e-ISSN: 2667-4165 • CİLT / VOLUME: III • SAYI / ISSUE: II • ARALIK / DECEMBER 2020

# AFYON KOCATEPE ÜNİVERSİTESİ ULUSLARARASI MÜHENDİSLİK TEKNOLOJİLERİ VE UYGULAMALI BİLİMLER DERGİSİ

# Afyon Kocatepe University International Journal of Engineering Technology and Applied Sciences

www.dergipark.org.tr/tr/pub/akuumubd

AFYON KOCATEPE ÜNİVERSİTESİ
ULUSLARARASI MÜHENDİSLİK TEKNOLOJİLERİ ve UYGULAMALI BİLİMLER DERGİSİ
Afyon Kocatepe University
International Journal of Engineering Technology and Applied Sciences

# Afyon Kocatepe University International Journal of Engineering Technology and Applied Sciences

https://dergipark.org.tr/tr/pub/akuumubd

www.ijetas.aku.edu.tr

**e-ISSN:2667-4165**

# Afyon Kocatepe University
# International Journal of Engineering Technology and Applied Sciences (AKU-IJETAS)

Volume: 3 / Number: 2 / December - 2020

***Owner / Publisher: Rector Prof. Dr. Mehmet KARAKAŞ for Afyon Kocatepe University***

***Editor in Chief Prof. Dr. Ayhan EROL***

***Co- Editor in Chief Assoc. Prof. Dr. Ahmet YÖNETKEN***

*Published Afyon Kocatepe University, December 2020,*

*ijetas@aku.edu.tr*



**CONTACT INFORMATION**

Afyon Kocatepe University International Journal of Engineering Technology and Applied Science Afyon Kocatepe University, Technology Faculty, 03200 Afyonkarahisar, TURKEY

Phone: +90-272-2281446 /ext.

Fax: +90-272 228 1449

e-mail : ijetas@aku.edu.tr, aerol@aku.edu.tr

# Welcome to AKU-IJETAS

Dear Researchers;

International Journal of Engineering and Applied Sciences ler has been published in Turkish and English since 2018 with 2 issues. Our journal will accept Turkish and English articles as 2 issues a year and the articles will be evaluated by at least two referees with the same system. Our magazine from December 2018; it offers many advantages to readers due to the practical and practical access to the authors as well as the process of publishing and publishing quickly and easily; The electronic journal (e-ISSN:2667-4165) accepts 2 numbers per year (June and December) in Turkish and English. The names of the judges evaluating the articles are not notified to the authors. The referees cannot see the names of the authors. The studies are evaluated as at least two referees. Our authors, who want to send articles, can register their original scientific articles online and follow the process by registering on our magazine page. Our journal is accepted as original and previously published research articles.

We are waiting for your contributions as both referee and writer. I thank you in advance for your support and I wish you success in your work.

**Prof. Dr Ayhan EROL**

**Chief Editor**

## Editörler/ Editorial Board

| | | |
|---|---|---|
| Adem KURT | Gazi University | TURKEY |
| Ahmet YILDIZ | Afyon Kocatepe University | TURKEY |
| Hazizan Md AKİL | Sains Malaysia University | MALAYSIA |
| Huseyin Ali YALIM | Afyon Kocatepe University | TURKEY |
| Huseyin AKBULUT | Afyon Kocatepe University | TURKEY |
| Muhammed YURUSOY | Afyon Kocatepe University | TURKEY |
| Mustaque HOSSAIN | Kansas State University, Manhattan | ABD |
| Ramazan KAÇAR | Karabük University | TURKEY |
| Rıdvan UNAL | Afyon Kocatepe University | TURKEY |
| Suleyman GUNDUZ | Karabük University | TURKEY |
| Ugur CALIGULU | Firat University | TURKEY |
| Yuksel OĞUZ | Afyon Kocatepe University | TURKEY |

## Danışma Kurulu / Advisory Board

| | | |
|---|---|---|
| Ahmet AKSOY | Akdeniz University | TURKEY |
| Alexander ONUFRAK | Pavol Jozef Safarik University | SLOVAKIA |
| Anas Sarwar QURESHI | Agriculture University | PAKISTAN |
| Artay YAGCI | Afyon Kocatepe University | TURKEY |
| Behçet GULENC | Gazi University | TURKEY |
| Bojan ZLENDER | Maribor University | SLOVENIA |
| Cahit GURER | Afyon Kocatepe University, | TURKEY |
| Dinçer BURAN | Süleyman Demirel University | TURKEY |
| Dunja PERIC | Kansas State University, Manhattan | ABD |
| Dusan ORAC | Kosice Technical University | SLOVAKIA |
| Elena Cristina RADA | Trento University | ITALY |
| Gabor PAY | University College of Nyiregyhaza | HUNGARY |
| Gratiela BOCA DANA | Technical University Cluj Napoca | ROMANIA |
| Huseyin BAYRAKCEKEN | Afyon Kocatepe University | TURKEY |
| Ioan ABRUDAN | Technical University Cluj Napoca | ROMANIA |
| Ivan KURIK, | Technical University Zilina | SLOVAKIA |
| Iveta VASKOVA | Kosice Technical University | SLOVAKIA |
| João Pedro SILVA | Leiria Polytechnic Institute | PORTUGAL |
| Lucian Ionel CIOCA | Lucian Blaga University of Sibiu | ROMANIA |
| Marco RAGAZZI | Trento University | ITALY |
| Martina HRUBOVCAKOVA | Kosice Technical University | SLOVAKIA |
| Matjaž ŠRAML | Maribor University | SLOVENIA |
| Merlinda EBIBI | Mother Teresa University | MACEDONIA |
| Mihai BANICA | Technical University Cluj Napoca | ROMANIA |
| Mircea HORGOS | Technical University Cluj Napoca | ROMANIA |
| Monica Lopez ALONSO | University of GRANADA | SPAIN |

| | | |
|---|---|---|
| Mustaque HOSSAIN | Kansas State University, Manhattan | ABD |
| Nadras OTHMAN | Sains University | MALAYSIA |
| Nicolae UNGUREANU | Technical University Cluj Napoca | ROMANIA |
| Neritan TURKESHI | Mother Teresa University | MACEDONIA |
| Olivera PETKOVSKA | Mother Teresa University | MACEDONIA |
| Olga OROSOVA | Pavol Jozef Safarik University | SLOVAKIA |
| Otar ZUMBURIDZE | GeorgiaTechnical University | GEORGIA |
| P. Trinatha RAO | Gitam University | INDIA |
| Peter MONKA | Technical University Kosice | SLOVAKIA |
| Prasanna RAMAKRISNAN | Neo Education Institu | MALAYSIA |
| Radu COTETIU | Technical University Cluj Napoca | ROMANIA |
| Regita BENDIKIENĖ | Kaunas Technology University | LİTVANIA |
| Renata PANOCOVA | Pavol Jozef Safarik Üniversity | SLOVAKIA |
| Robert CEP | Technical University Ostrava | CZECH |
| Serdar SALMAN | Marmara University | TURKEY |
| Serhat BASPINAR | Afyon Kocatepe University | TURKEY |
| Sermin OZAN | Fırat University | TURKEY |
| Sezai TAŞKIN | Celal Bayar University | TURKEY |
| Stanislaw LEGUTKO | Poznan University of Technology | POLAND |
| Tomasz NIZNIKOWSKI | Lomza State University Applied Science | POLAND |
| Tomaz TOLLAZZI | Maribor University | SLOVENIA |
| Yılmaz YALCIN | Afyon Kocatepe University | TURKEY |
| Zoran TRIFUNOV | Mother Teresa University | MACEDONIA |

**CONTENTS**

Afyon Kocatepe Üniversitesi Uluslararası Mühendislik Teknolojileri ve Uygulamalı Bilimler Dergisi

Afyon Kocatepe University International Journal of Engineering Technology and Applied Sciences

AKÜ IJETAS Cilt **3(2)** (2020) Aralık (45- 51pp) AKU J.Eng.App.Sci. **Vol 3(2)** (2020) December (45- 51pp)

Araştırma Makalesi / Research Article

*e-ISSN 2667-4165 (https://dergipark.org.tr/akuumubd)*

# Geri Dönüştürülmüş Al – Si Pistonlarda Mikroyapılarının Aşınma Etkisi

**B. N. G. Aliemeke[1] and M. H. Oladeinde[2]**

[1]*Department of Mechanical Engineering, Auchi Polytechnic, Nigeria*
[2]*Department of Production Engineering, University of Benin, Nigeria*
*eposta: aliemeka@yahoo.com, https://orcid.org/0000 0003 2681 3790*
*eposta: moladeinde@uniben.edu https://orcid.org/0000 0002 5030*

The arrival date:29.05.2020 ; Date of Acceptance:23.07.2020

**Anahtar Kelimeler**
Aşınma Oranı; Mikroyapı; mikrograf; α-alüminyum.

**Öz**

Aşınma, bir otomobil motor sisteminde olumsuz koşulları başlatma eğilimindedir. Piston, bir motordaki şiddetli termal gerilimlere dayanan önemli bir motor bileşenidir. Çevresel rahatsızlık oluşturan hurda pistonlar geri dönüştürülecek. 1,15 µg / m olarak belirlenen pistonların aşınma oranının azaltılmış bir değerinin, optimum çalışma koşulu için gerekli olan tokluğu ve sertliği koruyan gelişmiş bir mikro yapıya sahip olduğu fark edilmiştir. 6.04 µg / m'lik yüksek aşınma hızı değeri, daha az belirgin iğne şeklinde ötektik silikon parçacıklar üreten pistonların mikrografını vermiştir. Hurda jeneratör pistonlarından üretilen alüminyum alaşımlı döküm pistonun mikrografları. Sonuçlar piston alaşımının yapısal matrisinin arka planında birincil a-alüminyum parçacıklarını göstermektedir. Ayrıca, ötektik silikon partiküllerine yakın iğne şeklindeki partiküllerin ithal edilen piston alaşımlarının mikroyapısında dağıldığı fark edilir.

# Effect on Wear of the Microstructure of Recycled Al –Si Pistons

***Keywords***
Wear rate; Microstructure; micrograph ; α-aluminium.

**Summary**

Wear has a tendency of initiating adverse conditions in an automobile engine system. The piston is an important engine component which withstands severe thermal stresses in an engine. Scrap pistons which constitute environmental nuisance will be recycled. A reduced value of wear rate of the pistons determined to be 1.15µg/m was noticed to have an improved microstructure which retains toughness and hardness required for optimal working condition. While wear rate high value of 6.04µg/m yielded micrograph of pistons which produced less pronounced needle shaped eutectic silicon paricles. The micrographs of the aluminium alloy cast piston produced from scraps generator pistons. The result shows primary α-aluminium particles at the background of the structural matrix of the piston alloy. Also, needle shaped near eutectic silicon particles are noticed to be dispersed in the microstructure of the imported piston alloys.



## 1. Introduction

Wear is a major challenge in the automobile industry and its direct cost is estimated to be between 1 and 4% of the gross national product ( Agarwal, Parnaik and Sharma, 2013). Its effect can initiate adverse conditions in an automobile engine syste (Ameen, Hassan and Mubarak, 2011). To put forth great resistance to abrasive and sliding wear it will be important to have engine components produced in aluminium-silicon alloys (Hassan et al, 2011). So much effort and techniques has been expended to manufacture more durable materials to reduce the effect of wear on tools and engineering components.

Aluminium silicon alloys are reputed for great advantages like corrosion resistance, high thermal conductivity, good weldability and excellent castability (Dell, 2009). For many decades, engine pistons were manufactured from cast iron which was used for producing other engine components (Heuer, 2015). There is a departure in the usage of cast iron because of the improved mechanical properties of aluminium inherent in modern engineering (Yang, 2003).

Presently, automobile engine pistons are mostly manufactured from aluminium silicon alloys. In the recent past Yamaha generator piston had been produced using a Silumin aluminium alloy material which is chosen on the basis of high fatigue strength, high wear resistance and hardness (Ebhota et al,, 2015). Silumin is usually a term used in most countries for alloys based on Al-Si system. It is a series of lightweight, high-strength aluminium alloys with a silicon content within the range of 3-50%. Some of these aluminium alloys are casting ones which could be produced by rapid solidification processes and powder metallurgy. Putting into perspective the Aluminium Association designation system silumins are corresponding to alloys of two systems:which are 3xxx aluminium-silicon alloys containing magnesium and copper, and 4xxx-Binary aluminium-silicon alloys.One of the greatest advantages of silumin is its resistance to corrosion which makes it very applicable in humid environments(Vengatesvaran et al, 2018). The relevance of silicon and copper elements in eutectic aluminium alloys have been adjudged to be satisfactory in improving mechanical properties (Kumar and Grewal, 2013.

The microstructure of metallic material has the tendency of influencing physical properties such as toughness, strength, ductility, hardness, corrosion resistance and wear resistance( Manchanda, and Narang, 2005). In addition, the mechanical properties of aluminuim alloy such as strength formability, ductility, fatigue strength and surface hardness, amongst others enhances its performance in service. Studies have also shown that failure of aluminium can result from production methods, use of substandard material, poor design, manufacturing errors due to poor machining, or failure from a phenomenon called fatigue (Ajayi, 2013).

A microstrutural examination of wear rate of LM13 using centrifugal casting process was carried out by (Patel,2014). The study reveals that the silicon promotes fluidity during melting, enhances mechanical properties (tensile strength and hardness) and offer resistance to wear. The microstructural characterization of LM13 cast alloy showed presence of rod like shaped structure dispersed within the medium which accounts for toughness of the alloy (Kayser and Svendsen ,2008). The machined samples of the LM13 were tested for tensile strength, hardness, and wear rate. The result showed that the tensile strength, hardness and wear resistance of the LM13 cast alloy increase with silicon. The optimal mechanical properties and wear resistance occurred at 7000C pouring temperature and 1050 rpm mould rotation.

Xi et al. (2020) investigated The microstructure evolution and tribological property of SLM-processed AlSi10Mg/TiB2 composites. The result showed that Al-based composites with high manufacturing quality and uniform dispersion of TiB2 particles were dispersed throughout the structural matrix. Also, the composites showed high microhardness of 126 HV0.2 and wear rate of 5.2 × 10−4 mm3N−1m-1. The aim of this study is to investigate the microstructure and wear behavior of aluminum silicon alloy piston.

**2. Material ve Metod**

Optical metallurgical microscope of model L2003A having magnification strength of 400X shown in Fig.1 was employed to conduct microstructural analysis on the specimen obtained from the ingot. The specimens were dimensioned 16mm in diameter and 10mm in depth . The specimens were polished with various grade of emery clothe of P-60,P-120,P-220,P-400,P-800 and P-1200.A dry rough and fine polishing operation was carried out

using serium oxide. The etching operation was conducted by immersion and swabbing on each specimen for about 30 seconds using the Keller's reagent which constitutes 1% hydrofluoric acid, 1.5% hydrochloric acid, 2.5% trioxonitrate (v) acid and 95% distilled water (Kayser and Svendsen, 2009). The etchant was rinsed off the specimen properly with water and air blower was used to dry it properly before been introduced into the stage of the optical microscope machine. The magnification knob was adjusted to select a proper focal length between the workpiece and the magnification lens (Kayser, 2009). The microstructure of the samples was viewed on the pc screen via connected to the optical microscope.

**Fig. 1:** Optical Metallurgical Microscope

### *2.1 Wear Measurement*

The wear measurement experiment was performed on the pin on disc machine. The pin on disc machine employs three important parameters during wear test experimentation. The parameters are sliding distance, sliding speed and load.

The specimens from the ingots of cast aluminium alloy prepared in the form of a cylindrical pin with dimension 10.0mm diameter and 20.0mm length. These specimens were utilized as test samples for the experiment. The pin on disc wear machine is made up of steel disc of about 90mm in diameter and a thickness of 10mm.The cylindrical pins and the disc were thoroughly cleaned with water and dried with acetone before the commencement of the test. The steel disc was fixed on a rotating shaft which is connected to the shaft of an electric motor by means of belt and pulley. The test piece was weighed before the commencement of the experiment and properly positioned in the specimen holder of the machine. The positioned specimen is brought in contact to the flat surface disc. The machine is switched on and the shaft is made to rotate for about 50 minutes. In this experiment load of 50N was used. The sliding speed of the cylindrical pin and disc was maintained at 900 revolutions per minute. The weight of the specimen was taken using a weighing balance and the difference or loss in weight was recorded. The specimen test was repeated so as to bring about accuracy of test values gotten. The average weight loss was recorded ($M_1$-$M_2$). Wear rate values were determined by equation (1

$$Wear\ rate, W_R = \frac{M_1 - M_2}{2\pi R N_w t} \qquad (1)$$

where sliding distance=$2\pi RN_w t$, R=track radius, t= time taken, $N_w$=sliding speed, $M_1$= mass of specimen before wear experiment and $M_2$= mass of specimen after the wear experiment

**Table 1:** Chemical composition of Piston alloys

| Element | Cast piston(%) | Commercial available piston(%) |
|---|---|---|
| Si | 10.442 | 10.800 |
| Mg | 0.802 | 0.804 |
| Al | 78.984 | 83.048 |
| Ti | 0.930 | 1.022 |
| Cr | 0.009 | 0.018 |
| Mn | 0.671 | 0.765 |
| Fe | 1.888 | 1.262 |
| Ni | 1.004 | 1.012 |
| Cu | 2.538 | 2.600 |
| Zn | 1.782 | 1.785 |
| Sr | 0.800 | 0.824 |
| Pb | 0.173 | 0.222 |
| Sn | 0.027 | 0.030 |
| Sb | 0.022 | 0.106 |

The result from the test shows that the constituents of the local cast piston alloy and the commercially available piston are similar to LM 2 alloy. The local piston showed that its silicon and aluminium content are 10.44% and 79%

respectively while that of the imported piston is approximately 10.84% and 83%. Some aluminium was lost as result of formation of Theta precipitate and metal evaporation. It is apparent from the test result that the developed piston is near eutectic.

### *3.2 Wear Rate Test Result*

The wear rate test was carried out at the Material Science and Material Engineering Department, Obafemi Awolowo University, Ile-Ife, Nigeria. The wear test was conducted on a Pin-on-disc machine by keeping normal load of 50N and sliding speed of 900rpm constant. The sliding distance of the specimens were varied and mass loss recorded. The test values are shown in Table 2.

**Table 2:** Wear Test Values

| Experiment No | Random order of exp. | Initial mass(g) | Final mass(g) | Mass loss (mg) | Track diameter(mm) | Sliding distance(m) | Wear rate (µg/m) |
|---|---|---|---|---|---|---|---|
| 1 | 4 | 23.011 | 22.997 | 13.67 | 16.00 | 2262.20 | 6.04 |
| 2 | 2 | 22.010 | 21.998 | 11.31 | 15.00 | 2120.91 | 5.33 |
| 3 | 5 | 22.005 | 21.996 | 8.53 | 14.00 | 1979.50 | 4.32 |
| 4 | 7 | 20.004 | 19.997 | 7.38 | 13.50 | 1908.82 | 3.90 |
| 5 | 9 | 21.001 | 20.993 | 6.13 | 13.00 | 1838.11 | 3.34 |
| 6 | 1 | 21.003 | 20.998 | 5.03 | 12.50 | 1767.38 | 2.85 |
| 7 | 3 | 23.002 | 22.998 | 3.50 | 12.00 | 1696.71 | 2.06 |
| 8 | 6 | 22.002 | 21.999 | 2.55 | 11.00 | 1555.30 | 1.64 |
| 9 | 8 | 20.360 | 20.358 | 1.63 | 10.00 | 1413.91 | 1.15 |

The mathematical model obtained by the multilinear regression method using Minitab17 for the wear rate, $W_R$ in terms of pouring temperature A, vibration frequency B, vibration time C and runner size D is given as

$$W_R = 57.29 - 0.07213A - 0.03083B - 0.00922C - 0.000742D \quad (2)$$

### *3.3 Significance Test for the Wear Rate, $W_R$ Mathematical Model*

A statistical test of significance for the mathematical model developed for wear rate by the multiple linear regression was carried out to ascertain the relevance of the relationship between the response variable, $W_R$ and regressors, A, B, C and D. The test for significance of the wear rate regression model is shown in Table 3 and the probability plot is shown in Fig.2.

**Table 3:** Result for Significance Test for Wear Rate

| Term | Coef | SE Coef | T-value | P-value |
|---|---|---|---|---|
| Constant | 57.290 | 1.01 | 56.88 | 0.000 |
| A | -0.07213 | 0.00138 | -52.27 | 0.000 |
| B | -0.030830 | 0.00172 | -17.93 | 0.000 |
| C | -0.009220 | 0.00229 | -4.02 | 0.016 |
| D | -0.000742 | 0.00022 | -3.34 | 0.029 |

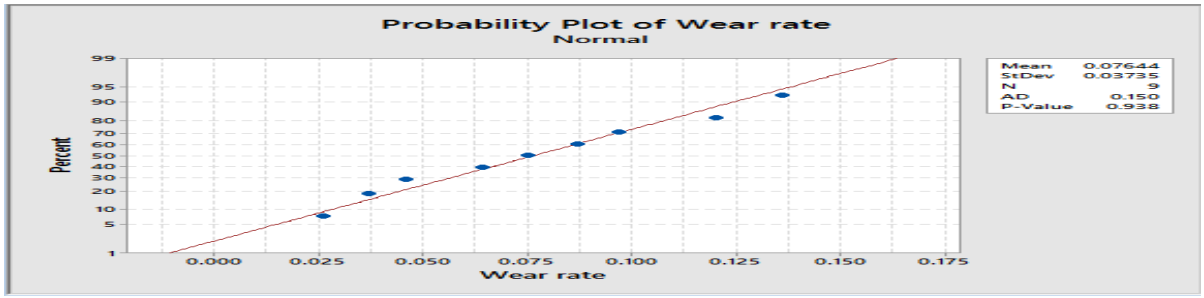



**Fig 2:** Normal Probability Plot for Wear Test Data

The Normality probability plot shown in Fig. 2 portrays that the residuals lie close to the ideal normal distribution diagonal line which interprets that the data are normally distributed. Also, the Anderson-Daling value and p-value which are 0.150 and 0.938 respectively indicate that there is insufficient evidence for any deviation and as such the normality condition have been satisfied.

### *3.4 Microstructural Result*

The microstructure experiment was conducted for the melted aluminium alloy scrap pistons. The microstructural images of the cast aluminium silicon piston alloys from each experiment are shown in Fig. 3.

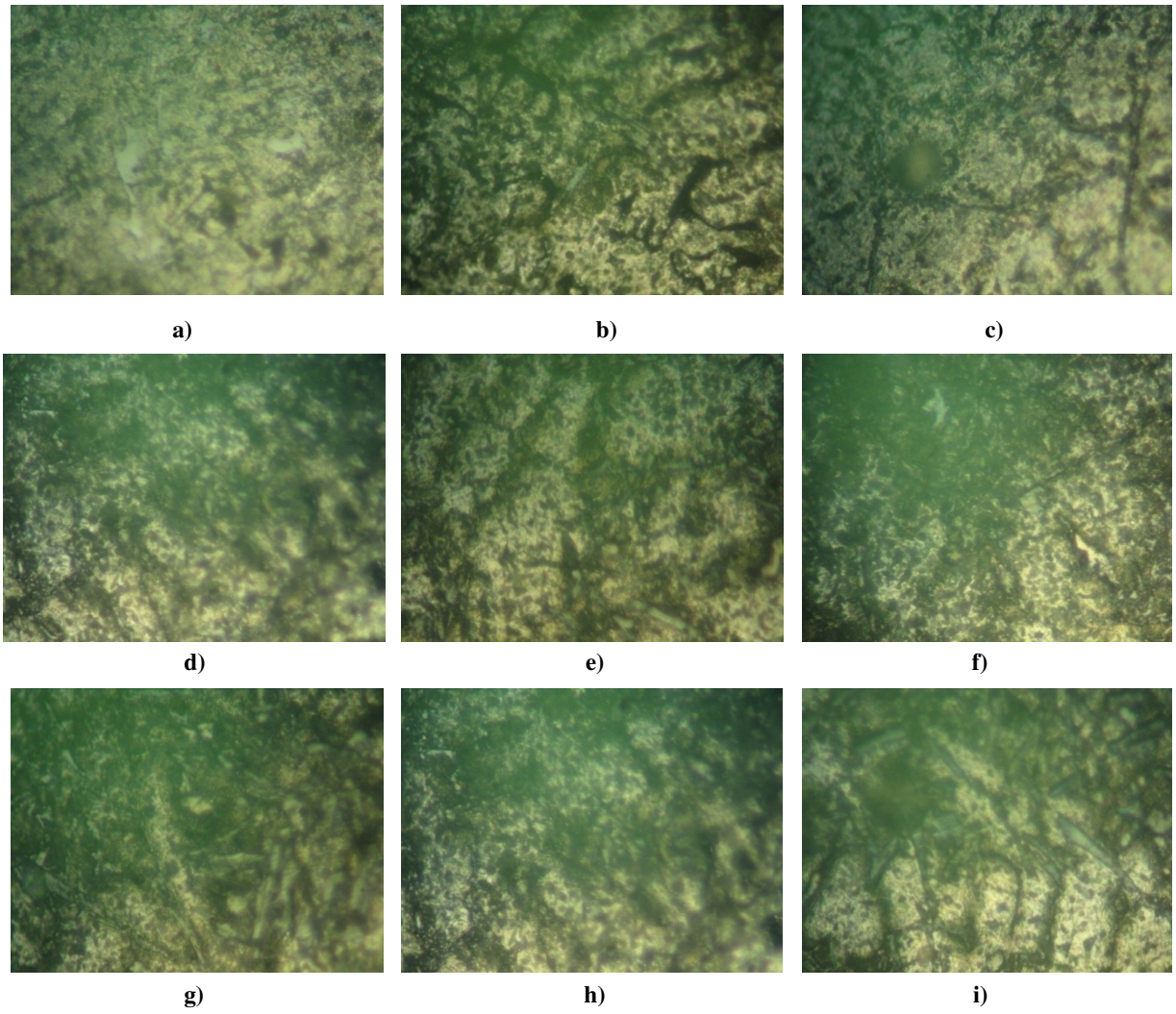



**Fig. 3:** (a-i) Micrograph of Cast Al-Si Piston Alloy obtained from melted scrap pistons from Taguchi Design Experiments (1-9) taken on magnification of 400X

The micrographs of the nine specimens show a basic microstructure which consist of primary $\alpha$-aluminium having even distribution of eutectic silicon grains and intermetallic particles dispersed within the structural matrix. It is noticed that the micrograph of Fig. 3. (a-d) had predominately primary $\alpha$-aluminium within the matrix. The microstructural view of Fig. 3(g,h and i) showed less pronounced presence of lamellar shaped eutectic silicon in the structure which accounts for toughness of aluminium alloy. The spikes are known to have been made less pronounced because of the action of wear before the metallic

piston recycling. Also the micrographs of commercially available pistons are shown in Fig.4.

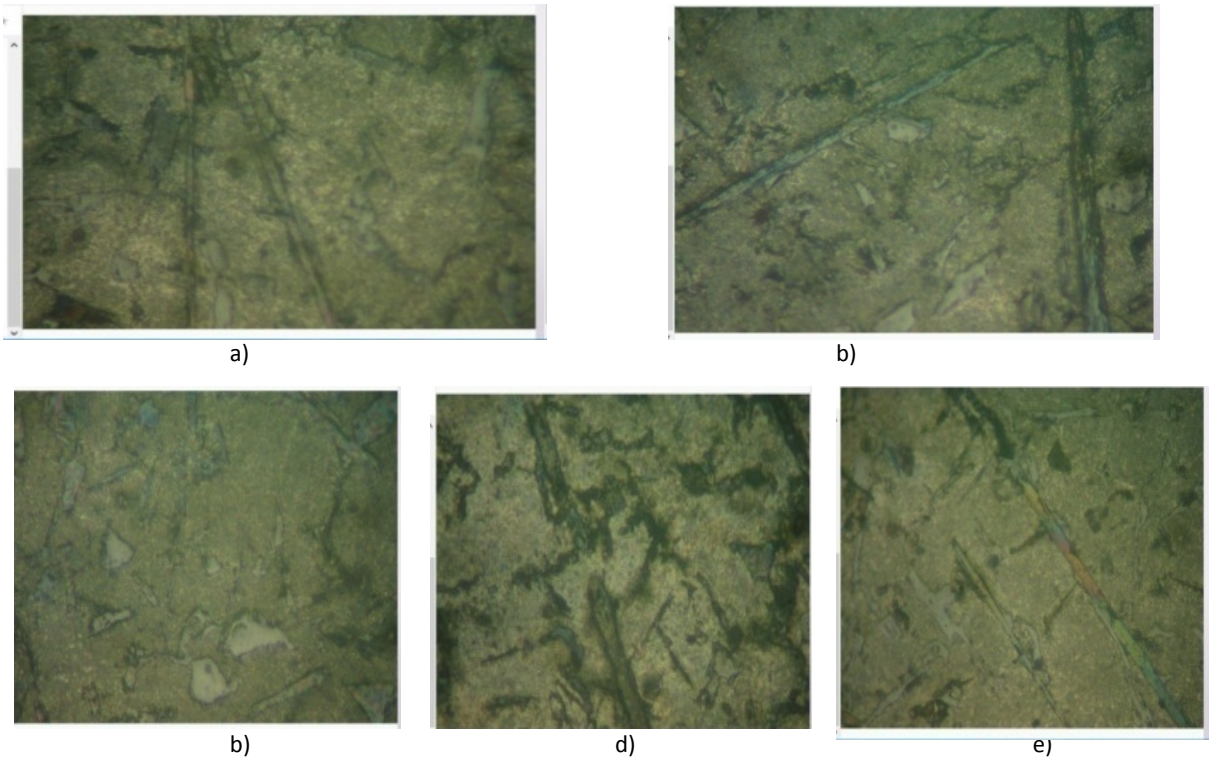



**Fig. 4**: Micrograph (400X) of Commercially available Yamaha Generator Pistons ( a-e)

The micrograph of the pistons in Figs. 4 (a, b, c, d and e) show that the primary □- aluminium particles are embedded in the structure of the aluminium alloy with uniform distribution of the intermetallic particles. The micrograph in Fig. 4 (c) shows globular near eutectic silicon contained in the structural matrix. Also a needle shaped near eutectic silicon particles are noticed to be present in the microstructure shown in Fig. 4(a, b, and d). The micrograph represented by Fig. 4 (e) depicts uniform distribution of long rod shaped eutectic silicon within the aluminium alloy.

**4. Conclusion**

The study showed that the higher the rate of wear of the piston the less pronounced the lamellar shaped spikes which accounts for the toughness of aluminum alloy. Also, the low wear rate of the piston gives rise to needle shaped eutectic silicon which connotes improved mechanical properties. The eutectic silicon particles also boosted the wear resistance strength of the pistons as seen from the chemical composition of the aluminium alloy. The aluminium metal from the melted scrap pistons was 78.98% while that of the commercially available piston was 83.334%. This occurrence is similar to the findings of Ozioko(2012) in the study of recycling motorcycle piston scraps. Also, the result is similar to that obtained by Mbuya (2010) in which the chemical composition of the melted scrap piston and commercially available pistons yielded little variation among the aluminium and the alloying elements. The result obtained in this study is similar to Kumar et al. (2020) in which the microstructure, mechanical and wear behavior under dry sliding of Silumin with particulate-reinforced Sic and TiB2 Metal matrix developed by stir casting showed excellent mechanical properties for AA6061.

## 5. Resources


Agarwal, G. , Parnaik A. and Sharma R. K., 2013. Parametric Optimization and Three-BodyAbrasive Wear Behavoir of Sic Filled Chopped Glass Fiber Reinforced Epoxy Composites. *International Journal of Composite Materials*, 3(2), 32-38.

Ajayi , J. A. , 2013. Carbon Steels in Structure and Properties of Engineering Alloys, Journal for *Materials Science and Engineering* , 3(9) 96–304.

Ameen, H. A. , Hassan, K. S. and Mubarak, E. M., 2011. Effect of loads, Sliding Speeds and Times on the Wear rate for Different Materials. *American Journal of Scientific and Industrial Research* 2(1) 99-106.

Dell, K. A., 2009, Metallurgy Theory and Practical Textbook. American Technical Society, Chicago, 20-50.

Ebhota, W. S., Ademola, E. , Abdulrahman, J. , Aduloju, S.C. and Owolabi, O. B., 2015. Designing for domestication of Yamaha CY80 Engine Piston Manufacturing Technology and Evaluation of Aluminium Alloy for Functionality*, International Journal of Advanced Scientific and Technical Research*, 1(5) 21-34.

Hassan, Z., Pandey, R. K. and Sehgal D. K., 2011. Wear characteristics in Al-SiC Particulate Composites and the Al-Si Piston Alloy, *Journal of Minerals and Materials Characterization and Engineering*, 10(14)1329-1335.

Heuer, J., 2015. Development and Testing of Carbon Pistons", Structured Materials Research Institute, Stuggart, Germany.

Kayser, T. P. and Svendsen, B. , 2008. Experimental and theoretical investigation on the microstructure of aluminum alloys during extrusion. *Proceedings on Applied Mathematics and Mechanics Conference,* 8(2)10431–10432.

Kayser,T. P. , Klusemann, F. and Svendsen, B,. 2009. Experimental and theoretical investigation of the microstructural evolution in aluminium alloys during extrusion", *Computational Methods and Experiments in Materials Characterization*, 4(2) 209–216.

Kumar, R and Grewal, C. , 2013. Improvement in Hardness of LM-6 Aluminium Green sand Castings by Taguchi Method. *Asian Journal of Engineering and Applied Technology,* 2(2)11-18.

Kumar, G. S., Reddy, M. Y. V. and Reddy, M. B. C., 2018 Experimental Investigation of Microstructure, Mechanical and Wear behavior under dry sliding of Silumin with particulate-reinforced Sic and TiB2 Metal matrix developed by stir casting, *Proceedings from International Conference on Engineering Thermodynamics and Mechanical Engineering*.

Manchanda, V. K. and Narang, G. B. S. ,2005. Materials and Metallurgy. (6th Edition), Khanna Publishers, 45- 70.

Mbuya, T. O., Odera, B. O., Ng'ang'a S. P. and Oduori F. M. ,2010, Effective Recycling of Cast Aluminium Alloys for Small Foundries. *Recycling for foundries*, 12(2)162 – 182

Ozioko, F. U. , 2012 . Synthesis and Study on Effect of Parameters on Dry Sliding Wear Characteristics of Al-Si Alloys, *Leonardo Electronic Journal of Practices and Technologies,* 20(6)39-48

Patel, V.J. ,2014. Tribological Investigation of LM 13 by Horizontal Centrifugal Casting Process, Doctoral Dissertation, Ganpat University, North Gujarat, 254

Vengatesvaran, K. Prithiviraj, N. and Periyasamy, N., 2018. Thermal Analysis and Material Optimization of Piston in I.C. Engine. International Journal for Applied Research in Innovative Ideas Engineering, 4 (3) 153-171.

Xi, L., Guo, S., Gu, D. , Guo. M. and Lin, K. , 2020. Microstructure development, tribological property and underlying mechanism of laser additive manufactured submicro-$TiB_2$ reinforced Al-based composites , Journal for Alloys and Compounds, 819(5)127-132

Yang, L. J., 2003. The Effect of Casting Temperature on the Properties of Squeeze Cast Aluminium and Zinc Alloys, *Journal of Materials Processing Technology*, 2(14) 391-396.

Afyon Kocatepe Üniversitesi Uluslararası Mühendislik Teknolojileri ve Uygulamalı Bilimler Dergisi

Afyon Kocatepe University International Journal of Engineering Technology and Applied Sciences

AKÜ IJETASCilt**3(2)** (2020) Aralık (52-59 s) AKU J.Eng.App.Sci. **Vol3(2)** (2020) Decemeber (52-59pp)

Araştırma Makalesi / Research Article

*e-ISSN 2667-4165 (https://dergipark.org.tr/akuumubd)*

# Makine Öğrenmesi Yöntemleri ile Web'den Bilgi Çıkarımı Sürecinin İyileştirilmesi

**Erkan Özhan**[1]

[1] *Tekirdağ Namık Kemal Üniversitesi, Çorlu Mühendislik Fakültesi, BilgisayarMühendisliği Bölümü, Çorlu, Tekirdağ.*

*e-posta: erkanozhan@gmail.com, ORCID ID: http://orcid.org/0000-0002-3971-2676*

Geliş Tarihi:22.08.2020 ; Kabul Tarihi:15.09.2020

***Anahtar kelimeler***
Web bilgi çıkarımı; Makine öğrenmesi; Sınıflandırma; Veri madenciliği

**Öz**

Web ortamı bilginin doğduğu, yayıldığı ve yaşadığı bir formata sahiptir. Gün geçtikte bilgi morfolojik olarak değişim geçirmekte ve bu değişimle birlikte avantajlar yanında istenilen anlamlı bilgiye ulaşmada zorluklar artmaktadır. Zaman, depolama, iletişim ve veri işleme maliyetleri açısından istenilen bilgiye en verimli şekilde ulaşmak kritik bir görevdir. Bunun yanında verinin yaşam süreci boyunca kullanılabilirliğini de artırabilir. Web sayfalarının "layout" adı verilen bölümlerinin sınıflandırılması bu sorunların çözümüne önemli katkılar sağlayabilir. Özellikle bu bölümlerdeki gereksiz içeriğin bilinmesi faydalı ve anlamlı bilgiye ulaşmayı kolaylaştırıcı ve maliyetleri düşürücü etki sağlayabilir. Bu çalışma makine öğrenmesi yöntemleri ile web sayfası bölümlerinin sınıflandırılması sürecini iyileştirmek amacıyla farklı algoritmalara odaklanmış ve bu algoritmaların iyileştirici sonuçlarını ortaya koymuştur. Elde edilen sonuçlara göre Random Forest ve KStar algoritmalarının süreci iyileştirici modeller olduğu görülmüştür. Random Forest algoritması %98.46 doğru sınıflandırma oranı sunarken, KStar hız faktörüyle öne çıkmıştır. Çalışmada karar ağacı ve entropi tabanlı algoritmaların başarımları da karşılaştırılmış ve bulgular hesaplama zamanlarıyla birlikte sunulmuştur.

# Improvingthe Information ExtractionProcessfromthe Web with Machine Learning Methods

***Keywords***
Web informationextraction; Machine learning; Classification; Data mining

**Summary**

The web environment has a format in which information is born, propagated and lived. Information changes morphologically day by day, and with this change, difficulties in reaching the desired meaningful information increase as well as advantages. It is a critical task to reach the desired information in the most efficient way in terms of time, storage, communication and data processing costs. In addition, it can increase the availability of data throughout its life cycle. Classification of the parts of web pages called "layout" can make important contributions to the solution of these problems. In particular, knowing the unnecessary content in these sections can facilitate access to useful and meaningful information and provide a cost-reducing effect. This study focuses on different algorithms in order to improve the process of classifying web page sections with machine learning methods and reveals the improvement results of these algorithms. According to the results, it has been seen that Random Forest and KStar algorithms have process improvement solutions. While the Random Forest algorithm offers 98.46% correct classification rate, KStar stands out with its speed factor. In the study, especially the performance of tree and entropy-based algorithms were compared and the findings were presented together with the computation times.

## 1. Giriş

Bilgisayarlar veya diğer cihazların iletişim ortamına dahil edilmesinin birincil nedeni bilgiyi paylaşmaktır. Bu iletişimin yönü tek yönlü olabildiği gibi çift yönlü de olabilir. Bir web sayfasındaki metin bloğunu okumak tek yönlü iken, bu yazıya yorum yapmak bir anda çift yönlü bilgi paylaşımına neden olur. Dünyada iletişim için kullanılan bu araçların çeşitliliği, yetenekleri, kapasiteleri ve iletişim-işlem hızları arttıkça iletişim sonucu ortaya çıkan bilgi yoğunluğu da dramatik bir şekilde artmaktadır. Bu yoğunluk ulaşılabilecek bilgi miktarı ve çeşitliliğinin artması anlamına gelir. Bu sonuç ilk bakışta olumlu görünse de istenilen bilgiye ulaşmak için çok sayıda ayıklama yapmanız gerekeceği anlamına gelir. Bu sorun yeni değildir ve en ilgili sonuçları ortaya çıkarmak için arama motorları geliştirilmiştir. Ancak arama motorları da devasa boyutlara ulaşmış, web kaynaklarını elde etmede ve onları değerlendirmede sorunlar yaşamaktadır. Web sayfaları yalnızca gerçek içerikten değil, aynı zamanda afişler (banner), gezinme öğeleri, reklamlar, telif hakkı vb. gibi diğer unsurlardan da oluşur (Wu, Liu and Fan, 2015). Web içerik çıkarımı (web content extraction) istenilen bilgiye veya ona en yakın olana ulaşmanın yollarını arar. Arama sonuçlarını en ilgili ve en hızlı bir şekilde ortaya çıkarmak için performans artırıcı çok sayıda teknik kullanırlar. Veri indirgenimi, haritalama, yüksek başarımlı hesaplama, veri madenciliği ve makine öğrenmesi (machine learning-ML) gibi çok sayıda fark yaratabilecek disiplinden faydalanmaktadırlar. Kısacası verinin elde edilmesi, depolanması, işlenmesi ve sunumunda iyileştirme çalıştırmaları sürekli yapılmaktadır.

İnternet ortamında bulunan veriler metin, resim, video, ses gibi birkaç farklı formatta olabilir. Bu verilerin sunumu için ise direkt veya dolaylı olarak dosyalar kullanılır. Veriler .html, .php, .asp vb. gibi dosyalar içerisinde gömülü olarak bulunabilir. Bunun yanında veri tabanlarından, sensörlerden vb. elde edildikten sonra sayfaya gömülerek yine web ortamında sunulabilir. Çoğu web sayfası; gazete, alışveriş kataloğu, başvuru formu gibi uzayıp giden bir listenin elektronik versiyonu gibidir (Duckett, 2011).

Web sayfaları layout adı verilen bölümlerden oluşur. Web öğelerinin konumlandığı bölgelere layout denir. Bu bölümlerin çeşitliliği ve sayısı oldukça değişkendir. Günümüzde web sayfası oluşturmak için çok sayıda biçimlendirme ve web programlama dili geliştirilmiştir. Web sayfaları için temel biçimlendirme dili HTML (Hypertext Markup Language)'dir. HTML aynı zamanda web sayfaları oluşturmak ve yönetmek için temel standarttır. Web sayfası (HTML belgesi), düğümlerin HTML öğeleri olduğu bir ağaç olarak temsil edilebilir (Štěpánek and Šimková, 2013). HTML için özgün ortam dışına gönderim formatı denebilir (Raggett, 1994). Bu formata göre metinler, resimler, tablolar, formalar vb. için geçerli olan biçimlendirme gereksinimleri "tag" adı verilen tanımlayıcı etiketlerle temsil edilebilir. Dahası bu etiketler yardımıyla özgün ortam dışına aktarıldığında biçimi bozulmadan tekrar bir araya getirilerek görüntülenebilir. Temelinde ise etiketler arasında kurulmuş hiyerarşik bir yapı vardır. Her etiket hiyerarşik yapıya uymak zorundadır.

Verilerin bir aygıttan diğerine aktarıldığında yeni konumunda nasıl görüntüleneceği ve depolanacağı çeşitlilik gösterse de anlaşılırlık ve tutarlılık temel gereksinimdir. Veriler TCP-IP (Transmission Control Protocol and Internet Protocol) protokollerine göre paketlere ayrılır, paketler telefon hattı üzerinden gönderilir ve alıcı bilgisayar tarafından kendi internet yazılımı kullanılarak etiketlere göre tekrar bir araya getirilir (Berners-Lee and Fischetti, 2000).

Bilgisayarlar için çok sayıda veri kaynağını HTML yapısının esaslarına göre gömülü veya doğrudan yazılarak barındıran sunucular (server) bilginin düğüm noktaları olarak düşünülebilir. Yoğun bir şekilde veri veya hizmet barındıran bu düğümler içerisinden istenilen bilgiyi tam olarak alabilmek için sıkı bir ayıklama yapmak kaçınılmazdır. Bu süreç sunucu için bant genişliği, işlemci ve bellek gibi kaynakların tüketiminde, istemci (client) içinse kaynak ve zaman maliyetlerini doğrudan etkiler. Bu nedenle istenilen bilginin az kaynak ile en doğru ve kısa sürede edilmesi amaçlanmaktadır. Bunun yanında web sayfalarından "yararlı ve ilgili" içeriğin çıkarılması, cep telefonu ve PDA taraması, görme engelliler için konuşma oluşturma ve metin

özetleme gibi birçok uygulamaya sahiptir (Gupta, Kaiser, Neistadt and Grimm, 2003). Web sayfalarının bölümlere ayrılması ve gürültünün (bilgilendirici olmayan bölüm) kaldırılması, duyarlılık analizi, metin özetleme ve bilgi erişimi gibi çeşitli uygulamalarda önemli ön işleme adımlarındandır (Pappas, Katsimpras and Stamatatos, 2012).

Sonuçta günümüzde web belgeleri üzerine yerleşmiş olan ve çok büyük miktarda kozmikleşmiş veri barındıran web ekosistemi ortaya çıkmıştır. Bu kozmik veriden anlamlı, işe yarar (knowledge) sonuçların çıkarılması için veri madenciliği (data mining) ve yapay zeka (artificial intelligence) tekniklerinin kullanılması kaçınılmazdır.

Veri madenciliği, büyük ve karmaşık veriler içerisinden anlamlı ve işe yarar bilgiyi ortaya çıkarmanın yöntemlerini inceleyen bir disiplindir. 1990'lardan bu yana, veri madenciliği kavramı, akademik alandan iş dünyasına veya tıbbi faaliyetlere kadar pek çok ortamda ortaya çıkmıştır (Gorunescu, 2011). Anlamlı bilginin keşfinde bir diğer ilgili disiplin ise yapay zekadır. Yapay zeka (Artificial Intelligence-AI), insanın öğrenme sürecine benzer olarak bilgisayarların veriden öğrenmesinin yöntemlerini inceleyen bilim dalıdır. Yapay zeka disiplininin altında yer alan makine öğrenmesi alt alanında denetimli (Supervised), denetimsiz (Unsupervised) ve Yarı-denetimli (semi-supervised) öğrenme olmak üzere üç tür öğrenme bulunmaktadır. Veri madenciliği ve yapay zeka veri içerisinden daha önce ortaya çıkarılmamış öngörülemeyen bilgiler gibi işe yarar bilgiyi ortaya çıkarmak için kullanırlar. Yapay zeka sistemleri, boyut ve karmaşıklık açısından giderek daha yetenekli olma eğilimindedir (Shuldiner, 2019). Sınıflandırma, kümeleme (clustering), birliktelik ilişkileri kurma gibi veriler üzerinde çok farklı görevleri yerine getirebilirler. Özellikle gelecekteki veriler üzerinde öngörü sunabilmesi ve faydalı örüntüler keşfedebilmesi onları cazip kılar.

Web içeriği çıkarma (web content extraction) teknikleri iki kategoride gruplanabilir: el yapımı kurallar ve otomatik ayıklama (Uzun, Serdar Güner, Kılıçaslan, Yerlikaya and Agun, 2014). Ele alınan verinin karmaşık, çok boyutlu ve büyük olması gibi anatomik özelliklerinden dolayı akıllı ve otomatik bir sistem geliştirmek oldukça faydalı ve başarılı olabilir.

Bu çalışmanın birinci amacı yapay zeka teknikleri ile layout-bölüm sınıflandırma işleminin başarımını artırmaktır. İkinci amacı ise algoritmaların başarım süreleri yanında işlem hızlarını da elde ederek analiz etmektir. Çalışmanın ikinci bölümünde önceki çalışmalar özetlenmiştir. Üçüncü bölümünde ise makine öğrenmesi teknikleri ve değerlendirme metrikleri hakkında bilgi verilmiştir. Dördüncü bölümde ise bulgular sunulmuş, son bölümde sonuçlar verilerek tartışılmıştır.

## 2. Önceki Çalışmalar

Web sayfalarındaki gerçek içeriği ayıklamak için çok sayıda akademik çalışma yapılmıştır. Wu ve ark. (Wu et al., 2015) yaptıkları araştırmada web sayfalarının DOM (Document Object Model-Belge Nesne Modeli) ağaç düğümü özelliklerini kullanarak birden çok özellik elde etmişler ve bu özellikleri makine öğrenimi yöntemini ile modellemeye çalışmışlardır. Araştırmacılar gerçek içeriğin uzamsal ve sürekli bir blokta yer aldığını gözlemlemişlerdir. Gupta ve ark. (Gupta et al., 2003) ise yine DOM ağacı ile orijinal verileri özetlemek yerine tanımlayarak ve koruyarak İçerik çıkarmaya çalışmışlardır. Weninger ve ark. farklı bir teknikle HTML belgesinin etiket oranlarını kullanarak çeşitli web sayfalarından içerik metni çıkarmak için Etiket Oranları (Content Extraction via Tag Ratios-CETR) adlı bir yöntem önermişlerdir. Uzun ve ark. (Uzun et al., 2014) ise yaptıkları araştırmada yedi farklı blok üzerinden web içeriğini otomatik olarak elde eden iCrawler adlı bir akıllı tarayıcı geliştirmişlerdir. Araştırmacılar daha sonra topladıkları içeriği makine öğrenmesi algoritmalarından DecisionTable (Karar tablosu) algoritması ile yüksek doğruluk oranı ile modellemeyi başarmışlardır. Yang ve Song(Yang and Song, 2010) ise heterojen yapıdaki web sayfaları ile başa çıkmada daha fazla uyarlanabilirliğe sahip gürültü ve karakteristiği gidermek üzere kullanılmayan aday düğümleri düzeltmeye dayalı bir yöntem önermişlerdir. Pappas ve ark. (Pappas et al., 2012), web sayfasının

görsel ve görsel olmayan özelliklerini hesaba katan ve kullanıcı tarafından oluşturulan içeriği (Haberler, Bloglar, Tartışmalar) içeren üç ana sayfa kategorisinden gürültülü bölümleri kaldırabilen bir algoritma önermişlerdir. Diğer yandan Bu ve ark. (Bu, Zhang, Xia and Wang, 2014) ana metin içeriğini web sayfalarından çıkarmak için bulanık ilişki kuralları (fuzzy association rules-FAR) kavramını kayan pencere (sliding window-SW) kavramıyla bütünleştiren istatistik tabanlı bir yaklaşım önermişlerdir. Uzun ve ark. (Uzun, Agun and Yerlikaya, 2013) gürültülü içeriği ortadan kaldırmak ve istenilen bilgiye ulaşmak için hibrit bir yöntem önermişlerdir. Lin ve ark. (Lin, Sheng, Vo and Tata, 2020)FreeDOM adını verdikleri araştırmada her site için örnek gerektiren ve web sitelerinin görsel yapısı üzerine inşa edilen sezgisel içerik çıkarım yöntemlerin sınırlılıkları olduğunu belirtmişlerdir. Araştırmacılar bu sorunu çözmek için FreeDOM'un web sayfalarının metin ve biçim bilgilerini birleştirerek sayfadaki her DOM düğümünün temsilini (Word embedding) öğrendiğini ve bu bilgiyi bir sinir ağı ile semantik ilişkiler elde etmek için kullandığını göstermişlerdir. Uçar ve ark. (Uçar, Uzun and Tüfekci, 2016) birbirini tamamlayan iki aşamalı bir algoritma önererek yüksek doğruluk elde etmişlerdir.

Bu çalışmada bu veri seti için daha önce kullanılmamış olan algoritmalar ve algoritma optimizasyon araçları kullanılmış ve başarım oranı aynı veriyi kullanan önceki çalışmalara göre artırılmıştır. Bunun yanında model hesaplama zamanları da çıkarılarak karşılaştırılıştır.

## 3. Materyal ve Metot

Çalışmada veri setinin analizi için farklı makine öğrenmesi algoritmaları eğitilmiştir. Daha sonra, testlerden elde edilen bulgular kaydedilerek değerlendirme metriklerine göre değerlendirilmiştir.

### *3.1 Makine Öğrenmesi Algoritma Testleri*

Çalışmanın makine öğrenmesi algoritmaları testi Weka (Frank et al., 2009) (Waikato Environment for Knowledge Analysis) adlı Waikato üniversitesi tarafından geliştirilmiş açık kaynak kodlu, Java tabanlı yazılım aracılığı ile yapılmıştır. Weka bünyesinde çok sayıda yapay zeka algoritmasını barındırmaktadır. Package Manager sayesinde yeni algoritma ve veri işleme araçları yüklemek mümkündür. Veriler üzerinde çok sayıda algoritma denenmiş ve en başarılı ilk 5 algoritma (Random Forest, Random Tree, JRIP, Bagging, KStar) tespit edilmiştir.

İlk algoritma Random Forest'tır. Random Forest algoritması, hem sınıflandırma hem de regresyon görevlerini gerçekleştirebilen çok yönlü ve akıllı bir makine öğrenmesi yöntemi olarak tanımlanabilir (Sullivan, 2018). Random Forest algoritması, ağaç indüksiyon algoritmasının rastgele bir varyantından türetilen ve bir karar ağaçları topluluğu (veya ormanı) oluşturmayı içerir (Louppe, 2014). Bu karar ağaçları kullanılarak özellikle büyük veriler üzerinde etkili çözümler elde edilebilir.

Ağaçlar, döngüsüz bağlı grafikler olarak tanımlanır ve özellikleri grafik (graph) teorisinin temelleridir (Drmota, 2009). Düğümler ve düğümlere komşu olan diğer düğümlerden meydana gelirler. Bu düğümlerden bir tanesi (root-r) kök düğüm olarak adlandırılır. Ancak köksüz de olabilirler. Ağaç tabanlı algoritmalarda ağacı ters çevrilmiş olarak düşünürsek bir verinin sınıfı kök düğümden başlanarak aşağı doğru her bir düğümdeki kritere göre yönlendirilerek bulunur. Random Tree sınıflandırıcı birkaç karar ağacını, önyüklemeye paralel olarak eğiten ve ardından bagging (torbalama) adı verilen işlemle bir araya getiren yöntemler topluluğudur (Misra, Li and He, 2019).

JRIP algoritması Cohen(Cohen, 1995) tarafından geliştirilmiş eğitim örnekleriyle hızlı ölçeklenebilen ve yüzbinlerce örnek içeren gürültülü veri kümelerini verimli bir şekilde işleyebilen kural tabanlı bir algoritmadır. Bu tür algoritmalar basit deterministik mantıksal kurallar üretir ve tüm örneklerin mükemmel bir doğrulukla sınıflandırılmasına izin verir (Nosofsky, Gluck, Palmeri, Mckinley and Glauthier, 1994). Kural tabanlı bir sınıflandırıcının kuralları bir karar ağacından çıkarılabilir (Yucalar, Ozcift, Borandag and Kilinc, 2020).

Bagging ensemble (birlikte) öğrenme temeline dayalı bir algoritmadır. Birlikte öğrenme yöntemleri, tek bir karar ağacı sınıflandırıcısından daha iyi tahmin performansı üretmek için birkaç karar ağacı sınıflandırıcısını birleştirir (Sarkar, 2019). Birliktelik modelinin arkasındaki ana ilke, bir grup zayıf öğrenicinin güçlü bir öğrenici oluşturmak için bir araya gelmesi ve böylece modelin doğruluğunun arttırılmasıdır. Bagging algoritması, önce bir sınıflandırıcılar komisyonu kurar ve ardından bunların sonuçlarını çoğunluk oylamasıyla toplar (Hsu and Srivastava, 2012).

KStar algoritması, olası tüm dönüşümler arasından rasgele seçim yaparak entropik ölçüm kullanır (Madhusudana, Kumar and Narendranath, 2016). Uzaklık ölçümü olarak entropi kullanımı sembolik niteliklerin, gerçek değerli niteliklerin ve eksik değerlerin ele alınmasında tutarlı bir yaklaşım sağlar (Cleary and Trigg, 1995).

### *3.2 Değerlendirme Metrikleri*

Makine öğrenmesi algoritmalarının performanslarını değerlendirmek için standart bazı metrikler kullanılır. Bunların sayısı çok olmakla birlikte kullanılacak olan alan ve beklentiye göre değerlendirme metrikleri özel olarak seçilebilir. Örneğin pozitif örneklerin önemi büyük ise ve yanlış pozitifliğin maliyeti düşük ise eşik değeri düşük tutularak tüm pozitifiler yakalanabilir. Çalışmada kullanılan algoritmalar için değerlendirme metriği olarak F-Measure, Kappa, RMSE ve Correctly Classified Instances seçilmiştir.

F-Measure, Precision (kesinlik) ve Recall değerlerinin harmonik ortalamasıdır. Recall model tarafından öğrenilen veya deneyimlenen bir şeyi hatırlama eylemi veya yeteneğini ifade eder. Precision ise modelin tahminlerindeki kesinliği ifade eder. Modelin çıktıları 4 durumdan biri olabilir bunlar:

- TP: Gerçekte pozitif iken model tarafından da pozitif olarak sınıflandırılanlar
- FP: Gerçekte negatif iken model tarafından pozitif olarak sınıflandırılanlar.
- TN: Gerçekte negatif iken model tarafından da negatif olarak sınıflandırılanlar.
- FN: Gerçekte pozitif iken model tarafından negatif olarak sınıflandırılanlardır.

Precision denklem 1'de gösterildiği gibi modelin sınıflandırdığı pozitif örnek sayısının toplam pozitif girdi sayısına oranıdır.

$$P = \frac{TP}{TP + FP} \qquad R = \frac{TP}{TP + FN} \tag{1}$$

Recall ise yine denklem 1'de gösterildiği gibi modelin pozitif sınıflandırdığı örnekler içerisinden kaç tanesinin gerçekten pozitif olduğunun ölçüsüdür. F-Measure veya literatürde F-Ölçüsü, F-Score olarak adlandırılan metrik ise denklem 2'deki gibi hesaplanır.

$$F - Measure = \frac{2 \times P \times R}{P + R} \tag{2}$$

Correctly Classified Intances değeri ise modelin doğru sınıflandırdığı örnek sayısının yüzdelik ifadesi olarak temsil edilen metriktir. Kappa ise gözlenen ve beklenen değerler arasındaki uyuşmayı gösterir. Yani modelin çıktısı ile beklenen çıktı (gerçek) arasındaki uyuşmayı temsil eder. Kappa -1 ile +1 arasında değer alabilir. Model için 1'e yakın bir kappa değeri istenir.

RMSE (Root mean squared error) ise denklem 3'te gösterildiği gibi modelin tahminleri ile gerçek değer arasındaki hataların karasel ortalaması alındıktan sonra toplanması ve sonucun karekökün alınması ile elde edilir. Bu değerlendirme hassasiyeti sağlar.

$$RMSE = \sqrt{\sum_{i=1}^{n} \frac{(\text{Gerçek değer-Model Tahmini})^2}{n}} \tag{3}$$

İki algoritmanın performansı birbiriyle neredeyse aynı ise, o zaman RMSE'ye bakarak hangisinin daha iyi olduğu ayırt edebilir (Pradham, Younan and King, 2008). RMSE'nin rakip algoritmaya göre düşük değerli olması model sonuçlarının daha doğru olduğunu gösterir(Aydin, Yucel and Sadikoglu, 2018).

### *3.3 Veri Seti*

Araştırmada Uzun ve ark. (Uzun et al., 2014) tarafından elde edilmiş olan 49 girdi özniteliği ve 7 çıktısı (sınıfı) olan veri seti kullanılmıştır. Veri setinin çıktıları web sayfalarında bulunan ve "main, menü,

links, summary, empty, headline, others" olarak etiketlenmiş bölüm adlarıdır. Araştırmacılar veri setini 2011 yılına kadar Goggle News'den rastgele seçilen 110 farklı Web alanından olmak üzere toplam 2414 web sayfasından elde etmişlerdir. Veri seti (Uzun, 2014) 14742 satırdan oluşmaktadır. Veri setinin sınıf dağılımları ise Tablo 1'de gösterilmiştir.

Tablo 1'e bakıldığında en yüksek değerin menü, links ve empty'e ait olduğu görülür. En düşük değer ise summary sınıfına aittir.

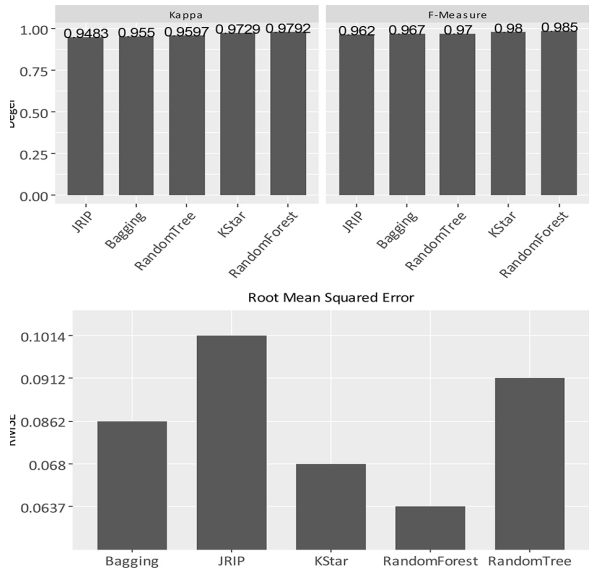
**Tablo 1.** Veri setinin sınıf dağılımı

| Sınıf Adı | Sayısı |
|---|---|
| Main | 549 |
| Headline | 553 |
| Summary | 73 |
| Others | 1889 |
| Menu | 5643 |
| Links | 4054 |
| Empty | 1981 |

Veri setinin sınıf dağılımı incelendiğinde dengeli bir dağılım olmamasına karşın sonraki bölümde paylaşılacak olan sonuçlara göre analiz gerekliliklerini sağladığı yani yanlılık ile karşılaşılmadığı görülmektedir.

## 4. Bulgular

Yapılan makine öğrenmesi algoritma testlerine göre elde edilen performans metrikleri ve değerleri Şekil 1'de sunulmuştur.

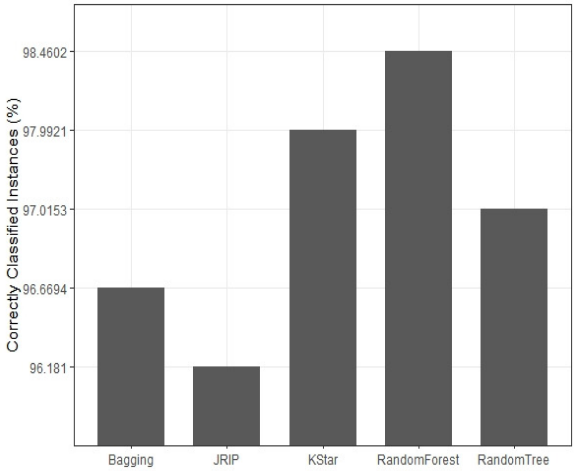



**Şekil 1.** Algoritmaların Kappa ve F-Measure değerleri

Şekil 1'in üst bölümü Kappa ve F-measure değerlerini göstermektedir. Random Forest ve KStar algoritmalarının Kappa ve F-measure değerlerine göre en iyi sınıflandırıcı algoritmalar olduğu görülmektedir.

Bunun yanında RMSE değerlerine bakıldığında en az hata oranına sahip olan algoritmaların sırasıyla Random Forest ve KStar olduğu görülmektedir.

Uzun ve ark. (Uzun et al., 2014)'nın Decision Table algoritmasıyla elde ettikleri doğru sınıflandırma (Accuracy) değeri %96.87'dir. Bu çalışmada elde edilen bulgular Şekil 2'de gösterilmiştir. Şekil 2'de gösterildiği gibi Random Forest algoritması ile elde edilen doğru sınıflandırma oranı %98.46'dir. Bunun yanında KStar (%97.99), Random Tree (%97.01) algoritmalarının da önceki çalışmaya göre daha iyi sonuçlar ürettiği görülmüştür.

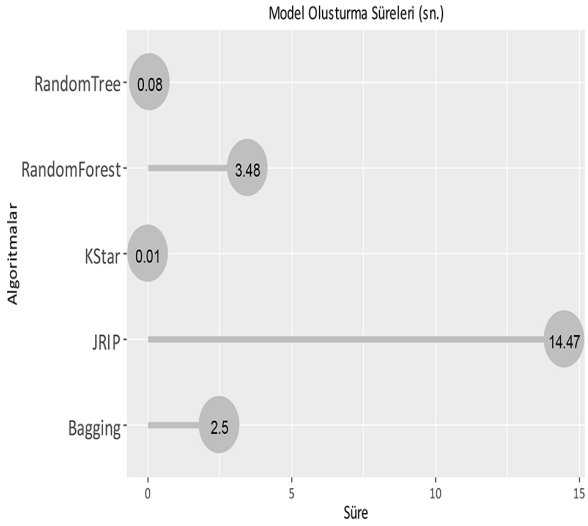



**Şekil 2.** Algoritmaların Correctly Classified Instances Değerleri

Başarımı diğerlerine göre daha düşük olan algoritmalar ise JRIP ve Bagging algoritmalarıdır.

Özellikle entropi temelli uzaklık ölçümü kullanan KStar ve kural tabanlı ağaç algoritmalarının veri setinin sınıflandırılmasında yüksek başarı gösterdikleri söylenebilir. Veriler tür olarak sayısal verilerdir. Sadece çıktılar nominal türdedir.

Şekil 2'deki Correctly Classified Instances ve Şekil 3'teki F-Measure, Kappa ve RMSE değerleri düşünüldüğünde hız/performans açısından KStar algoritmasının Random Forest'ın alternatifi olarak kullanılabileceği söylenebilir.




**Şekil 3.** Algoritmaların model oluşturma süreleri

Şekil 3'te görüldüğü gibi KStar algoritması en hızlı model kurma zamanına sahip algoritmadır.

## 5. Sonuçlar ve Tartışma

Web sayfalarından içeriğin ayıklanması ve çıkarılması özellikle erişim ve işlem hızı, depolama, bant genişliği ve istenilen bilgiye en doğru şekilde ulaşmak açısından son derece önemli bir görevdir. Web sayfalarının layout'larının tespit edilmesi bilgi çıkarımı için iyi bir başlangıç olabilir. Bu çalışmada çeşitli web sayfalarından toplanarak oluşturulan veri seti üzerinde bölüm sınıflandırması işlemi makine öğrenmesi algoritmalarıyla iyileştirilmeye çalışılmıştır. Sınıflandırma başarım oranı Random Forest algoritması ile %1.59 oranında iyileştirilmiştir.

Çalışmada ayrıca sınıflandırıcı makine öğrenmesi algoritmalarının model oluşturma süreleri de analiz edilmiş bu analizler sonucunda hem başarım hem de hesaplama süresi bakımından alternatif olarak KStar algoritmasının kullanılabileceği görülmüştür.

Çalışmada elde edilen sonuçlar makine öğrenmesi algoritmalarının yapısal açıdan değerlendirildiğinde ağaç ve entropi tabanlı algoritmaların bu veri seti üzerinde daha başarılı sonuçlar verdiğini göstermiştir.

Gelecekte araştırmacılar bu veri seti üzerinde performans artırıcı çalışmalar yürütebilirler. Bunun yanında veri setinin ve barındırdığı özniteliklerin geliştirilmesi ve iyileştirilmesi de düşünülebilir.

**Teşekkür**

Bu araştırmada kullanılan verileri sağlayan ve açık erişim şekilde yayınlayarak paylaşan Tekirdağ Namık Kemal Üniversitesi, Çorlu Mühendislik Fakültesi Bilgisayar Mühendisliği Bölümü öğretim üyesi Doç. Dr. Erdinç Uzun'a teşekkürlerimi sunarım.

## 6. Kaynaklar

Aydin, E. S., Yucel, O. and Sadikoglu, H. (2018). Chapter 2.6 - Numerical Investigation of Fixed-Bed Downdraft Woody Biomass Gasification. I. Dincer, C. O. Colpan and O. B. T.-E. Kizilkan Energetic and Environmental Dimensions (Eds.), (pp. 323–339). Academic Press. doi:https://doi.org/10.1016/B978-0-12-813734-5.00018-4

Berners-Lee, T. and Fischetti, M. (2000). *Weaving the Web: The Original Design and Ultimate Destiny of the World Wide Web* (1st ed.). New York, NY, USA: Harper Business.

Bu, Z., Zhang, C., Xia, Z. and Wang, J. (2014). An FAR-SW based approach for webpage information extraction. *Information Systems Frontiers*, *16*(5), 771–785. doi:10.1007/s10796-013-9412-2

Cleary, J. G. and Trigg, L. E. (1995). K*: An Instance-based Learner Using an Entropic Distance Measure. *Machine Learninginternational Workshop Then Conference*, *5*, 1–14. doi:10.1.1.51.4098

Cohen, W. W. (1995). Fast Effective Rule Induction. A. Prieditis and S. B. T.-M. L. P. 1995 Russell (Eds.), (pp. 115–123). San Francisco (CA): Morgan Kaufmann. doi:https://doi.org/10.1016/B978-1-55860-377-6.50023-2

Drmota, M. (2009). *Random trees: An interplay between combinatorics and probability*. *Random Trees: An Interplay Between Combinatorics and Probability*. doi:10.1007/978-3-211-75357-6

Duckett, J. (2011). *HTML and CSS: Design and Build Websites*. (C. Long, Ed.). Indianapolis, Indiana: John Wiley & Sons.

Frank, E., Hall, M., Holmes, G., Kirkby, R., Pfahringer, B., Witten, I. H. and Trigg, L. (2009). Weka-A Machine Learning Workbench for Data Mining. *Data Mining and Knowledge Discovery Handbook* in . doi:10.1007/978-0-387-09823-4_66

Gorunescu, F. (2011). *Data Mining*. Intelligent Systems Reference Library (Vol. 12). Berlin, Heidelberg: Springer Berlin Heidelberg. doi:10.1007/978-3-642-19721-5

Gupta, S., Kaiser, G., Neistadt, D. and Grimm, P. (2003). DOM-Based Content Extraction of HTML Documents. *Proceedings of the 12th International Conference on World Wide Web* in , WWW '03 (pp. 207–214). New York, NY, USA: Association for Computing Machinery.

doi:10.1145/775152.775182

Hsu, K.-W. and Srivastava, J. (2012). Improving Bagging Performance through Multi-algorithm Ensembles. L. J. Huang J.Z., Bailey J., Koh Y.S. (Ed.), *New Frontiers in Applied Data Mining* in (pp. 471–482). Berlin, Heidelberg: Springer Berlin Heidelberg. doi:10.1007/978-3-642-28320-8_40

Lin, Y., Sheng, Y., Vo, N. H. and Tata, S. (2020). FreeDOM: A Novel Transferable Neural Architecture for Structured Data Extraction over Web Documents. *KDD 2020* in .

Louppe, G. (2014, July). *Understanding Random Forests from Theory to Practice*. https://arxiv.org/pdf/1407.7502.pdf from retrieved.

Madhusudana, C. K., Kumar, H. and Narendranath, S. (2016). Condition monitoring of face milling tool using K-star algorithm and histogram features of vibration signal. *Engineering Science and Technology, an International Journal*, *19*(3), 1543–1551. doi:https://doi.org/10.1016/j.jestch.2016.05.009

Misra, S., Li, H. and He, J. (2019). *Machine Learning for Subsurface Characterization*. Elsevier Science. https://books.google.com.tr/books?id=WdO1DwAAQBAJ from retrieved.

Nosofsky, R. M., Gluck, M. A., Palmeri, T. J., Mckinley, S. C. and Glauthier, P. (1994). Comparing modes of rule-based classification learning: A replication and extension of Shepard, Hovland, and Jenkins (1961). *Memory & Cognition*, *22*(3), 352–369. doi:10.3758/BF03200862

Pappas, N., Katsimpras, G. and Stamatatos, E. (2012). Extracting Informative Textual Parts from Web Pages Containing User-Generated Content. *Proceedings of the 12th International Conference on Knowledge Management and Knowledge Technologies* in , i-KNOW '12. New York, NY, USA: Association for Computing Machinery. doi:10.1145/2362456.2362462

Pradham, P., Younan, N. H. and King, R. L. (2008). 16 - Concepts of image fusion in remote sensing applications. T. B. T.-I. F. Stathaki (Ed.), (pp. 393–428). Oxford: Academic Press. doi:https://doi.org/10.1016/B978-0-12-372529-5.00019-6

Raggett, D. (1994). A review of the HTML + document format. *Computer Networks and ISDN Systems*, *27*(2), 135–145. doi:https://doi.org/10.1016/0169-7552(94)90127-9

Sarkar, P. (2019, 14 October). Bagging and Random Forest in Machine Learning: How do they work? 18 August 2020 tarihinde https://www.knowledgehut.com/blog/data-science/bagging-and-random-forest-in-machine-learning from retrieved.

Shuldiner, A. (2019). Chapter 8 - Raising Them Right: AI and the Internet of Big Things. W. Lawless, R. Mittu, D. Sofge, I. S. Moskowitz and S. B. T.-A. I. for the I. of E. Russell (Eds.), *Artificial Intelligence for the Internet of Everything* in (pp. 139–143). Academic Press. doi:https://doi.org/10.1016/B978-0-12-817636-8.00008-9

Štěpánek, J. and Šimková, M. (2013). Comparing Web Pages in Terms of Inner Structure. *Procedia - Social and Behavioral Sciences*, *83*, 458–462. doi:10.1016/j.sbspro.2013.06.090

Sullivan, W. (2018). *Decision Tree and Random Forest - Machine Learning and Algorithms: The Future Is Here!* CreateSpace Independent Publishing Platform. https://books.google.com.tr/books?id=x-u_tAEACAAJ from retrieved.

Uçar, E., Uzun, E. and Tüfekci, P. (2016). A novel algorithm for extracting the user reviews from web pages. *Journal of Information Science*, *43*(5), 696–712. doi:10.1177/0165551516666446

Uzun, E. (2014). iCrawler/Dataset at master · erdincuzun/iCrawler. 18 August 2020 tarihinde https://github.com/erdincuzun/iCrawler/tree/master/Dataset from retrieved.

Uzun, E., Agun, H. V. and Yerlikaya, T. (2013). A hybrid approach for extracting informative content from web pages. *Information Processing & Management*, *49*(4), 928–944. doi:https://doi.org/10.1016/j.ipm.2013.02.005

Uzun, E., Serdar Güner, E., Kılıçaslan, Y., Yerlikaya, T. and Agun, H. V. (2014). An effective and efficient Web content extractor for optimizing the crawling process. *Software: Practice and Experience*, *44*(10), 1181–1199. doi:10.1002/spe.2195

Wu, S., Liu, J. and Fan, J. (2015). Automatic Web Content Extraction by Combination of Learning and Grouping. *Proceedings of the 24th International Conference on World Wide Web* in , WWW '15 (pp. 1264–1274). Republic and Canton of Geneva, CHE: International World Wide Web Conferences Steering Committee. doi:10.1145/2736277.2741659

Yang, D. and Song, J. (2010). Web Content Information Extraction Approach Based on Removing Noise and Content-Features. *2010 International Conference on Web Information Systems and Mining* in (Vol. 1, pp. 246–249). doi:10.1109/WISM.2010.82

Yucalar, F., Ozcift, A., Borandag, E. and Kilinc, D. (2020). Multiple-classifiers in software quality engineering: Combining predictors to improve software fault prediction ability. *Engineering Science and Technology, an International Journal*, *23*(4), 938–950. doi:https://doi.org/10.1016/j.jestch.2019.10.005

# AFYON KOCATEPE ÜNİVERSİTESİ ULUSLARARASI MÜHENDİSLİK TEKNOLOJİLERİ ve UYGULAMALI BİLİMLER DERGİSİ

Afyon Kocatepe Üniversitesi
Ahmet Necdet Sezer Kampüsü
Teknoloji Fakültesi
AFYONKARAHİSAR
Tel: +90 272 228 14 46
Belgegeçer: +90 272 228 14 49
E-posta: ijetas@aku.edu.tr

www.ijetas.aku.edu.tr