learning algorithms have been accepted as important by researchers for the risk prediction model of the disease. In the study, a database created with information from 520 subjects was used to estimate the probability of developing diabetes. In the study, Multilayer Perceptron Artificial Neural Networks (MLPNN), Support Vector Machines (SVM), Decision Trees (DT), Ensemble Learning Algorithms (ELA), Linear Discriminant Analysis (LDA), k-NN Methods were used as machine learning methods. Among these methods, k-NN algorithm provided the highest accuracy and 99,81% accuracy was achieved with this algorithm. A diabetes early diagnosis kit was developed by including the algorithm providing the highest accuracy value into a computer user interface developed within the scope of the study.

**Keywords:** Diabetes risk, Early stage, Support Vector Machine, Ensembles Learning Algorithms, Decision Trees.

### 1. Giriş (Introduction)

İnsan vücudundaki insülin hormonunun gereğinden az ya da yetersiz üretilmesi ve üretilen insülinin etkili kullanılamaması gibi durumlarda kronik bir rahatsızlık durumu ortaya çıkar. Bu rahatsızlık Diyabet (Şeker

Hastalığı) olarak adlandırılır. Diyabet insanoğlunun yaşam kalitesini önemli derecede etkileyen, dünyada ve Türkiye’de görülme sıklığı giderek artan önemli bir hastalıktır (TEMĐ, 2013; TSHGM, 2011). Diyabet hastalığı organ kayıplarına ve ölümlere sebep olabilmektedir. Özellikle sinir sistemi, böbrek, kalp, gözler, uzuvlar ve kan damarlarının tahribatına yol

\* Sorumlu yazar. Gürkan Bilgin  
E-posta adresi: gbilgin@mehmetakif.edu.tr

Alındı : 9 Şubat 2021  
Kabul : 7 Mart 2021

açmakta ve önemli kayıplara sebebiyet verebilmektedir (TEMD, 2013; Harding vd., 2019; Islam vd., 2020). Bu sebeple diyabetin önlenmesi veya vereceği tahribatın en aza indirilebilmesi için erken tanısı ve takibi büyük önem kazanmaktadır.

Genellikle tip 1 ve tip 2 diyabet olarak adlandırılan iki tür diyabet biliyoruz. Tip 1 diyabet, bağışıklık sistemi yanlışlıkla pankreas beta hücrelerine saldırdığında ve vücuda çok az insülin salındığında veya bazen vücuda hiç insülin salınmadığında ortaya çıkar. Öte yandan tip 2 diyabet, vücudumuz uygun insülin üretmediğinde veya vücut insüline dirençli hale geldiğinde ortaya çıkar. Bazı araştırmacılar diyabeti tip 1, tip 2 ve gestasyonel diyabet olarak ikiye ayırdı (The 6 Different Types of Diabetes, 2018). Diyabetin yaygın semptomları poliüri, polidipsi, polifaji, ani kilo kaybı (genellikle tip 1), halsizlik, obezite (genellikle tip 2), gecikmiş iyileşme, görsel bulanıklık, kaşıntı, sinirlilik, genital pamukçuk, kısmi parezi, kas sertliği ve alopesidir. vb. (The 6 Different Types of Diabetes, 2018; Statistics About Diabetes, 2018).

Diğer taraftan, Diabetes Australia, Harris vd. tarafından daha önce 12 yıla kadar belirtildiği gibi, diyabetin klinik tanıdan 7 yıl öncesine kadar var olabileceğini yayınlamışlardır (Diabetes Australia, 2018; Harris vd., 2018). Bununla beraber, diabetes Australia'ya göre, Tip 2'nin erken teşhisinde başarısız olunması Avustralya sağlık sistemine her yıl 700 milyon dolardan fazlaya mal olabilmektedir (Diabetes Australia, 2018). Düşük ve orta gelirli ülkeler, yaygınlığı endişe verici bir oranda artan diyabet gibi bu kadar maliyetli bir hastalığı yönetmenin yükünü kaldıramaz. Bu nedenle, erken teşhis ve uygun terapötik tedavinin başlatılması, hasta sonuçlarında çok önemli bir rol oynayabilir ve gayri safi ulusal harcamaları ve üretim kaybını azaltabilir.

Dolayısıyla, diyabetin erken tanısının konulabilmesi ve gerekli önlemlerin alınabilmesini, laboratuvar ortamlarında alınan kan örneklerinin, hekimlerin kontrol ve takibinde değerlendirilmesi ile gerçekleştirilmektedir. Fakat kan örneklerinin yanı sıra diyabet hastalığının belirtisi olan farklı semptomlar da mevcuttur. Hastalardan elde edilen bu semptomların dijital ortamlarda işlenerek değerlendirilmesi için veri setleri oluşturulmaktadır. Oluşturulan veri setlerinin erken tanı için kullanılabilmesinde dijital ortamdaki sınıflandırma yaklaşımları bir hayli önem kazanmaktadır. Veri madenciliği, tahmin için kullanılan önemli bir bilgisayar bilimi alanıdır. Veri analizi yoluyla, önceden bilinen verilerden yeni verilerin keşfedilmesi sürecidir (George, vd., 2015). Günümüzde birçok gelişmiş ve karmaşık veri seti analizlerine olanak sağlayan sistemler, algoritmalar ve metodolojiler makine öğrenmesinin tıbbi alandaki uygulamalarına olanak sağlamaktadır (Islam vd., 2020; Goldenberg ve Punthakee, 2013; Kononenko, 2001). Teknolojinin sağladığı depolanmış büyük miktardaki akıllı verilerin işlenmesinde makine öğrenmesi tekniklerinin kısa

zaman içinde yaygın olarak kullanılacağı düşünülmektedir (Ogurtsova vd., 2017).

Gerçekleştirilen bu çalışmada, hekimler tarafından erken tanısı konulmaya çalışılan söz konusu diyabet durumu erken dönemde tespit edilmeye çalışılmış, elde edilen veri setleri farklı makine öğrenme metotlarına uygulanarak doğru sınıflandırmanın en yüksek oranda elde edilmesi hedeflenmiş ve metotların karşılaştırmaları yapılmıştır. Çalışmada, makine öğrenmesi metotları olarak Çok Katmanlı Algılayıcı Yapay Sinir Ağları (ÇKAYSA), Destek Vekör Makinaları (DVM), Karar Ağaçları (KA), Topluluk Öğrenme Algoritmaları (TÖA), Doğrusal Ayrımcı Analizi (DAA), k-NN Metotları kullanılmıştır. En yüksek doğruluk değeri sağlayan algoritmanın çalışma kapsamında geliştirilmiş olan bir bilgisayar kullanıcı arayüzü içerisine dâhil edilmesiyle bir diyabet erken tanı kiti geliştirilmesi planlanmıştır.

Sağlık sektöründe gelişen teknoloji ile birçok çalışmada makine öğrenme algoritmaları kullanılarak farklı sağlık problemlerine çözüm üretmektedir. Herhangi bir sınıflandırma problemine çözüm üretmek için geleneksel algoritmaların zor olduğu veya mümkün olmadığı pek çok durumda makine öğrenmesi belirleyici olmuştur (Özer, 2020). Makine öğrenmesinin en önemli özelliklerinden birisi doğrusal olmayan ve çok karmaşık veri setlerini kullanarak, kararlı ve performansı yüksek tahminler yapabilmesidir (Hastie vd., 2009). Literatürdeki tıbbi tanı ile ilgili yapılan çalışmalardan birisinde, sınıflandırmanın tıp alanındaki tanı pratiğine rehberlik ettiğini ifade edilmiştir (Jutel, 2011). Bir başka çalışmada hastaları risk gruplarına göre ayırabilmek için hasta verileri kullanılmış ve analizlerin gerçekleştirilmesi için sınıflandırma modeli önerilmiştir (Parikh vd., 2016). Kullanılan yüksek boyutlu veri setlerinin güvenilirliğinin ve etkinliğinin artırılarak işlevsel hale getirilmesinde yapay zekâ metodları etkin rol oynamaktadır (Cichosz vd., 2015; Tran vd., 2019; Zhong vd., 2019; Bishop 2006). Örneğin bir akciğer kanseri alt tiplerinin belirlenmesinde nefes örnekleri kullanılmış ve sınıflandırma için k-NN algoritması kullanılmıştır. Sonuçta yüksek başarı oranlarına ulaşılmıştır (Wang vd., 2020).

Diyabet hastalığının erken tanısı için makine öğrenmesi ile sınıflandırma yapmak yüksek doğrulukla sonuçlar vermektedir. Sınıflandırmaların yapılabilmesi için çok farklı makine öğrenmesi algoritmaları kullanılmaktadır. Literatürdeki farklı diyabet hastalığının tanısı için yapılan çalışmalara bakılırsa, 25-78 yaş aralığındaki 250 adet hastadan elde edilen veri seti, Sapon vd. tarafından yapılan bir çalışmada Bayesian Regülasyon algoritmasında kullanılmıştır ve %88,8 oranında bir doğruluk başarısına ulaşılmıştır (Sapon vd., 2011). 2013 yılında gerçekleştirilen bir başka diyabet tahmini çalışmasında farklı metodolojiler kullanılıp karşılaştırmalar yapılmıştır. Karşılaştırmalarda kısmi En Küçük Kareler yöntemi (EKK)- DAA kombinasyonu %74 ile en yüksek doğrulukta tahmin değeri üretmiştir (Karthikeyani ve

Begum, 2013). Parashar vd. DDA-DVM kombinasyonu ve İleri Beslemeli Sinir Ağları (İBSA) sınıflandırma teknikleri kullanılarak karşılaştırmalar yapılmıştır. DVM kombinasyonu doğruluğu %75 değerine ulaşmıştır (Parashar vd., 2014). Ahmed tarafından yapılan çalışmada insülin, ilaç tedavisi ve diyet gibi diyabetik tedavi planlarını sınıflandırmak için yeni bir model geliştirilmiştir. 318 medikal kayıt ile gerçekleştirilen çalışmada 9 farklı öznelik kullanılmıştır. Çalışma sonunda geliştirilen J48 algoritması ile %70,8 başarımlı değeri elde edilmiştir (Ahmed, 2016). Mercaldo vd. diyabet teşhisine yardımcı olmak ve hızlandırmak için, Dünya Sağlık Örgütü kriterlerine göre seçilmiş bir dizi özellik kullanılarak diyabetten etkilenen hastaları sınıflandırabilen bir yöntem önermiştir (Mercaldo vd., 2017). Mercaldo vd. çalışmalarında diyabet tanısı ve sınıflandırılma için Hoefding Tree algoritmasını kullanmışlardır. Diyabet tanısı konusunda klinik veriler kullanılarak gerçekleştirilen bir başka çalışmada, diyabetin tahmini, komplikasyonları, genetik arka plan ve çevre açısından sistematik bir inceleme yapılmıştır. Çeşitli makine öğrenme algoritmaları kullanılmış, DVM en başarılı ve yaygın olarak kullanılan algoritma olarak ortaya çıkmıştır (Kavakiotis vd., 2017). 2018 yılında Joshi ve Chawan tarafından yapılan çalışmada glikoz, yaş, kan basıncı ve vücut kitle indeksi gibi 7 farklı öznelik kullanılarak diyabet tahmini gerçekleştirilmiştir. Analizler için makine öğrenmesi algoritması olarak DVM, lojistik Regresyon ve YSA kullanmışlardır. Sınıflandırma performanslarına bakıldığında en iyi sonuçlar DVM ile elde etmişlerdir (Joshi ve Chawan, 2018). Aynı yıl yapılan bir başka Kaur ve Kumari Ulusal Diyabet, Sindirim ve Böbrek Hastalıkları Enstitüsüne ait Pima Hindistan nüfuslu en az 25 kadın hastanın oluşturduğu veri kümesi ile çalışmıştır. Çalışmada Gözetimli (Supervised) Makine Öğrenmesi algoritmaları ile tahmin çalışmaları yapmışlardır (Kaur ve Kumari, 2018). Al Helal vd. diyabet tanısında farklı sınıflandırma algoritmaları kullanmışlar ve sonuçlarını k-NN ile %66,19, Naive Bayes ile %72,66 ve Rastgele Orman (RO) ile %73,72 olarak tespit etmişlerdir (Al Helal vd., 2019). Farklı bir çalışmada, belirlenmiş sınıflandırma algoritmaları iki farklı veri setine uygulanmış veri setlerinin doğruluk yüzdesine etkisini ortaya koymuşlardır. Ayrıca bahsi geçen çalışmada, Lojistik Regresyon algoritması ile %96 en yüksek doğruluk başarısına ulaşılmıştır, diğer en yüksek skorlara ise sırasıyla %94 DAA ve %93 AdaBoost algoritması ile ulaşılmıştır (Mujumdar ve Vaidehi, 2019). Güncel bir diyabet tanı çalışmasında, sayısal ve metinsel veri kümesini işlemek için İBSA modeli önerilmiştir. Bu yaklaşım ile Pima Indian Diabetes veri setinde %97,27 eğitim ve %96,09 test doğruluğu elde edilmiştir (Frimpong vd., 2021). Bi vd. çalışmalarında hastaların yaşam tarzı müdahalelerinin etkisini ölçmek için yeni bir makine öğrenimi modeli önermiştir, 6405 kadın ve 5913 erkek hastadan elde edilen veri setine uygulanmıştır. Önerilen yöntemin başarısının ÇKAYSA

ve Destek Vektör Regresyon (DVR)' a göre daha yüksek olduğu gözlemlenmiştir (Bi vd., 2021). Benzer bir başka çalışmada, 500 diyabet hastası ve 268 sağlıklı insan olmak üzere toplam 768 örnekle yapılan çalışmadan oluşturulan veri setinin otomatik sınıflandırılması için ÇKA Derin Öğrenme Algılayıcısı ve DVM yöntemleri kullanılmıştır. Derin öğrenme algılayıcı sınıflandırıcısı, diyabet veri setiyle iyi performans göstermiş %77,474 başarımlı değeri ulaşmıştır (Thaiyalnayaki, 2021). Tiwari vd. temel özelliklerle diyabet tahmini yapmak için Rastgele Orman ve Özyinelemeli Eleme yoluyla önemli özellik seçimi yapmıştır. XGBoost ve YSA kıyaslandığında XGBoost YSA' ya göre %78,91 oran ile daha yüksek doğruluğu sağlamıştır (Tiwari ve Singh, 2021). Aynı yıl yapılan farklı bir çalışmada Fan vd. diyabet hastalığı tahmini için hastaların dijital dil görüntüleri ve mide semptomlarını kullanmışlardır. Verilerin sınıflandırılmasında RO ve DVM kullanılmış, sonuç olarak RF sınıflandırma modelinin DVM sınıflandırıcı modelinden önemli ölçüde daha iyi olduğu tespit edilmiştir (Fan vd., 2021).

## 2. Materyal ve Yöntem (Material and Method)

Önerilen sistemin temel yapısı Şekil 1'de gösterilmektedir. Şekilde gösterildiği gibi, hastaların semptomları hakkındaki bilgileri içeren veri seti, çalışma kapsamında tasarlanan bir bilgisayar arayüzüne hekim tarafından girilecektir. Arayüze girilen bu bilgiler program vasıtasıyla sayısal değerlere çevrilip makine öğrenme metot girişlerine uygulanacaktır. Daha sonra veriler sırasıyla, DAA, DVM, TÖA, KA, k-NN Metotları, ÇKAYSA algoritmaları ile eğitilecek ve her birinin performansı değerlendirilecektir. Performans değerlendirmesi çapraz doğrulama, duyarlılık ve özgüllük değerleri ile gerçekleştirilecektir. Daha sonra en yüksek performans değerini sağlayan algoritma arayüz yazılımına dahil edilecektir. Burada seçilen algoritma gelen verileri tahmin edecektir. Elde edilen sonuçlar, tekrardan arayüz üzerinden çıkışta hekim ekranında gösterilecektir.



Şekil 1. Çalışmanın Genel Yapısı (The Structure of the Study)

### 2.1. Veri Tabanı (Database)

Çalışmada kullanılan veri tabanı, 520 denek üzerinde yapılan diyabetle ilgili semptomların sorgulandığı bir rapor kümesinden oluşmaktadır. Veri tabanında, diyabet belirtileri sayılabilecek semptomlara sahip insanlardan elde edilen veriler bulunmaktadır. Bu veri tabanı, yakın zamanda diyabet tanısı konmuş denekler veya henüz diyabet olmamış ve semptom

gösteren deneklere doğrudan yapılan bir anket çalışması ile oluşturulmuştur. Veriler, Bangladeş Sylhet Sylhet Diyabet Hastanesinden doğrudan anket yöntemi kullanılarak hastalardan toplanmıştır (Islam vd., 2020). Çalışmada kullanılan bu veri kümesine, UCI Makine Öğrenmesi veri tabanı üzerinden ulaşım sağlanmıştır (Dua ve Graff, 2019). Veri tabanı Tablo 1’de de gösterildiği gibi 16 özellikten oluşmaktadır. Tablonun son satırında ise diyabet riski pozitif veya diyabet riski negatif olarak gösterilmektedir.

**Tablo 1.** Özelliklerin Açıklaması (The Explanation of Attributes)

Özellikler	Değerler
Yaş	1. 20-35, 2. 36-45, 3. 46-55, 4. 56-65, 5. 65 yaşüstü
Cinsiyet	1.Erkek, 2.Kadın
Poliüri	1.Evet, 2.Hayır
Polidipsi	1.Evet, 2.Hayır
Ani kilo kaybı	1.Evet, 2.Hayır
Zayıflık	1.Evet, 2.Hayır
Polifaji	1.Evet, 2.Hayır
Genital pamukçuk	1.Evet, 2.Hayır
Görsel bulanıklık	1.Evet, 2.Hayır
Kaşıntı	1.Evet, 2.Hayır
Sinirlilik	1.Evet, 2.Hayır
Gecikmiş iyileşme	1.Evet, 2.Hayır
Kısmi parezi	1.Evet, 2.Hayır
Kas sertliği	1.Evet, 2.Hayır
Alopesi	1.Evet, 2.Hayır
Obezite	1.Evet, 2.Hayır
Sınıf	1.Pozitif, 2.Negatif.

**Şekil 2.** Diyabet Tahmini İçin Oluşturulan Veri Giriş Arayüzü (Data Input Interface Created For Diabetes Prediction)

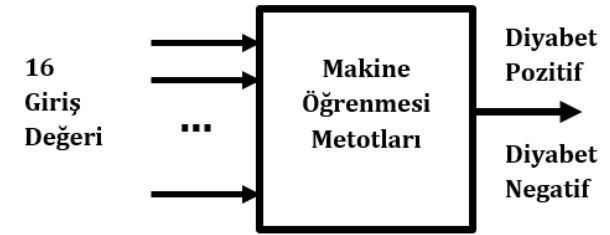
Pozitif olması hastanın diyabet olduğu, negatif olması ise denegin diyabet olmadığını ifade etmektedir. Veri tabanında 320 pozitif bulunurken 200 denekte

negatif olarak veri sağlanmıştır. Veri tabanın oluşturulurken yaş grupları da 5 ayrı şekilde ifade edilmiştir. Cinsiyet erkek ve kadın olarak yer alırken, diğer tüm özellikler ise evet veya hayır olarak değerlendirilmiştir. Bu özelliklerde ise “Evet” ifadesi semptomun bulunduğunu ifade ederken, “Hayır” seçeneği ise semptomun bulunmadığı anlamına gelmektedir. Bu özellikler ve seçenekler kullanılarak, veri girişi için bir bilgisayar arayüzü oluşturulmuştur. Bu arayüz detayları şekil 2’de gösterilmektedir.

Şekil 2’de görüldüğü üzere seçenekler sırayla yerleştirilmiş ve seçme işlemi için hekim denetimine sunulmuştur.

## 2.2. Önerilen Yöntemler (Proposed Methods)

Çalışmanın bundan sonraki kısmında ise, hastalardan elde edilen veriler, farklı makina öğrenmesi metotlarına sırasıyla uygulanmıştır. Bu metotlar DAA DVM, TÖA, KA, k-NN Metotları, ÇKAYSA algoritmaları şeklinde sıralanmaktadır. Bu metotların sisteme nasıl uyarlandığı Şekil 3’te gösterilmektedir.



**Şekil 3.** Makine Öğrenmesi Metotlarının Sisteme Uygulanması (Application of Machine Learning Methods to the System)

### 2.2.1. Doğrusal Ayrışım Analizleri (Linear Discriminant Analysis)

DAA algoritmaları genellikle modelleri iki sınıf arasında sınıflandırmak için kullanılır. Ancak, birden çok kategoride sınıflandırmak için genişletilebilir algoritmalar. Çalışmadaki kategori sayısı iki olduğu DAA algoritması yeterli olabilmektedir. DAA, tüm sınıfların doğrusal olarak ayrılabilir olduğunu varsaymaktadır. Bu varsayımdan yola çıkarak, çoklu doğrusal ayırım işlevine göre, özellik uzayında birkaç hiper düzlemi temsil eden sınıfları ayırt etmek için oluşturulur. Çalışmada sadece iki sınıf olduğu için, DAA bir düzlem çizer ve iki kategorinin ayrılmasını optimum seviyeye ulaştıracak şekilde verileri bu alt düzleme yansıtır. Bu alt düzlem, aynı anda ele alınan iki kritere göre oluşturulur: bunlardan birincisi, iki sınıfın ortalamaları arasındaki mesafeyi maksimize etmek; ikincisi ise, her kategori arasındaki varyasyonu en aza indirmektir (Vaibhaw vd., 2020). Çalışmadaki DAA modeli üzerinde, Doğrusal, quadratik, diyagonal doğrusal, diyagonal quadratik, psödo-doğrusal ve psödo-quadratik ayrışım tipleri uygulanmış ve en başarılı sonuç quadratik tipi ile elde edilmiştir.

### 2.2.2. Destek Vektör Makinaları (Support Vector Machines)

DVM, bir sınıfın tüm veri noktalarını başka bir sınıfın tüm noktalarından ayıran en iyi hiper düzlemi algılayarak verileri sınıflandırmaktadır. Optimum seviyedeki hiper düzlem seçimi, iki kategori arasındaki en büyük marjın olduğu anlamına gelmektedir. Marjın, iç kısmında bir veri noktası olmayan bir hiper düzleme paralel olan maksimum genişlik anlamına gelmektedir. Destek vektörleri ise, hiper düzlem ayırımına en yakın verilerden geçen noktalarlardır. DVM'da, sınıflandırma fonksiyonu denklem 1'deki şekilde ifade edilebilmektedir (Mohammed vd., 2016; Huang vd., 2018).

$$f(z) = \text{sign}[\sum_{j=1}^n a_j y_j x_j z + b] \quad (1)$$

Denklem (1) 'de,  $x_j$  parametreleri destek vektörlerini ifade ederken,  $a_j$  parametreleri optimum Lagrange çarpanlarını,  $y_j$  parametreleri ise sınıf etiketleri olarak (-1) veya (+1)'e eşittir. B değişkeni ise fonksiyonun eğilim (bias) parametresi olarak ifade edilmektedir. Metot üzerinde, optimal bir hiper düzlem tanımlanması için, marjın genişliğinin maksimum seviyeye çekilmesi gerekmektedir. Sonuçlar, doğrusal, Gauss fonksiyonu, Radyal Tabanlı fonksiyon ve farklı derecelerde polinom fonksiyonları da dâhil olmak üzere dört farklı çekirdek fonksiyonu kullanılarak test edilmiştir. Çalışma kapsamında, en iyi sonuç Radyal Tabanlı fonksiyon ile elde edilmiştir. Marjı ihlal eden gözlemlere uygulanan maksimum cezayı kontrol eden yumuşak-marj sabiti (YMS), 10-5'ten 10<sup>5</sup>'e arasında seçilerek Bayes optimizasyon algoritması olarak seçilmiştir. Çalışmada kullanılan YMS parametre değeri 22,956 olarak seçilmiştir. Diğer taraftan, çekirdek (Kernel) ölçeği (ÇÖ) 10-5 ile 10<sup>5</sup> arasında ayarlanır ve aynı şekilde Bayesian optimizasyon algoritması kullanılarak seçilmiştir. Optimizasyon sonucunda ÇÖ ise 0,53067 olarak belirlenmiştir.

### 2.2.3. Topluluk Öğrenme Algoritmaları (Ensemble Learning Algorithms)

Topluluk öğrenme algoritmaları, tahmine dayalı analitik çalışmalarda en başarılı makine öğrenme algoritmalarından biridir. Bu tür algoritmalar somut bir problemin çözümü için geliştirilmiş modellerin bir araya gelmesi ile oluşmaktadır. Bu tür modellerin bir araya gelmesindeki temel amaç doğruluk değerlerinin artırılmasıdır. TÖA temel olarak, torbalama (Bagging), artırma (Boosting), istif (Stacking) şeklinde ortaya çıkmaktadır. Çalışmada farklı yöntemler test edilmiştir. Bu metotlar sırasıyla, Bag, Subspace, AdaBoostM1, AdaBoostM2, GentleBoost, LogitBoost, LPBoost, RobustBoost, RUSBoost, TotalBoost yöntemleridir. Bu yöntemler arasında en başarılı sonuç üreten metot Uyarlanabilir lojistik regresyon (LogitBoost) yöntemi olarak gözlemlenmiştir. LogitBoost yöntemi, Jerome Friedman, Trevor Hastie ve Robert Tibshirani tarafından

formüle edilmiş bir yükseltme algoritmasıdır (Friedman vd., 2000). Logitboost algoritması aşağıda ifade edilen üç adımda gerçekleştirilir.

Adım 1: Ağırlık vektörü ve  $F(x)$  sınıflandırma fonksiyonu  $w_i = \frac{1}{N}$   $i = 1, 2, \dots, N$ ,  $F(x) = 0$  ve olasılık ise  $p(x_i) = \frac{1}{2}$  olsun.

Adım 2: Bu işlem  $m=1, 2, \dots, M$  için tekrar yazılırsa;

- Çalışma cevabı ve ağırlıklarının hesaplanması;

$$z_i = \frac{y_i^* - p(x_i)}{p(x_i)(1-p(x_i))} \quad (2)$$

$$w_i = p(x_i)(1 - p(x_i)) \quad (3)$$

Şeklinde olsun.

- $f_m(x)$  fonksiyonu,  $w_i$  ağırlıklarını kullanarak  $z_i$ ' den  $x_i$ ' ye ağırlıklı en küçük kareler regresyonu ile uydurulur.
- Daha sonra ise  $F(x) \leftarrow F(x) + \frac{1}{2} f_m(x)$  ve  $p(x) \leftarrow (e^{F(x)}) / (e^{F(x)} + e^{-F(x)})$  olacak şekilde güncellenir.

Adım 3: Sonunda ise sınıflandırıcı çıkışı aşağıdaki şekilde ifade edilir.

$$\text{sign}[F(x)] = \text{sign}[\sum_{m=1}^M f_m(x)] \quad (4)$$

Çalışmada, Logitboost metodunun öğrenme hızı deneysel olarak test edilmiş ve en iyi sonuç 0,5 öğrenme hızı ile elde edilmiştir.

### 2.2.4. Karar Ağaçları (Decision Trees)

Karar ağaçlarının temel yapısı düğümler, dallar ve yapraklar olarak adlandırılan üç temel bölümden oluşmaktadır. Bu ağaç yapısında, her özellik bir düğüm ile temsil edilmektedir. Dallar ve yapraklar ise, ağaç yapısının diğer bileşenlerini oluşturmaktadır. KA yapısının temel prensibi, veriyi küçük parçalara bölerek en kısa sürede elde etmektir. KA'da, ağaçtaki her bir düğüm bir sınıfı belirtmekte veya düğümdeki test verilerini oluşturan olası çıktılara göre numune alanını ayırarak bir test bölümü oluşturmaktadır. Bölünen her bir alt küme, yeni bir alt ağaçla çözülerek yeni bir alt sınıflandırma problemi ortaya çıkaracaktır. Yapraklar olarak adlandırılan düğümler, sonuç düğümünün sınıfını içerir. Yaprak düğümü olmayan noktalar ise karar düğümleri olarak tanımlanır. Bu karar düğümleri yeni bir özellik oluşturur ve bu nedenle, bu özelliklerin her olası değeri için aynı ağacın dallarını bölerek başka bir karar ağacı oluşturur (Rokach ve Maimon, 2005). Minimum yaprak düğüm sayısı 1 ile 20 arasında ve

maksimum dal düğüm sayısı ise 1 ile 39 arasında seçilir. Bölme kriteri ise, Gini'nin Çeşitlilik İndeksi (GÇİ) sapma parametreleri kullanılarak Bayesian optimizasyon algoritması ile belirlenmektedir.

#### 2.2.5. k-NN Metodu (k-NN Method)

K-NN algoritması, en basit ve temel makine öğrenimi algoritmalarından birisi olarak gösterilmektedir. Metot içindeki k değeri, tam sayı olarak en yakın özellikleri seçmek için belirlenmektedir. Test verileri ile eğitim verileri arasındaki mesafe Euclid, Manhattan, Chebyshev, Minkowski veya Hamming mesafe yöntemleriyle ölçülmektedir. Test verilerine en yakın k-eğitim verileri arasında en yakın komşular arasında en yaygın olan sınıflandırmaya dahil edilmiştir (Mohammed vd., 2016). Bu çalışmada, Bayes optimizasyon algoritması ve 30 tekrardan sonra en düşük hatayı veren değerler sınıflandırıcı parametresi kullanılarak 1'den 5'e kadar değişen k değerleri optimize edilmiş ve optimizasyon sonrası k değeri 1 olarak seçilmiştir. Diğer taraftan mesafe yöntemi de optimizasyon sonrası Minkowski mesafe yöntemi olarak belirlenmiştir. En başarılı doğruluk bu değerler ile elde edilmiştir.

#### 2.2.6. Çok Katmanlı Algılayıcı Yapay Sinir Ağları (Multilayer Perceptron Artificial Neural Networks)

Yapay Sinir Ağları, örneklerden olaylar arasındaki ilişkileri belirleyen ve daha önce hiç görmedikleri örnekler hakkında öğrendikleri bilgileri kullanan, birbirine bağlı yapay sinir hücrelerinden oluşan bilgisayar sistemleri olarak tanımlanır. YSA'nın sahip olduğu bilgi, her bir proses elemanı ile olan bağlantılarda ağırlık değerleri ile ağda depolanır ve ağa yayılır (Hu ve Hwang, 2002). Çalışma sırasında farklı YSA metotları test edilmiş ve en başarılı doğruluk değerini veren ÇKAYSA yapısı çalışmaya dahil edilmiştir. ÇKAYSA bir giriş katmanından, bir veya daha fazla gizli katman ve bir çıkış katmanından oluşmaktadır (Hu ve Hwang, 2002). ÇKAYSA'nın çıktısı ise aşağıdaki matematiksel formül ile ifade edilebilmektedir.

$$y = f\left(\sum_{i=1}^N X_i W_{ij} + b_j\right), (j = 1, 2, \dots, M) \quad (5)$$

Bu denklemde M katman sayısıdır, giriş katmanındaki nöron sayısı N ile gösterilir,  $X_i$  gizli bir katmandaki i. Nöronu ifade eder,  $W_{ij}$  her giriş için ağırlık olarak açıklanır,  $b_j$  algılayıcı, f aktivasyon fonksiyonudur ve y j. katmandaki algılayıcı çıktısıdır. Çalışmada ÇKAYSA için farklı katman sayıları farklı nöron sayıları ve aktivasyon fonksiyonları test edilmiştir. En yüksek doğruluk 2 katmanlı her iki katmanında 10 nöron bulunan ÇKAYSA yapısı ile elde edilmiştir. Her iki katman ve çıkış katmanında tanjant sigmoid fonksiyonu kullanılmıştır. Tespit edilen özellikler farklı YSA yapıları ile analiz edilmiş ve

sonuçlar gözlemlenmiştir. ÇKAYSA'da ağırlık matrisini seçmek için geri yayılım eğitim hatası yöntemi olarak Levenberg-Marquardt algoritması kullanılmıştır.

### 3. Deneysel Sonuçlar (Experimental Results)

Çalışmada, makine öğrenmesi metotları kullanılmış ve doğruluk değerleri birbirleri ile kıyaslanmıştır. Kıyaslama ölçütlerinde bilindik doğrulama yöntemleri kullanılmıştır. Buna göre, doğrulama için 3, 5 ve 10 kat çapraz doğrulama (KÇD) işlemleri gerçekleştirilmiş ve her doğrulama sonucunda duyarlılık, özgüllük ve doğruluk değerleri hem tüm veri tabanı için hem de sadece test verileri için kaydedilmiştir. Çalışmada hesaplanan duyarlılık, özgüllük ve doğruluk değerleri şu şekilde hesaplanmaktadır.

$$\text{Duyarlılık}(\%) = \frac{DP}{DP+YN} \times 100 \quad (6)$$

$$\text{Ozgulluk}(\%) = \frac{DN}{DN+YP} \times 100 \quad (7)$$

$$\text{Dogruluk}(\%) = \frac{DP+DN}{DN+DP+YN+YP} \times 100 \quad (8)$$

Bu formüllerde geçen parametreler aşağıdaki gibi tanımlanmaktadır;

- DP: Gerçekte pozitif sınıf içinde, tahminde ise pozitif sınıf içinde yer alan değer sayısı.
- YN: Gerçekte pozitif sınıf içinde, tahminde ise negatif sınıf içinde yer alan değer sayısı.
- YP: Gerçekte negatif sınıf içinde, tahminde ise pozitif sınıf içinde yer alan değer sayısı.
- DN: Gerçekte negatif sınıf içinde, tahminde ise negatif sınıf içinde yer alan değer sayısı.

Çalışmada kullanılan farklı makine öğrenmesi yöntemlerinin detaylı doğruluk bilgisi ile performanslarının karşılaştırılması Tablo 2'de gösterilmiştir.

Sonuçlar karşılaştırıldığında DVM metodunun giriş verileri göz önünde bulundurulduğunda diyabet riskinin pozitif veya negatif olacağı yönündeki değerlendirmesinde çok etkili bir metot olduğu görülmektedir. Çapraz doğrulama işlemi 3-KÇD, 5-KÇD ve 10-KÇD olarak değerlendirilmiştir. Ayrıca sonuçlar duyarlılık ve özgüllük yönünden de değerlendirilmiştir.



**Tablo 2.** Metotların k-Kat Çapraz Doğrulama Karşılaştırmaları (k-Fold Cross Validation Comparisons of Methods)

KÇD	METOT	VERİLERİN TÜMÜ			TEST VERİLERİ		
		Duy. (%)	Özg. (%)	Doğ. (%)	Duy. (%)	Özg. (%)	Doğ. (%)
3-KÇD	DAA	93,76	96,59	94,74	90,51	95,67	92,12
	DVM	98,76	97,72	98,33	96,60	93,53	95,19
	TÖA	99,48	97,86	98,85	98,41	93,76	96,54
	KA	96,57	93,03	95,13	94,54	89,87	92,50
	k-NN	99,79	97,87	99,04	99,36	93,85	97,12
	ÇKAYSA	99,06	97,86	98,59	97,17	93,71	95,77
5-KÇD	DAA	96,71	95,76	96,35	94,57	95,42	94,81
	DVM	99,37	98,51	99,04	96,86	92,75	95,19
	TÖA	99,56	99,40	99,50	97,86	97,01	97,50
	KA	97,29	94,64	96,12	94,47	90,19	91,92
	k-NN	99,87	99,01	99,54	99,37	95,26	97,69
	ÇKAYSA	99,24	93,29	96,27	97,40	88,31	93,08
10-KÇD	DAA	96,89	95,53	96,37	94,48	95,25	94,42
	DVM	99,66	99,50	99,60	96,77	95,21	95,96
	TÖA	99,81	99,65	99,75	98,17	96,63	97,50
	KA	98,12	95,60	97,12	96,56	91,62	94,42
	k-NN	99,91	99,65	99,81	99,07	96,66	98,08
	ÇKAYSA	99,75	99,30	99,58	97,46	93,42	95,77

Her KÇD bölümünde elde edilen duyarlılık ve özgüllük değerleri kabul edilebilecek seviyede olup gayet iyi sonuç vermişlerdir. Diğer taraftan doğruluk değerleri karşılaştırıldığında en yüksek tahmin doğruluk değerinin her KÇD bölümünde k-NN metoduna ait olduğu ve %99,81 şeklinde doğruluk elde edildiği aşikârdır. Dolayısıyla, çalışmada gerçekleştirilen optimizasyon teknikleri ve deneysel gözlemler diyabet riski sınıflandırılmasında k-NN yönteminin çok etkili olduğunu ortaya koymaktadır. Tablo 2'ye göre, DVM den başka, DVM ve TÖA yöntemlerinin de gayet başarılı sonuçlar verdiği görülmektedir. Diğer taraftan bakıldığında ise DAA yönteminin bu tür bir sınıflandırmada diğer metotlara göre hayli yetersiz kaldığı görülmektedir.

Bu çalışmada elde edilen sonuçlar literatürde gerçekleştirilmiş olan diğer çalışmalar ile karşılaştırıldığında, DVM yönteminin hayli başarılı bir yöntem olduğu burada da ortaya çıkmaktadır. Sonuçların diğer metotlarla karşılaştırılabilmesi açısından kesinlik ve F1 skoru değerleri de hesaplanmıştır. Bu değerler aşağıdaki formüller ile ifade edilebilmektedir.

$$Kesinlik(\%) = \frac{DP}{DP+YP} \times 100 \quad (9)$$

$$F1 \text{ skoru}(\%) = 2 \times \frac{Kesinlik \times Duyarlılık}{Kesinlik + Duyarlılık} \quad (10)$$

Elde edilen kesinlik, hassasiyet, F1 ve doğruluk skoru değerleri literatürdeki diğer çalışmalar ile karşılaştırılmış ve Tablo 3'te gösterilmiştir.

Bildirilen çalışmaların büyük çoğunluğunda, diyabet riskinin öngörülmesinde sınıflandırma doğruluğunu %80'lerin üzerinde olduğunu belirtmek gerekir. Tabloda görüldüğü üzere, İslam vd., 4 farklı metot kullanarak, en yüksek doğruluğu Rasgele orman (RO) yöntemi ile elde etmişler ve bu yöntemle %97,4 doğruluğu yakalamışlardır (İslam vd., 2020). Diğer taraftan Özer, yaptığı çalışmada, uzun kısa dönem bellek ağlarını (LSTM) kullanarak aynı veri seti üzerinde %98,9 başarı elde etmişlerdir. Frimpong vd. 2021' deki çalışmasında İleri Beslemeli YSA modeli kullanarak %96,09 doğruluk ve %94,88 F1 skorunu yakalamıştır (Frimpong vd., 2021). Bunlara karşın, gerçekleştirilen bu çalışmada ise k-NN yöntemi kullanılmış ve bu yöntem ile %99,81 oranında bir başarı elde edilmiştir. Ayrıca %99,83 F1 skoru yakalama başarısı gösterilmiştir. Ayrıca, alışılmış öznitelikler kullanılarak yapılan sınıflandırma çalışmalarından farklı olarak Fan ve arkadaşları dil görüntülerinden elde ettikleri öznitelikleri kullanarak yaptıkları çalışmalarında 3 birleşik özellik seti kullanmıştır.

**Tablo 3.** Metotların k-Kat Çapraz Doğrulama Karşılaştırmaları (k-Fold Cross Validation Comparisons of Methods)

Çalışma	Yöntem	Has. (%)	Kes. (%)	F1 Skoru (%)	Doğ. (%)
(Kaur ve Kumari, 2018)	DVM-Lineer	87	88	87	89
	k-NN	90	87	88	88
	YSA	88	85	86	86
	RBF-DVM	83	85	83	84
	MDR	83	82	84	83
(Mujumdar ve Vaidehi, 2019)	Lojistik Regresyon	-	-	-	96
	DAA	-	-	-	94
	Gradient Boost Sınıflandırıcı	-	-	-	93
	AdaBoost Sınıflandırıcı	-	-	-	93
	Gaussian Naive Bayes	-	-	-	93
	Rastgele Orman	-	-	-	91
(Islam vd., 2020)	Naive Bayes	87,4	87,9	87,5	-
	Lojistik Regresyon	92,4	92,4	92,4	-
	J48	95,6	95,7	95,6	-
	Rasgele Orman	97,4	97,4	97,4	-
(Özer, 2020)	LSTM	99,37	98,43	98,9	-
(Frimpong, 2021)	İleri Beslemeli YSA	93,28	96,53	94,88	96,09
(Fan vd., 2021)	Rastgele Orman	93,5	-	-	94,2
	DVM-Lineer	79,8	-	-	88,8
(Tiwari ve Singh, 2021)	XGBoost Sınıflandırıcı	59,33	-	-	78,91
	YSA	45,22	-	-	71,35
<b>Bu çalışma</b>	<b>k-NN</b>	<b>99,91</b>	<b>99,75</b>	<b>99,83</b>	<b>99,81</b>

Rastgele Orman (%94,2) ve DVM-Lineer (%88,8) metotlarında en yüksek skorlara ulaşmışlardır (Fan vd., 2021).

Birçok sınıflandırma probleminde olduğu gibi diyabetin erken tanısı için gerçekleştirilen bu çalışmada da veri setlerinin niteliği ve miktarı önem kazanmaktadır. Güncel bir çalışmada 9 farklı öznelik kullanılmış XGBoost Sınıflandırıcı ve YSA yöntemleri ile yüksek başarılar elde edilmiştir. XGBoost ile 78,91 ve YSA ile 71,35' tir (Tiwari ve Singh, 2021). Kaur ve Kumari yapmış olduğu çalışmada makine öğrenmesi için 8 öznelik kullanmış, en yüksek doğruluk değerlerine sırasıyla DVM-Lineer ve k-NN metodları ile %89 ve %88 elde etmiştir. Aynı çalışmada en yüksek F1 skoru değeri k-NN ile %88' dir (Kaur ve Kumari, 2018). Bir başka çalışmada Mujumbar ve Vaidehi, 10 farklı öznelik kullanmış en yüksek skorları %96, %94 ve %93 olarak sırasıyla Lojistik Regresyon, DAA ve Gradient Boost Sınıflandırıcı algoritmaları ile elde etmişlerdir (Mujumbar vd., 2019). Buna karşın, bu çalışma, öznelik çıkarımı için 16 farklı özellik kullanılarak ön plana çıkmış ve %99,81 doğruluk oranına ulaşmıştır. Bu sonuçlar ile yalnız kullanılan sınıflandırma metotlarının değil, öznelik çıkarımı için kullanılan özelliklerin niteliğinin de sonuçta ne kadar etkili olduğu ortaya konulmuştur.

Diyabet riski erken tahmininde, kişilerin konumu, yaşı veya eğitim geçmişi ne olursa olsun, bu kişilere küresel olarak erişilebilir bir online sistem tasarımı şüphesiz ki önemli bir katkı olacaktır. Güncel teknolojide, internet, veri ve hizmet aramanın en yaygın

yolu haline gelmektedir. Diyabet riskinin erken tespitinde, kullanıcı semptomlarının sisteme girilmesini sağlayacak basit bir web sitesi yapılabilir. Bu web sitesi hem diyabete yakalanma olasılığı tahminlerini yapabilir hem de diyabetik belirtiler ile ilgili bazı yararlı sağlık ipuçları sağlayabilir nitelikte olabilir. Bu ihtiyaç için çalışma kapsamında geliştirilen bir arayüz çalışması ana sayfası Şekil 4'te gösterilmektedir.

**Şekil 4.** Demo Arayüz Tasarımı (Demo Interface Design)

#### 4. Sonuç ve Tartışma (Result and Discussion)

Farklı yaş gruplarından insanlar her geçen gün diyabet riski ile karşı karşıyadır. Bu çalışmada, diyabetin erken aşamada tespit edilmesinin tedavide çok önemli bir rol oynadığını ve erken teşhis için ipuçlarının yakalanmasının çok önemli olduğu vurgulanmaktadır. Beslenmede şeker dengesinin sağlanması, düzenli fiziksel aktivite gerçekleştirilmesi ve sağlıklı yaşam tarzına geçiş yapmak gibi bazı basit bilgilendirmeler kısmen de olsa obeziteyi önleyebilecek imkânlar sağlayabilecektir. Diyabeti tahmin etmek için geliştirilen makine öğrenmesi yöntemleri ve araçları geliştirilince ve bu yöntemlerin girişine uygulanacak veri sayısı ve denek sayısı arttıkça sistemin gerçekliği ve erken tanı ve teşhise maliyeti düşük bir şekilde vereceği destekte, bu tıbbi sağlık hizmetindeki rolünü yadsınamaz hale getirecektir. Bu çalışmanın literatüre olan temel katkısı, diyabetik riskinin erken tahmini için yapılan en iyi algoritmayı bulmaktır. Çalışmada, k-NN algoritmasının yüzdelerik çapraz doğrulama değerlendirme testinde en iyi doğruluk performansı sergilediği gösterilmiştir. Çalışma sonucunda, k-NN algoritmasının doğruluk performans değeri %99,81 olarak hesaplanmıştır. Bununla beraber F1 skoru % 99,83 olarak elde edilmiş ve diğer çalışmalara göre üstünlüğü ifade edilmiştir. Diğer taraftan ise, son kullanıcı tarafında küçük anket tarzı bir sorgulama ile bilgisayar arayüzü üzerinden diyabet ön tahmini yapılması sağlanmıştır.

#### Teşekkür (Acknowledgment)

Bu çalışma Burdur Mehmet Akif Ersoy Üniversitesi Bilimsel Araştırma Projeleri Koordinasyon Birimi tarafından desteklenmiştir.

#### Kaynaklar (References)

Ahmed, T. M., 2016. Developing a predicted model for diabetes type 2 treatment plans by using data mining. *Journal of Theoretical and Applied Information Technology*, 90(2), 181.

Al Helal, M., Chowdhury, A. I., Islam, A., Ahmed, E., Mahmud, M. S., & Hossain, S., 2019. An optimization approach to improve classification performance in cancer and diabetes prediction. In 2019 International Conference on Electrical, Computer and Communication Engineering (ECCE), pp. 1-5, IEEE.

Bi, S., Ding, X., Yu, S., Guo, B., Mu, L., Wang, B., 2021. A machine learning model for quantifying the effect of lifestyle interventions for patients with type 2 diabetes mellitus. In *Journal of Physics, Conference Series*, Vol. 1732, No. 1, p. 012006, IOP Publishing.

Bishop, C. M., 2006. *Pattern Recognition and Machine Learning* Springer-Verlag New York. Inc. Secaucus, NJ, USA.

Cichosz, S. L., Johansen, M. D., Hejlesen, O., 2016. Toward big data analytics: review of predictive models in

management of diabetes and its complications. *Journal of diabetes science and technology*, 10(1), 27-34.

Dua, D., Graff, C., 2019. UCI Machine Learning Repository [http://archive.ics.uci.edu/ml]. Irvine, CA: University of California, School of Information and Computer Science.

Failure to detect type 2 diabetes early costing \$700 million per year, *Diabetes Australia*, 8 July 2018. <https://www.diabetesaustralia.com.au>

Fan, S., Chen, B., Zhang, X., Hu, X., Bao, L., Yang, X., Liu, Z., Yu, Y., 2021. Machine Learning Algorithms in Classifying TCM Tongue Features in Diabetes Mellitus and Symptoms of Gastric Disease. *European Journal of Integrative Medicine*, 101288.

Friedman, J., Hastie, T., Tibshirani, R., 2000. Additive logistic regression: a statistical view of boosting. *Annals of Statistics*, 28 (2), 337–407.

Frimpong, E. A., Oluwasanmi, A., Baagyere, E. Y., Zhiguang, Q., 2021. A feedforward artificial neural network model for classification and detection of type 2 diabetes. In *Journal of Physics: Conference Series*, Vol. 1734, No. 1, p. 012026, IOP Publishing.

George, T., Rufus, E., Alex, Z.C., 2015. Simulation of microwave induced thermo-acoustical imaging technique for cancer detection. *Journal of Engineering and Applied Sciences (ARPN)*, 10.

Goetz, T., 2010. *The Decision Tree: Taking Control of Your Health in the New Era of Personalized Medicine*, New York, NY, USA: Rodale.

Goldenberg, R., Punthakee, Z., 2013. Definition, classification and diagnosis of diabetes, prediabetes and metabolic syndrome. *Canadian journal of diabetes*, 37, S8-S11.

Harding, J. L., Pavkov, M. E., Magliano, D. J., Shaw, J. E., Gregg, E. W., 2019. Global trends in diabetes complications: a review of current evidence. *Diabetologia*, 62(1), 3-16.

Harris, M.I., Klein, R., Welborn, T.A., Knudman, M. W., 1992. Onset of NIDDM occurs at least 4–7 yr before clinical diagnosis. *Diabetes Care*, 15(7), 815–819.

Hastie, T., Tibshirani, R., Friedman, J., 2009. *The elements of statistical learning: data mining, inference, and prediction*.

Hu, H.Y., Hwang, J.N., 2002. *Handbook of Neural Network Signal Processing*, New York, NY, USA: CRC Press.

Huang, S., Cai, N., Pacheco, P. P., Narrandes, S., Wang, Y., Xu, W., 2018. Applications of support vector machine (SVM) learning in cancer genomics. *Cancer Genomics-Proteomics*, 15(1), 41-51.

Islam, M. F., Ferdousi, R., Rahman, S., Bushra, H. Y., 2020. Likelihood prediction of diabetes at early stage using data mining techniques. In *Computer Vision and Machine Intelligence in Medical Image Analysis* pp. 113-125, Springer, Singapore.

Joshi, T. N., Chawan, P. P. M., 2018. Diabetes Prediction Using Machine Learning Techniques. *International Journal of Engineering Research and Application (Ijera)*, vol. 8, no.1, pp. 9-13, 2018.

Jutel, A., 2011. Classification, disease, and diagnosis. *Perspectives in biology and medicine*, 54(2), 189-205.

Karthikeyani, V., Begum, I. P., 2013. Comparison a performance of data mining algorithms (CPDMA) in prediction of diabetes disease. *International journal on computer science and engineering*, 5(3), 205.

- Kaur, H., Kumari, V., 2018. Predictive modelling and analytics for diabetes using a machine learning approach. *Applied Computing and Informatics*.
- Kavakiotis, I., Tsave, O., Salifoglou, A., Maglaveras, N., Vlahavas, I., Chouvarda, I., 2017. Machine learning and data mining methods in diabetes research. *Computational and structural biotechnology journal*, 15, 104-116.
- Kononenko, I., 2001. Machine learning for medical diagnosis: history, state of the art and perspective. *Artificial Intelligence in Medicine*, vol. 23, no. 1, pp. 89-109.
- Mercaldo, F., Nardone, V., Santone, A., 2017. Diabetes Mellitus Affected Patients Classification and Diagnosis through Machine Learning Techniques. *Procedia Computer Science*, 112, 2519-2528.
- Mohammed, M., Khan, M. B., Bashier, E. B. M., 2016. *Machine learning: algorithms and applications*. Crc Press.
- Mujumdar, A., Vaidehi, V., 2019. Diabetes Prediction Using Machine Learning Algorithms. *Procedia Computer Science*, 165, 292-299.
- Ogurtsova, K., da Rocha Fernandes, J. D., Huang Y., Linnenkamp, U., Guariguata, L., Cho, N. H., Cavan, D., Shaw, J. D., Makaroff, L. E., 2017. IDF Diabetes Atlas: global estimates for the prevalence of diabetes for 2015 and 2040. *Diabetes research and clinical practice*, 128, 40-50.
- Özer, İ., 2020. Uzun Kısa Dönem Bellek Ağlarını Kullanarak Erken Aşama Diyabet Tahmini. *Mühendislik Bilimleri ve Araştırmaları Dergisi*, 2 (2) , 50-57.
- Parashar, A., Burse, K., Rawat, K., 2014. A Comparative approach for Pima Indians diabetes diagnosis using lda-support vector machine and feed forward neural network. *International Journal of Advanced Research in Computer Science and Software Engineering*, 4(11), 378-383.
- Parikh, R. B., Kakad, M., Bates, DW., 2016. Integrating predictive analytics into high-value care: the dawn of precision delivery. *JAMA*, 315, 651-652.
- Rokach, L., Maimon, O., 2005. Decision trees. In *Data mining and knowledge discovery handbook*, pp. 165-192, Springer, Boston, MA.
- Sapon, M. A., Ismail, K., Zainudin, S., 2011. Prediction of diabetes by using artificial neural network. In *Proceedings of the 2011 International Conference on Circuits, System and Simulation*, Singapore Vol. 2829.
- Specht, D., 1991. A general regression neural network, *IEEE Trans. Neural Netw.* 2, 568-576.
- Statistics About Diabetes: American Diabetes Association, 22 Mar 2018. <https://www.diabetes.org>.
- Thaiyalnayaki, K., 2021. Classification of Diabetes Using Deep Learning and SVM Techniques. *International Journal of Current Research and Review*, Vol, 13(01), 146.
- The 6 Different Types of Diabetes: (5 Mar 2018). *The diabetic journey*. [https:// thediabeticjourney.com/the-6-different-types-of-diabetes](https://thediabeticjourney.com/the-6-different-types-of-diabetes).
- Tiwari, P., Singh, V., 2021. Diabetes disease prediction using significant attribute selection and classification approach. In *Journal of Physics: Conference Series*, Vol. 1714, No. 1, p. 012013.
- Tran, B. X., Latkin, C. A., Giang, V. T., Huong, L. T. N., Son, N., Ming-Xuan, T., Zhi-Kai, L., Cyrus, S. H. H., Roger, C. M. H., 2019. The Current Research Landscape of the Application of Artificial Intelligence in Managing Cerebrovascular and Heart Diseases: A Bibliometric and Content Analysis. *International Journal of Environmental Research and Public Health*, 16,2699.
- Türk Endokrinoloji ve Metabolizma Derneği (TEMED), 2013. *Diabetes mellitus ve komplikasyonlarının tanı, tedavi ve izlem kılavuzu (6.baskı)*. Ankara, BAYT Bilimsel Araştırmalar Basın Yayın, 2012, 15-42.
- TC Sağlık Bakanlığı Temel Sağlık Hizmetleri Genel Müdürlüğü (TSHGM), 2011. *Türkiye Diyabet Önleme ve Kontrol Programı Eylem Planı (2011- 2014)*. Ankara: Sağlık Bakanlığı Yayın No:816.
- Vaibhaw, Jay Sarraf, P.K. Pattnaik, Chapter 2 - Brain-computer interfaces and their applications, Editor(s): Valentina Emilia Balas, Vijender Kumar Solanki, Raghendra Kumar, *An Industrial IoT Approach for Pharmaceutical Industry Growth*, Academic Press, 2020, Pages 31-54, ISBN 9780128213261.
- Wang, C., Long, Y., Li, W., Dai, W., Xie, S., Liu, Y., Zhang, Y., Liu, M., Tian, Y., Li, Q., Duan, Y., 2020. Exploratory study on classification of lung cancer subtypes through a combined K-nearest neighbor classifier in breathomics. *Scientific reports*, 10(1), 1-12.
- Zhong, G., Ling, X., Wang, L. N., 2019. From shallow feature learning to deep learning: Benefits from the width and depth of deep architectures. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 9(1), e1255.