p-ISSN 2521-4209
e-ISSN 2521-4217

جامعة التنمية البشرية
UNIVERSITY OF HUMAN DEVELOPMENT

UHD Journal of Science and Technology

A Scientific periodical issued by University of Human Developement

Vol.5 No.(2) December 2021

2021 2721

www.jst.uhd.edu.iq

Kurdistan Regional Government
University of Human Development

UHD Journal of Science and Technology

A periodic scientific journal issued by University of Human Development

Introduction

UHD Journal of Science and Technology (UHDJST) is a semi-annual journal published by the University of Human Development, Sulaymaniyah, Kurdistan Region, Iraq. UHDJST member of ROAD, e-ISSN: 2521-4217, p-ISSN: 2521-4209 and a member of Crossref, DOI: 10.21928/issn.2521-4217. UHDJST publishes original research in all areas of Science, Engineering, and Technology. UHDJST is a Peer-Reviewed Open Access journal with Creative Commons Attribution Non-Commercial No Derivatives License 4.0 (CC BY-NC-ND 4.0). UHDJST provides immediate, worldwide, barrier-free access to the full text of research articles without requiring a subscription to the journal, and has article processing charge (APC). UHDJST applies the highest standards to everything it does and adopts APA citation/referencing style. UHDJST Section Policy includes three types of publications: Articles, Review Articles, and Letters.

By publishing with us, your research will get the coverage and attention it deserves. Open access and continuous online publication mean your work will be published swiftly, ready to be accessed by anyone, anywhere, at any time. Article Level Metrics allow you to follow the conversations your work has started.

UHDJST publishes works from extensive fields including, but not limited to:

- Pure Science
- Applied Science
- Medicine
- Engineering
- Technology

Scope and Focus

UHD Journal of Science and Technology (UHDJST) publishes original research in all areas of Science and Engineering. UHDJST is a semi-annual journal published by the University of Human Development, Sulaymaniyah, Kurdistan Region, Iraq. We believe that if your research is scientifically valid and technically sound then it deserves to be published and made accessible to the research community. UHDJST aims to provide a service to the international scientific community enhancing swap space to share, promote and disseminate the academic scientific production from research applied to Science, Engineering, and Technology.

SEARCHING FOR PLAGIARISM

We use plagiarism detection: detection; According to Oxford online dictionary, Plagiarism means: *The practice of taking someone else's work or ideas and passing them off as one's own.*

Section Policies

No.	Title	Peer Reviewed	Indexed	Open Submission
1	Articles: This is the main type of publication that UHDJST will produce	☑	☑	☑
2	Review Articles: Critical, constructive analysis of the literature in a specific field through summary, classification, analysis, comparison.	☑	☑	☑
3	Letters: Short reports of original research focused on an outstanding finding whose importance means that it will be of interest to scientists in other fields.	☑	☑	☑

PEER REVIEW POLICIES

At UHDJST we are committed to prompt quality scientific work with local and global impacts. To maintain a high-quality publication, all submissions undergo a rigorous review process. Characteristics of the peer review process are as follows:

- The journal peer review process is a "double-blind peer review".
- Simultaneous submissions of the same manuscript to different journals will not be tolerated.
- Manuscripts with contents outside the scope will not be considered for review.
- Papers will be refereed by at least 2 experts as suggested by the editorial board.
- In addition, Editors will have the option of seeking additional reviews when needed. Authors will be informed when Editors decide further review is required.
- All publication decisions are made by the journal's Editors-in-Chief on the basis of the referees' reports. Authors of papers that are not accepted are notified promptly.
- All submitted manuscripts are treated as confidential documents. We expect our Board of Reviewing Editors, Associate Editors and reviewers to treat manuscripts as confidential material as well.
- Editors, Associate Editors, and reviewers involved in the review process should disclose conflicts of interest resulting from direct competitive, collaborative, or other relationships with any of the authors, and remove oneself from cases in which such conflicts preclude an objective evaluation. Privileged information or ideas that are obtained through peer review must not be used for competitive gain.
- Our peer review process is confidential and the identities of reviewers cannot be revealed.

Note: UHDJST is a member of CrossRef and CrossRef services, e.g., CrossCheck. All manuscripts submitted will be checked for plagiarism (copying text or results from other sources) and self-plagiarism (duplicating substantial parts of authors' own published work without giving the appropriate references) using the CrossCheck database. Plagiarism is not tolerated.

For more information about CrossCheck/iThenticate, please visit
http://www.crossref.org/crosscheck.html.

OPEN ACCESS POLICY

This journal provides immediate open access to its content on the principle that making research freely available to the public supports a greater global exchange of knowledge. Open Access (OA) stands for unrestricted access and unrestricted reuse which means making research publications freely available online. It access ensures that your work reaches the widest possible audience and that your fellow researchers can use and share it easily. The mission of the UHDJST is to improve the culture of scientific publications by supporting bright minds in science and public engagement.

UHDJST's open access articles are published under a Creative Commons Attribution CC-BY-NC-ND 4.0 license. This license lets you retain copyright and others may not use the material for commercial purposes. Commercial use is one primarily intended for commercial advantage or monetary compensation. If others remix, transform or build upon the material, they may not distribute the modified material. The main output of research, in general, is new ideas and knowledge, which the UHDJST peer-review policy allows publishing as high-quality, peer-reviewed research articles. The UHDJST believes that maximizing the distribution of these publications - by providing free, online access - is the most effective way of ensuring that the research we fund can be accessed, read and built upon. In turn, this will foster a richer research culture and cultivate good research ethics as well. The UHDJST, therefore, supports unrestricted access to the published materials on its main website as a fundamental part of its mission and a global academic community benefit to be encouraged wherever possible.

Specifically:

- The University of Human Development supports the principles and objectives of Open Access and Open Science
- UHDJST expects authors of research papers, and manuscripts to maximize the opportunities to make their results available for free access on its final peer-reviewed paper
- All manuscript will be made open access online soon after final stage peer-review finalized.
- This policy will be effective from 17th May 2017 and will be reviewed during the first year of operation.
- Open Access route is available at http://journals.uhd.edu.iq/index.php/uhdjst for publishing and archiving all accepted papers,
- Specific details of how authors of research articles are required to comply with this policy can be found in the Guide to Authors.

ARCHIVING

This journal utilizes the LOCKSS and CLOCKSS systems to create a distributed archiving system among participating libraries and permits those libraries to create permanent archives of the journal for purposes of preservation and restoration.

LOCKSS: Open Journal Systems supports the LOCKSS (Lots of Copies Keep Stuff Safe) system to ensure a secure and permanent archive for the journal. LOCKSS is open source software developed at Stanford University Library that enables libraries to preserve selected web journals by regularly polling registered journal websites for newly published content and archiving it. Each archive is continually validated against other library caches, and if the content is found to be corrupted or lost, the other caches or the journal is used to restore it.

CLOCKSS: Open Journal Systems also supports the CLOCKSS (Controlled Lots of Copies Keep Stuff Safe) system to ensure a secure and permanent archive for the journal. CLOCKSS is based upon the open-source LOCKSS software developed at Stanford University Library that enables libraries to preserve selected web journals by regularly polling registered journal websites for newly published content and archiving it. Each archive is continually validated against other library caches, and if the content is found to be corrupted or lost, the other caches or the journal is used to restore it.

PUBLICATION ETHICS

Publication Ethics and Publication Malpractice Statement

The publication of an article in the peer-reviewed journal UHDJST is to support the standard and respected knowledge transfer network. Our publication ethics and publication malpractice statement is mainly based on the Code of Conduct and Best-Practice Guidelines for Journal Editors (Committee on Publication Ethics, 2011) that includes;

- General duties and responsibilities of editors.
- Relations with readers.
- Relations with the authors.
- Relations with editors.
- Relations with editorial board members.
- Relations with journal owners and publishers.
- Editorial and peer review processes.
- Protecting individual data.
- Encouraging ethical research (e.g. research involving humans or animals).
- Dealing with possible misconduct.
- Ensuring the integrity of the academic record.
- Intellectual property.
- Encouraging debate.
- Complaints.
- Conflicts of interest.

ANIMAL RESEARCHES

- For research conducted on regulated animals (which includes all live vertebrates and/or higher invertebrates), appropriate approval must have been obtained according to either international or local laws and regulations. Before conducting the research, approval must have been obtained from the relevant body (in most cases an Institutional Review Board, or Ethics Committee). The authors must provide an ethics statement as part of their Methods section detailing full information as to their approval (including the name of the granting organization, and the approval reference numbers). If an approval reference number is not provided, written approval must be provided as a confidential supplemental information file. Research on non-human primates is subject to specific guidelines from the Weather all (2006) report (The Use of Non-Human Primates in Research).
- For research conducted on non-regulated animals, a statement should be made as to why ethical approval was not required.
- Experimental animals should have been handled according to the highest standards dictated by the author's institution.
- We strongly encourage all authors to comply with the '*Animal Research: Reporting In Vivo Experiments*' (ARRIVE) guidelines, developed by NC3Rs.
- Articles should be specific in descriptions of the organism(s) used in the study. The description should indicate strain names when known.

ARTICLE PROCESSING CHARGES

UHDJST is an Open Access Journal (OAJ) and has article processing charges (APCs). The published articles can be downloaded freely without a barrier of admission.

Address

University of Human Development, Sulaymaniyah-Kurdistan Region/Iraq
PO Box: Sulaymaniyah 6/0778

Contact

Principal Contact

Dr. Aso Darwesh

Editor-in-Chief

University of Human Development – Sulaymaniyah, Iraq

Phone: +964 770 148 5879

Email: jst@uhd.edu.iq

Support Contact

UHD Technical Support

Phone: +964 770 247 3391

Email: jst@uhd.edu.iq

Contents

ORIGINAL RESEARCH ARTICLE UHD JOURNAL OF SCIENCE AND TECHNOLOGY

The Luminosity Function of Galaxies in Some Nearby Clusters

Mariwan A. Rasheed[1,2] and Khalid K. Mohammad[2]

[1]Development Center for Research and Training, University of Human Development, Sulaimani, Kurdistan Region, Iraq, [2]Department of Physics, College of Science, University of Sulaimani, Sulaimani, Kurdistan Region, Iraq

ABSTRACT

In the present work, the galaxy luminosity function (LF) has been studied for a sample of seven clusters in the redshift range ($0.0 \lesssim z \lesssim 0.1$), within Abell radius ($1.5\ h^{-1}$ Mpc), in the five SDSS passbands *ugriz*. In each case, the absolute magnitude distribution is found and then fitted with a Schechter function. The fitting is done, using the χ^2 – minimization method to find the best values of Schechter parameters Φ^* (normalization constant), M^* (characteristic absolute magnitude), and α (faint-end slope). No remarkable changes are found in the values of M^* and α, for any cluster, in any passband. Furthermore, the LF does not seem to vary with such cluster parameters as richness, velocity dispersion, and Bautz–Morgan morphology. Finally, it is found that M^* becomes brighter toward redder bands, whereas almost no variation is seen in the value of α with passband, being around (-1.00).

Index Terms: *Galaxies, Clusters, Luminosity function, Galaxy formation, Galaxy evolution*

1. INTRODUCTION

Galaxies come in a diversity of sizes and cover a very wide range of luminosities, extending from the faintest dwarfs to the most luminous giant ellipticals. To know how these galaxies are distributed with respect to their luminosities, the luminosity function (LF) is used. It is one of the most important techniques used for studying galaxy formation and evolution. A suitable approximation to this function was given by Paul Schechter in 1976 [1]. It can be written as

$$\Phi(L) = \left(\frac{\Phi^*}{L^*}\right)\left(\frac{L}{L^*}\right)^{\alpha} exp\left(-\frac{L}{L^*}\right) \quad (1)$$

Access this article online

DOI: 10.21928/uhdjst.v5n2y2021.pp1-10

E-ISSN: 2521-4217

P-ISSN: 2521-4209

where, L^* is a characteristic luminosity, indicating the change from power law ($L < L^*$) to exponential law ($L > L^*$), α is the faint-end slope, and Φ^* is a normalization constant for the distribution. These parameters may take different values for different morphological types and also for different environments. Considering an interval dL in luminosity, $\Phi(L)$ dL gives the number density of galaxies.

Galaxy clusters are ideal systems for studying the galaxy LF due to the existence of a large number of galaxies at almost the same distance. Many studies have thus been devoted to the LF of cluster galaxies to discover the influence of environment on their evolution. After the earlier works on the LF, carried out by Hubble (1936) [2], [3], Zwicky (1942) [4], Oemler (1974) [5], and others, Schechter (1976) [1] proposed the analytic expression given by Equation (1), which is called the Schechter function. He suggested that the cluster LF is universal in shape. This universality has been supported by various studies [6], [7], [8]. However, studies carried out by others [9], [10], [11] have demonstrated that the shape of the cluster LF is not universal.

Corresponding author's e-mail: Khalid K. Mohammad, Department of Physics, College of Science, University of Sulaimani, Sulaimani, Kurdistan Region, Iraq. E-mail: khalid.mohammad@univsul.edu.iq

Received: 21-04-2021 Accepted: 01-07-2021 Published: 03-07-2021

The LF of cluster galaxies has been compared to that of field galaxies through several studies. Some of these studies found them to be identical [12], [13], [14], while others found them to be different [8], [15], [16]. The cluster LF has been found to vary with cluster-centric radius [11], [17]. This is because different galaxy morphological types have different LFs [18] and that the mixture of these morphological types varies with cluster-centric radius, according to the morphology-density relation [19]. In fact, studying the variation of the cluster LF with such characteristics as cluster-centric radius, galaxy morphologies, and, also, galaxy colors is very important in constraining theories of galaxy formation and evolution.

In the present work, we study the LF of a sample of seven Abell-type galaxy clusters having redshifts in the range ($0.0 \lesssim z \lesssim 0.1$). A detailed description of the sample is given in Section 2, and the results and discussion are presented in Section 3. Our conclusions are outlined in Section 4. Throughout the work, ΛCDM parameters ($\Omega_M = 0.27$, $\Omega_\Lambda = 0.73$, $H_0 = 73$ km s^{-1} Mpc^{-1}) are used.

2. SAMPLE AND DATA

In this work, we consider a sample of seven nearby galaxy clusters, selected from Abell catalogue [20] within the redshift range ($0.0 \lesssim z \lesssim 0.1$). Their basic data are given in Table 1. All possible member galaxies within Abell radius (R_A= 1.5 h^{-1} Mpc) of each cluster were taken into account. For membership confirmation, redshift data were obtained from the Sloan Digital Sky Survey (SDSS-DR9) [21] database (for A1656, A2199, and A2147) and the NASA/IPAC Extragalactic Database (NED) (for A2255 and A2144). For the other two clusters A85 and A2029, redshift data were obtained from Agulli *et al.* (2016) [22] and Sohn e*t al.* (2017) [23], respectively. Petrosian magnitudes, taken from the SDSS database, were used for calculating the absolute magnitudes in the five bands *u* (3551Å), *g* (4686Å), *r* (6166Å), *i* (7480Å), and *z* (8932Å). These magnitudes were then corrected for galactic foreground extinction, using values given by Schlafly and Finkbeiner (2011) [24], and, also, K-corrected, using a method given by Chilingarian *et al.* 2010 [25] and Chilingarian and Zolotukhin (2012) [26].

TABLE 1: The basic data of the cluster sample.

Cluster	Equ. J2000.0		Redshift	Velocity dispersion σ (km/s)	Richness class	Bautz–Morgan type
	R.A.	Dec.				
A1656	125, 948.7	+275,850	0.0231	970	2	II
A2199	162, 838.0	+393,255	0.0302	733	2	I
A2147	160, 218.7	+160,112	0.0350	859	1	III
A0085	004, 150.1	−091,809	0.0551	963	1	I
A2029	151, 055.0	+054,312	0.0773	1247	2	I
A2255	171, 231.0	+640,533	0.0806	998	2	II-III
A2142	155, 820.6	+271,337	0.0909	1008	2	II

With both of these corrections taken into consideration, the relation between absolute and apparent magnitudes for any passband can be written as:

$$M = m - 5\log_{10}(D_L) - 25 - K(z) - A_\lambda / \sin(b) \quad (2)$$

where, D_L is the luminosity distance, $K(z)$ is the K correction, A_λ is the galactic foreground extinction, and b is the galactic latitude.

3. RESULTS AND DISCUSSION

It is convenient to write the LF in terms of absolute magnitude, M, rather than luminosity [27]. These two quantities are related through the expression

$$M^* - M = 2.5\log\left(\frac{L}{L^*}\right) \quad (3)$$

Hence, the LF becomes [28]

$$\Phi(M) = (0.4\ln 10)\Phi^* 10^{0.4(\alpha+1)(M^*-M)} * exp\left(-10^{0.4(M^*-M)}\right) \quad (4)$$

where, M^* is the characteristic absolute magnitude corresponding to L^*.

Figures 1-5 show the absolute magnitude distributions of galaxies in the *ugriz* bands, within R_A= 1.5 h^{-1} Mpc, for the whole cluster sample, each fitted with a Schechter function. The fitting is done using the χ^2 – minimization method, and for each case, we vary the magnitude bins until we get the best χ^2 that gives the optimal values of Schechter parameters. Table 2 summarizes the results of the best-fitting Schechter parameters Φ^*, M*, and α, for the whole clusters, in all passbands.

Since Φ^* is just a normalization constant which defines the overall density of galaxies, we focus our attention only on

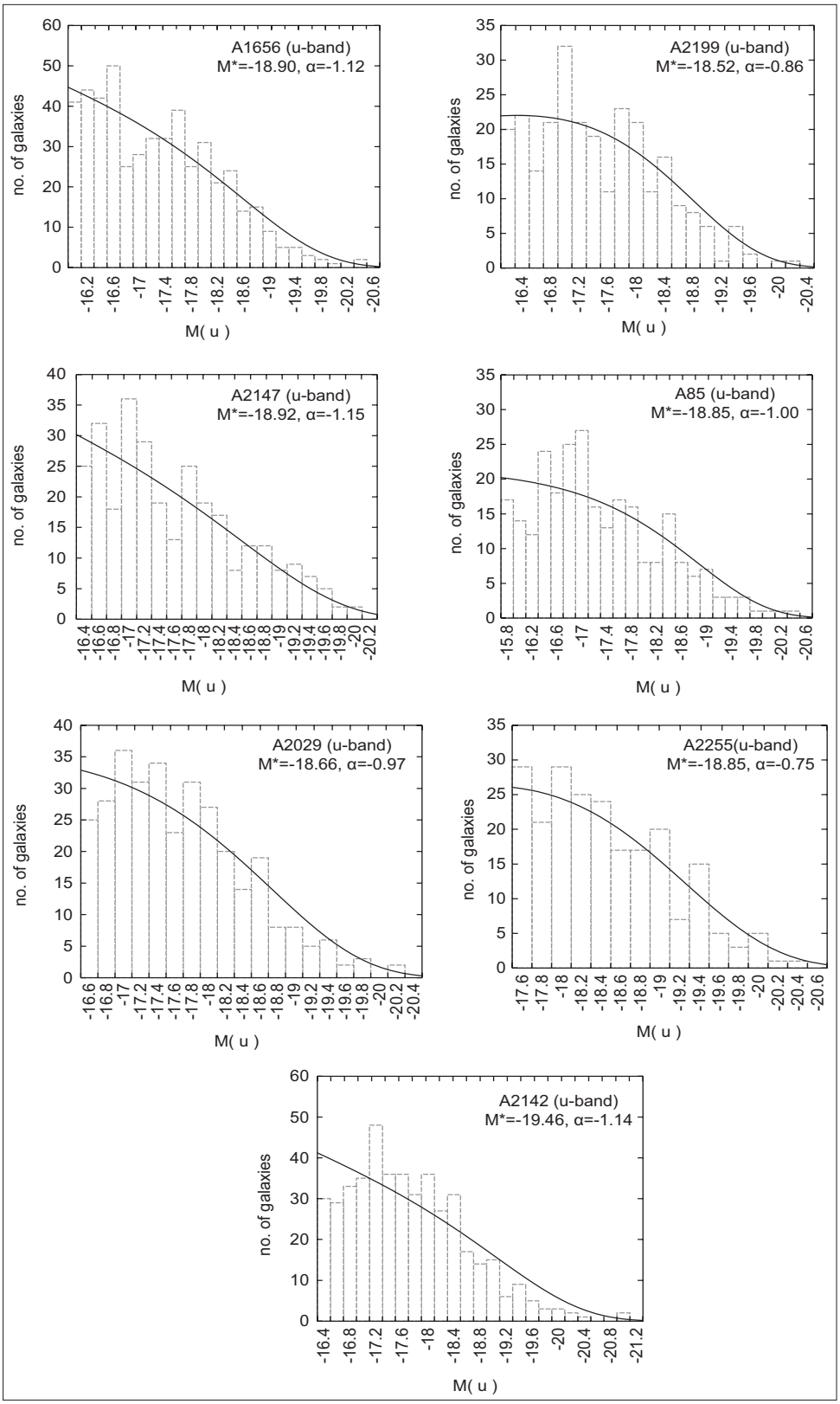

Fig. 1. The luminosity distributions (histograms) of the cluster sample in the u-band, fitted with Schechter functions (solid curves).

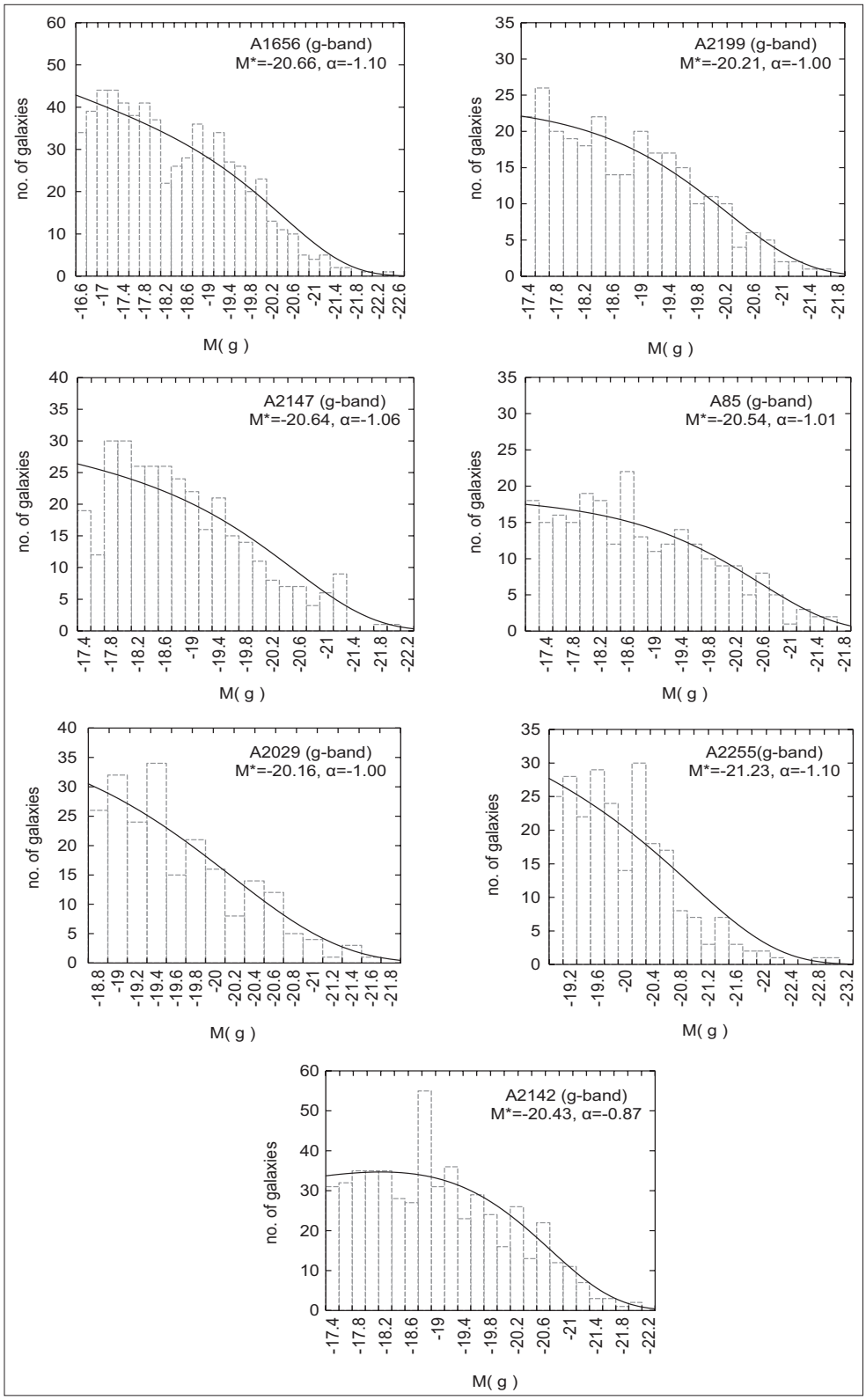

Fig. 2. The luminosity distributions (histograms) of the cluster sample in the g-band, fitted with Schechter functions (solid curves).

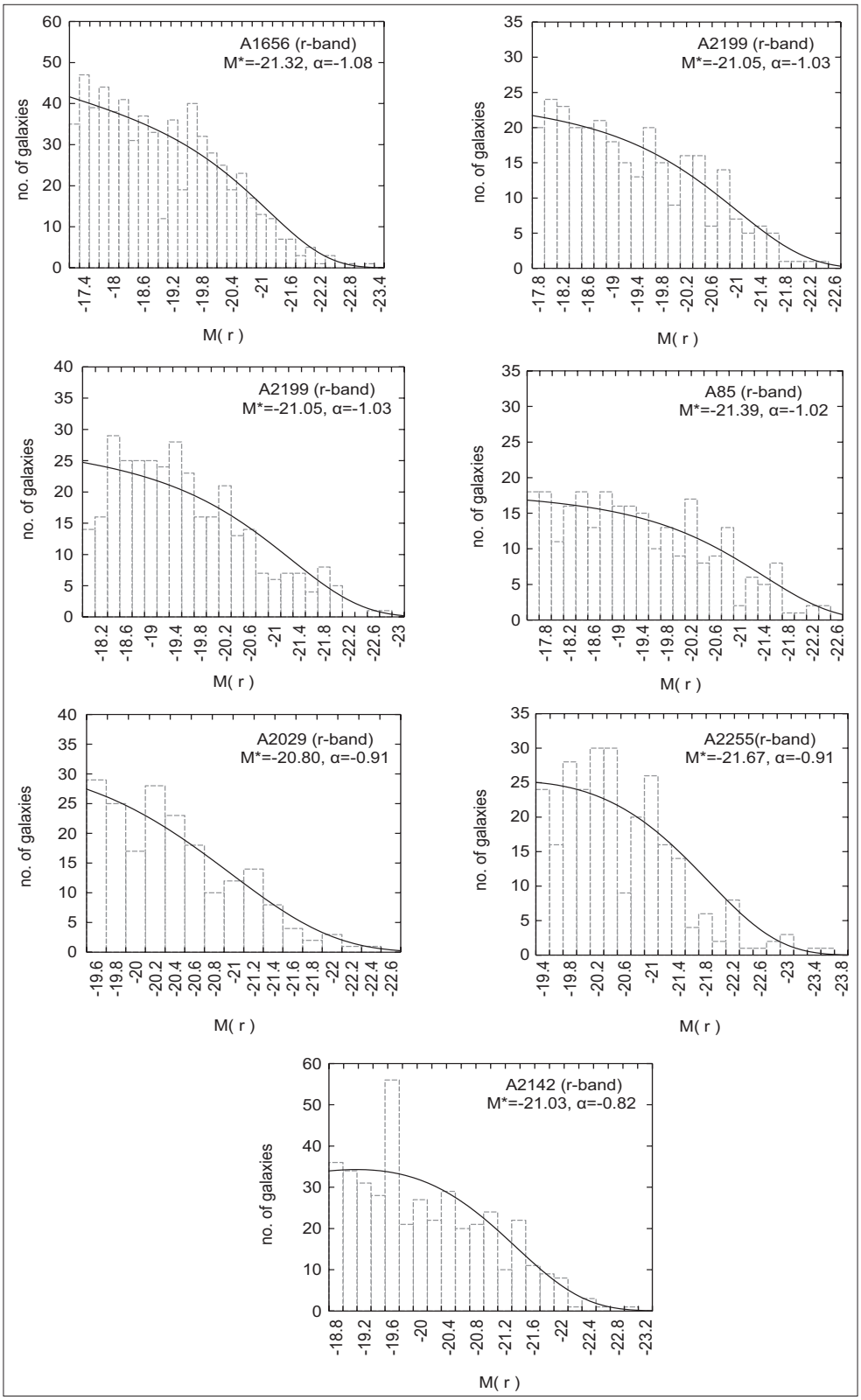

Fig. 3. The luminosity distributions (histograms) of the cluster sample in the r-band, fitted with Schechter functions (solid curves).

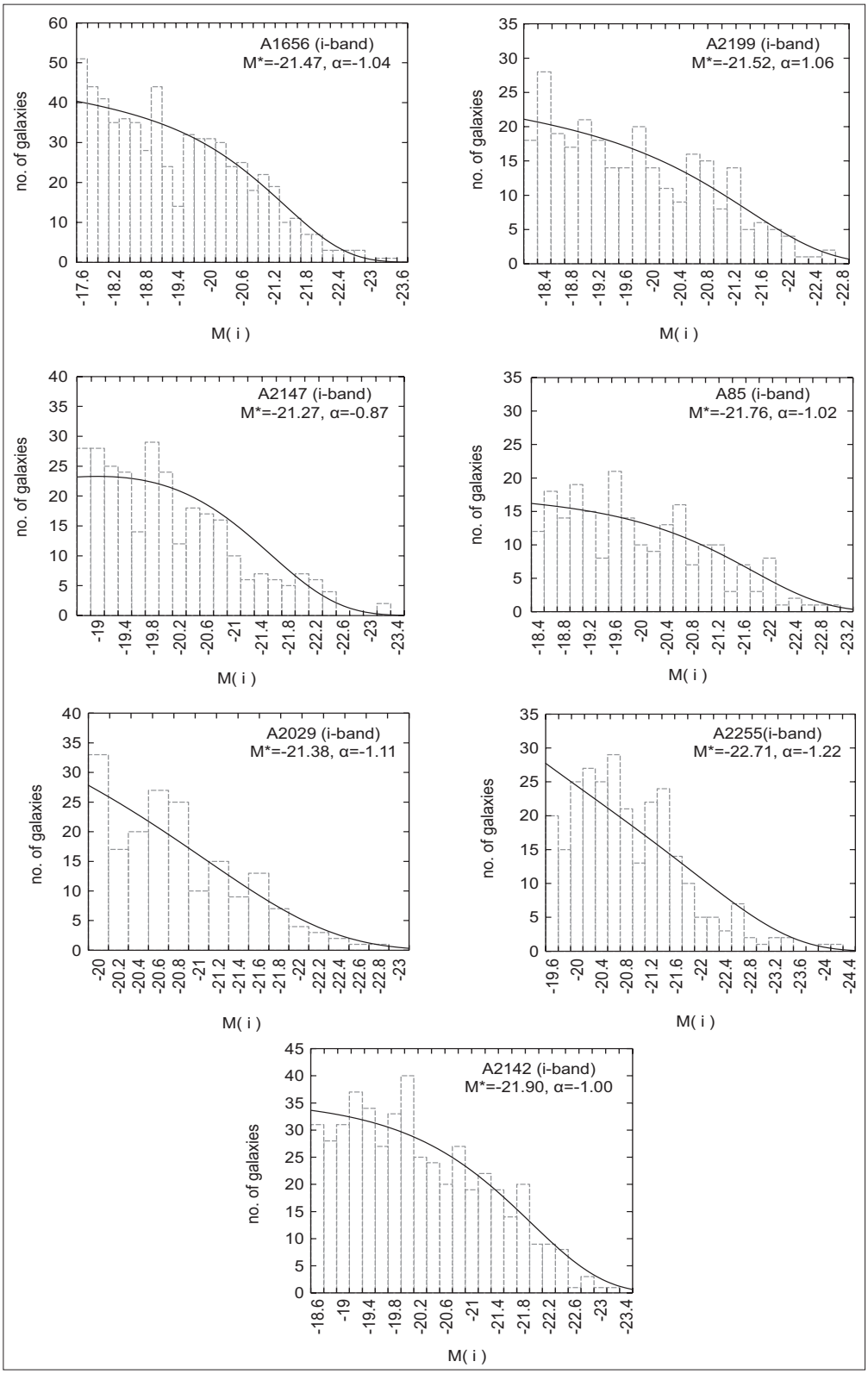

Fig. 4. The luminosity distributions (histograms) of the cluster sample in the i-band, fitted with Schechter functions (solid curves).

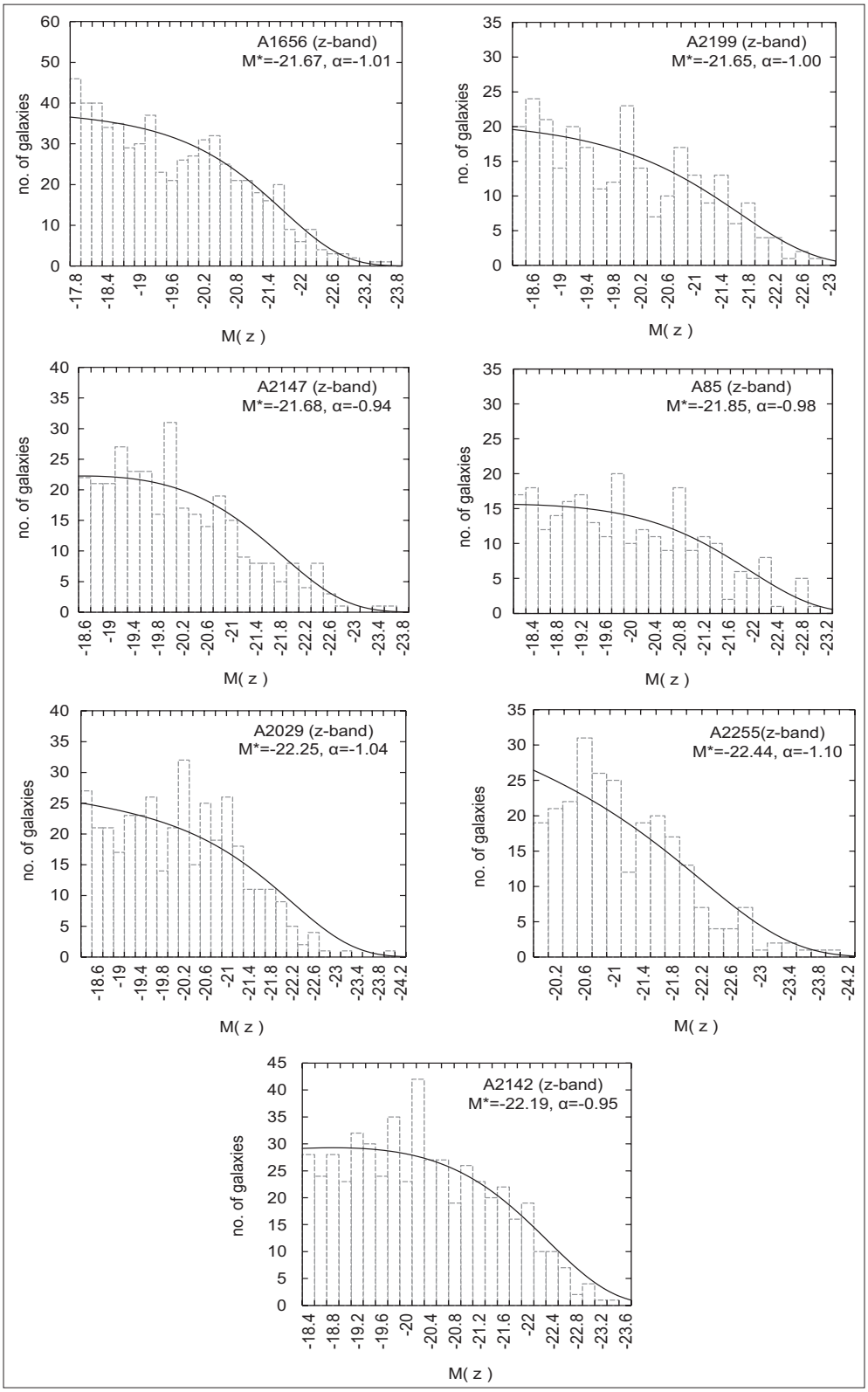

Fig. 5. The luminosity distributions (histograms) of the cluster sample in the z-band, fitted with Schechter functions (solid curves).

the characteristic absolute magnitude, M*, and the faint-end slope α, as the shape of the LF is defined by these two parameters [29]. No remarkable variations are seen in both of these parameters, for all clusters, in each band, within the redshift range considered in this work. Furthermore, by noting the basic data listed in Table 1, we conclude that, in each band, the LF does not vary with such cluster characteristics as velocity dispersion (in agreement with Propis *et al.* [8]), richness (in agreement with Colless [6], Propis *et al.* [8]), and Bautz–Morgan morphology (in agreement with Colless [6], Propis *et al.* [8], Lugger [30]). The above results confirm the universality of the cluster LF, in agreement with several other works (for example, Colless [6]). For this reason, we can deal with the mean values of the Schechter parameters M* and α, for the whole clusters. These values are listed in Table 3.

It is obvious from Table 3 that the characteristic absolute magnitude, M*, becomes brighter towards redder bands, while no remarkable change is noted in the value of the faint-end slope with passband. The reason for this variation of galaxy LF with passband is the contribution of different mechanisms in galaxy evolution. At ultraviolet, for example, the shape of the LF is strongly influenced by star formation since most of the flux is generated by young stars [31]. On the other hand, the LF in the red bands determines the typical stellar distribution [28]. The results in the present work are in good agreement with the previous works [32], [33]. The flat faint-end slope (α~−1) obtained in the present work (Table 3) agrees well with the one obtained by Blanton *et al.* (2003) [32]. This flat faint-end slope is a result of the disruption of a large number of dwarf galaxies inside clusters during the first stages of cluster formation [10]. At the bright end of the LF, the exponential decrease of the number density of galaxies is caused by various feedback processes quenching star formation in massive galaxies. The mechanisms proposed for this quenching are either the effect of supernova explosions or an accreting supermassive black hole. In either case, the gas content is heated and then ejected out of the galaxy, quenching star formation process.

TABLE 2: The best-fitting Schechter parameters for the cluster sample in the *ugriz* bands.

Cluster	Schechter parameters	u-band	g-band	r-band	i-band	z-band
A1656	ϕ*	5.19	4.46	4.69	5.28	5.40
	M*	−18.90	−20.66	−21.32	−21.47	−21.67
	α	−1.12	−1.10	−1.08	−1.04	−1.01
A2199	ϕ*	5.03	3.54	3.15	2.76	2.99
	M*	−18.52	−20.21	−21.05	−21.52	−21.65
	α	−0.86	−1.00	−1.03	−1.06	−1.00
A2147	ϕ*	3.52	3.44	3.48	5.20	4.21
	M*	−18.92	−20.64	−21.31	−21.27	−21.68
	α	−1.15	−1.06	−1.03	−0.87	−0.94
A0085	ϕ*	3.19	2.63	2.42	2.35	2.61
	M*	−18.85	−20.54	−21.39	−21.76	−21.85
	α	−1.00	−1.01	−1.02	−1.02	−0.98
A2029	ϕ*	6.04	6.09	6.30	4.79	3.35
	M*	−18.66	−20.16	−20.80	−21.38	−22.25
	α	−0.97	−1.00	−0.91	−1.11	−1.04
A2255	ϕ*	7.09	3.81	5.07	2.36	3.52
	M*	−18.85	−21.23	−21.67	−22.71	−22.44
	α	−0.75	−1.10	−0.91	−1.22	−1.10
A2142	ϕ*	4.36	7.67	8.37	5.17	5.31
	M*	−19.46	−20.43	−21.03	−21.90	−22.19
	α	−1.14	−0.87	−0.82	−1.00	−0.95

TABLE 3: The mean values of the Schechter parameters M* and α for the cluster sample in the *ugriz* bands.

Band	M*	α
u	−18.88±0.11	−1.00±0.06
g	−20.55±0.13	−1.02±0.03
r	−21.22±0.11	−0.97±0.04
i	−21.72±0.18	−1.05±0.04
z	−21.96±0.12	−1.00±0.02

4. CONCLUSIONS

The galaxy LFs of some nearby clusters were studied in all of the SDSS passbands *ugriz*. In each case, a Schechter function was fitted to the bright end of the distribution, using the χ^2 – minimization technique, to obtain the best-fitting Schechter parameters, Φ*, M*, and α. For each passband, no noticeable variations were observed in the values of M* and α in any cluster. Further, it was found that the LF does not change with such cluster parameters as richness, velocity dispersion, and Bautz–Morgan morphology. From the mean values of M* and α, it was found that M* becomes brighter toward redder bands, whereas no remarkable change was noted in the value of α with passband, being about (−1.00).

5. ACKNOWLEDGMENT

Funding for SDSS-III has been provided by the Alfred P. Sloan Foundation, the Participating Institutions, the National Science Foundation, and the U.S. Department of Energy Office of Science. The SDSS-III web site is http://www.sdss3.org/.

SDSS-III is managed by the Astrophysical Research Consortium for the Participating Brazilian Participation Group, Brookhaven National Laboratory, Carnegie Mellon University, University of Florida, the French Participation

Group, the German Participation Group, Harvard University, the Instituto de Astrofisica de Canarias, the Michigan State/ Notre Dame/JINA Participation Group, Johns Hopkins University, Lawrence Berkeley National Laboratory, Max Planck Institute for Astrophysics, Max Planck Institute for Extraterrestrial Physics, New Mexico State University, New York University, Ohio State University, Pennsylvania State University, University of Portsmouth, Princeton University, the Spanish Participation Group, University of Tokyo, University of Utah, Vanderbilt University, University of Virginia, University of Washington, and Yale University.

This research has made use of the NED which is operated by the JET propulsion Laboratory, California Institute of Technology, under contact with the National Aeronautics and Space Administration.

REFERENCES

[1] P. Schechter. "An analytic expression for the luminosity functions for galaxies. *Astrophysical Journal*, vol. 203, pp. 297-306, 1976.

[2] E. Hubble. "The luminosity function of nebulae. I. The luminosity function of resolved nebulae as indicated by their bright stars". *Astrophysical Journal*, vol. 84, p. 158, 1936.

[3] E. Hubble. "The luminosity function of nebulae. II. The luminosity function as indicated by residuals in velocity-magnitude relations". *Astrophysical Journal*, vol. 84, p. 270, 1936.

[4] F. Zwicky. "On the large scale distribution of matter in the universe". *Physical Review*, vol. 61, pp. 489-503, 1942.

[5] A. Oemler. "The systematic properties of clusters of galaxies. I. Photometry of 15 clusters". *Astrophysical Journal*, vol. 194, pp. 1-20, 1974.

[6] M. Colless. "The dynamics of rich clusters-II. Luminosity functions". *Monthly Notices of the Royal Astronomical Society*, vol. 237, pp. 799-826, 1989.

[7] E. J. Gaidos. "The galaxy luminosity function from observations of twenty Abell clusters". *Astrophysical Journal*, vol. 113, pp. 117-129, 1997.

[8] R. De Propis, M. Colless, S. P. Driver, W. Couch, J. A. Peacock, I. K. Baldry, C. M. Baugh, J. Bland-Hawthorn, T. Bridges, R. Cannon, S. Cole, C. Collins, N. Cross, G. B. Dalton, G. Efstathiou, R. S. Ellis, C. S. Frenk, K. Glazebrook, E. Hawkins, C. Jackson, O. Lahav, I. Lewis, S. Lumsden, S. Maddox, D. S. Madgwick, P. Norberg, W. Percival, B. Peterson, W. Sutherland and K. Taylor. "The 2dF galaxy redshift survey: The luminosity function of cluster galaxies". *Monthly Notices of the Royal Astronomical Society*, vol. 342, pp. 725, 2003.

[9] A. Dressler. "A comprehensive study of 12 very rich clusters of galaxies. I. Photometric technique and analysis of the luminosity function". *Astrophysical Journal*, vol. 223, pp. 765-787, 1978.

[10] O. López-Cruz, H. K. C. Yee, J. P. Brown, C. Jones and W. Forman. "Are luminous CD halos formed by the disruption of dwarf galaxies"? ApJ, vol. 475, p. L97, 1997.

[11] P. Popesso, A. Biviano, H. Böhringer and M. Romaniello. "RASS-SDSS galaxy cluster survey. IV. A ubiquitous dwarf galaxy population in clusters". *Astronomy Astrophysics*, vol. 445, pp. 29-42, 2006.

[12] R. De Propris, P. R. Eisenhardt, S. A. Stanford and M. Dickinson. "The infrared luminosity function of galaxies in the Coma cluster". *Astrophysical Journal*, vol. 503, p. L45, 1998.

[13] L. Cortese, G. Gavazzi, A. Boselli, J. Iglesias-Paramo, J. Donas and B. Milliard. "The UV luminosity function of nearby clusters of galaxies". *Astronomy Astrophysics*, vol. 410, p. L25, 2003.

[14] L. Bai, G. H. Rieke, M. J. Rieke, J. L. Hinz, D. M. Kelly and M. Blaylock. "Infrared luminosity function of the Coma cluster". *Astrophysical Journal*, vol. 639, pp. 827, 2006.

[15] C. A. Valotto, M. A. Nicotra, H. Muriel and D. G. Lambas. "The luminosity function of galaxies in clusters". *Astrophysical Journal*, vol. 479, p. 90, 1997.

[16] M. Yagi, N. Kashikawa, M. Sekiguchi, M. Doi, N. Yasuda, K. Shimasaku and S. Okamura. Luminosity functions of 10 nearby clusters of galaxies. II. Analysis of the luminosity function". *Astrophysical Journal*, vol. 123, p. 87, 2002.

[17] S. M. Hansen, T. A. McKay, R. H. Wechsler, J. Annis, E. S. Sheldon and A. Kimball. "Measurement of galaxy cluster sizes, radial profiles, and luminosity functions from SDSS photometric data". *Astrophysical Journal*, vol. 633, p. 122, 2005.

[18] B. Binggeli, A. Sandage and G. A. Tammann. "The luminosity function of galaxies". *Annual Review of Astronomy and Astrophysics*, vol. 26, pp. 509-560, 1988.

[19] A. Dressler. "Galaxy morphology in rich clusters: Implications for the formation and evolution of galaxies". *Astrophysical Journal*, vol. 236, pp. 351-356, 1980.

[20] G. O. Abell, H. G. Corwin and R. P. Olowin. "A catalog of rich clusters of galaxies". *Astrophysical Journal Supplement Series*, vol. 70, p. 1, 1989.

[21] C. P. Ahn, R. Alexandroff, C. A. Prieto, S. F. Anderson, T. Anderton, B. H. Andrews, É. Aubourg, S. Bailey, E. Balbinot, and R. Barnes. "The ninth data release of the sloan digital sky survey: First spectroscopic data from the SDSS-III baryon oscillation spectroscopic survey". *Astrophysical Journal Supplement Series*, vol. 203, p. 21, 2012.

[22] I. Agulli, J. A. L. Aguerri, R. Sánchez-Janssen, C. Dalla Vecchia, A. Diaferio, R. Barrena, L. Dominguez Palmero and H. Yu. "Deep spectroscopy of nearby galaxy clusters I. Spectroscopic luminosity function of Abell 85". *Monthly Notices of the Royal Astronomical Society*, vol. 458, p. 1590-1603, 2016.

[23] J. Sohn, M. J. Geller, H. J. Zahid, D. G. Fabricant and A. Diaferio. "The velocity dispersion function of very massive galaxy clusters: Abell 2029 and Coma". *Astrophysical Journal Supplement Series*, vol. 229, p. 20, 2017.

[24] E. F. Schlafly and D. P. Finkbeiner. "Measuring reddening with Sloan Digital Sky Survey stellar spectra and recalibrating SFD". *Astrophysical Journal*, vol. 737, p. 103, 2011.

[25] I. V. Chilingarian, A. L. Malchoir and I. Y. Zolotukin. "Analytical approximations of K-corrections in optical and near-infrared bands". *Monthly Notices of the Royal Astronomical Society*, vol. 405, pp. 1409-1420, 2010.

[26] I. Chilingarian and I. Zoloyukin. "A universal ultraviolet-optical colour-colour-magnitude relation of galaxies". *Monthly Notices of the Royal Astronomical Society*, vol. 419, pp. 1727-1739, 2012.

[27] M. S. Longair. *Galaxy Formation*. Springer, Germany, 2008.

[28] P. Schneider. *Extragalactic Astronomy and Cosmology: An Introduction*, Springer, Germany, 2006.

[29] H. Karttunen, P. Kröger, H. Oja, M. Poutanen and K. J. Donner. *Fundamental Astronomy*. Springer, Germany, 2007.

[30] P. M. Lugger. Luminosity functions for nine Abell clusters". *Astrophysical Journal*, vol. 303, pp. 535-555, 1986.

[31] R. De Propris, M. Bremer and S. Phillips. "Luminosity functions of cluster galaxies. The near-ultraviolet luminosity function at < z > ~ 0.05". *Astronomy Astrophysics*, vol. 1807, p. 10775, 2018.

[32] M. R. Blanton, J. Brinkmann, I. Csabai, M. Doi, D. Eisenstein, M. Fukugita, J. E. Gunn, D. W. Hogg and D. J. Schlegel. "Estimating fixed-frame galaxy magnitudes in the Sloan Digital Sky Survey". *Astronomical Journal*, vol. 125, pp. 2348-2360, 2003.

[33] P. Popesso, H. Böhringer, M. Romaniello and W. Voges. "RASS-SDSS galaxy cluster survey. II. A unified picture of the cluster luminosity function". *Astronomy Astrophysics*, vol. 433, pp. 415-429, 2005.

ORIGINAL RESEARCH ARTICLE UHD JOURNAL OF SCIENCE AND TECHNOLOGY

Network Intrusion Detection using a Combination of Fuzzy Clustering and Ant Colony Algorithm

Yadgar Sirwan Abdulrahman

IT Department Kurdistan Technical Institute, Sulaymaniyah, Kurdistan Region, Iraq

ABSTRACT

As information technology grows, network security is a significant issue and challenge. The intrusion detection system (IDS) is known as the main component of a secure network. An IDS can be considered a set of tools to help identify and report abnormal activities in the network. In this study, we use data mining of a new framework using fuzzy tools and combine it with the ant colony optimization algorithm (ACOR) to overcome the shortcomings of the k-means clustering method and improve detection accuracy in IDSs. Introduced IDS. The ACOR algorithm is recognized as a fast and accurate meta-method for optimization problems. We combine the improved ACOR with the fuzzy c-means algorithm to achieve efficient clustering and intrusion detection. Our proposed hybrid algorithm is reviewed with the NSL-KDD dataset and the ISCX 2012 dataset using various criteria. For further evaluation, our method is compared to other tasks, and the results are compared show that the proposed algorithm has performed better in all cases.

Index Terms: Intrusion detection, Data mining, Fuzzy clustering, Ant colony

1. INTRODUCTION

Unusual behavior detection refers to a finding patterns process in a dataset that does not have the expected behavior. Network intrusion is also known as a set of unusual behaviors in the network. Mode detection provides essential information in a variety of applications that will improve network performance.

An intrusion detection system (IDS) is a device or software application that monitors the network to look for suspicious activity, threats, or policy breaching, and on encountering such activities, it alerts the security personnel. IDS monitors inbound as well as outbound network flow for abnormal behavior and then alert the admin or user that a network intrusion might be occurring. It performs the task by comparing signatures of a known malware against the system.

It monitors the user behavior, system processes, and system configurations for any unusual behavior. Security personnel is alerted on security breaches with data consisting of the addresses of the source, the target, and the type of attack.

The problem of intrusion detection is a complicated issue. A compromise must be made between detection accuracy, detection speed, intrinsic network dynamics, and high data volume for processing, and the methods used must be able to distinguish between state and abnormal behaviors. Be normal behaviors in the network. The IDS's primary purpose can be considered network display for any mode such as DoS, U2R, R2L, some of which are listed in Table 1 [1].

Access this article online

DOI: 10.21928/uhdjst.v5n2y2021.pp11-19

E-ISSN: 2521-4217

P-ISSN: 2521-4209

Corresponding author's e-mail: Yadgar Sirwan Abdulrahman, *IT Department Kurdistan Technical Institute, Sulaymaniyah, Kurdistan Region, Iraq.* E-mail: yadgar.abdulrahman@kti.edu.krd

Receied: 01-04-2021 Accepted: 07-07-2021 Published: 16-07-2021

TABLE 1: The network attack types classification.

Attack class	Attack name	Attack description
Probing	Probe	An attacker performs port scanning and monitoring activities to gather information or find vulnerabilities in a network
DoS	Denial of Services	An attacker fills a busy network resource (such as memory or bandwidth) with repeated requests, causing network resources to overflow and users' requests not to be answered
User to root	U2R	An attacker accesses a regular account and searches for a vulnerability to gain unauthorized root access to the system
R2L	Remote to Local	An attacker gains access to a system through a remote network and attempts to gain unauthorized local access through a remote system

Among these modes, distributed denial of service (DDoS) is one of the security threats associated with computer networks, especially the internet, which targets access to network resources, and the purpose of this mode is to disrupt network service. One of the most dangerous and most recent situations on the internet is not to disrupt the service, but to force the network and server to be unable to provide regular service to target network bandwidth. It is done with a victim who drowns the victim's network or processing capacity in information packets and prevents users and customers from accessing the service. One of the most common and significant threats on the internet today is a denial of service by interfering with configuration information. Routers and IP source fraud occurs, leading to reduced network performance.

For years, experts have warned about the poor security of internet-connected devices and equipment, and poor security has made them vulnerable configuration of equipment and the heterogeneity of operating systems such devices a very convenient yet easy target for attackers. One of the main exploits of hackers and destroyers of these devices and equipment is to capture them to execute the distributed model. During this state, an army of these hacked devices bombards it by sending simultaneous requests to the victim's server, which is called this type of hacked equipment with a net. Receiving a request from thousands and sometimes tens of thousands of devices with different IP addresses at the same time will eventually lead to slowing down or even stopping the server service to users. When a DDOS attack occurs, the first step is to determine what layer of the open systems interconnection model the attack is on; the mode is usually on layers 3 and 4 of the network and the scope of an attack depends on features such as volume and the number of packets sent per second. Layers 3 and 4 are very difficult to control. Dispose of it. The issue of identifying and providing a suitable solution is one of the biggest challenges facing network security professionals.

The methods of diagnosis and prevention that have been presented so far have either not been effective or have not been adequately responded to by attackers with increasing level of knowledge, and most of the detection is in the form of statistical methods and monitoring and control of network traffic. In the case of this type of attack and high traffic, if two attacks occur with different traffics combined in the network, this type of method will no longer be a good answer us, and because the attack speed and the network traffic volume is very high in a short time, it should be possible to detect and deal with the attack as soon as possible, in other words, the system notices the denial of service when the network attack increases and affects traffic volume and it is no longer possible to deal with, and when it detects an attack on the traffic volume network, it will be more careful with the defense layers on the internet, before the defense layer service can react. For example, if we can detect the mode in network routers and these hand-held routers can make a proper diagnosis, the probability of service denial mode is reduced and further risks are avoided. In this problem, the attack detection importance in the lower layers of the network, such as the network layer and the data link layer, is seen more, and in the higher layers, such as the application layer, it requires data packets to be examined, which will take many of our resources and time.

A computer network attack detection system is one of the most important parts to prevent illegal intrusions in the network. Detecting and detecting intrusion can reduce the misuse of individuals' personal information as well as prevent financial risks for users and service companies. Various algorithms and methods for this classification have been proposed in the previous works, each of which has its advantages and disadvantages. In this study, we present our proposed framework along with the improvements in the algorithms used to distinguish DoS/DDoS mode from normal network mode in the ISCXIDS2012 dataset, which are fully described in Section 3. This study tries to provide a suitable framework for attack detection using fuzzy clustering and feature selection.

This paper's structure is as follows: Section 2, we mentioned an overview of the related work. Section 3 describes the

problem and the proposed method. Section 4 will present the experiments and results and finally, the conclusion is stated in Section 5.

2. RELATED WORK

Much work has been done in the field of intrusion detection in computer networks, and in this section, we will briefly mention some of these studies.

In Chitrakar and Chuanhe [2], to solve the problem of high data volume requirements in works that use k-means or k-means clustering and Forward Neural Network, a combination of support vector machine (SVM) with k-means clustering The Kyoto 2006+ dataset is used in this work and the simulation results show that the use of SVM in any volume of data has higher classification accuracy. To evaluate the operations performed in this work, sensitivity criteria. And False Alarm has been used in Chitrakar and Chuanhe [2], to detect network intrusion, the combined method of K-means clustering with Naive Bayes classification with the same working criteria and the same dataset (Kyoto 2006+ dataset) has been used. The results of this work were somewhat weaker than the work done in Chitrakar and Chuanhe [2]. In Saifullah [3], they propose a defense mechanism to detect an attack using a distributed algorithm that runs a moderate load valve in the opposite direction of the router. The valve has a medium load because the traffic intended for the server is controlled [3]. The operation (increase or decrease) is performed using a perforated bucket in the router based on the number of connected users, who are directly or indirectly connected to the router. At the beginning of the algorithm, the remaining capacity is underestimated by the router. The remaining initialization capacity is the minimum or normal value at the beginning of the algorithm. The speed is updated (increase or decrease), sends to small routers based on server feedback, and finally multiplies all routers in descending order. The convergence of the whole server is loaded with an acceptable capacity range.

In Syarif *et al.* [4] there are three objectives: (1) effective feature selection and dimension reduction, –(2) a strong algorithm selection in the classification field, and –(3) unconventional detection, using clustering algorithms based on segmentation; to achieve the first goal genetic algorithm and particle swarm optimization (PSO) are used. For the classification operation, the nearest neighbor classification method has been used, and finally, by comparing different types of clustering methods, the expectation maximization method has had the best performance. The results for false-positive and accuracy criteria have been investigated using four classification methods, the highest accuracy being related to the decision tree. In Revathi and Malathi [5], just like [4] methods of optimizing collective intelligence have been used. This work uses simplified collective intelligence, which is a kind of simplified and improved PSO algorithm (SSO) along with the Random Forest method on the KDDcup 1999 dataset. In Singh and Singh [6] an attack detection method is presented in MANET based on the ACO algorithm. In this work, after the intrusion detection by the ant colony algorithm, the genetic algorithm has been used to retrieve the network, in which the number of recovered nodes has been investigated for the number of 10–80 replications and the probability of mutation between 0.2 and 0.4.

In both papers Cheng *et al.* [7] and Xia *et al.* [8] IP Address Interaction (IAI), the algorithm is proposed due to sudden traffic changes, the interaction between addresses, the asymmetry between many addresses, source distribution, and focus of targeted goals, IAI algorithm is designed to describe the network stream validity important features. The SVM classifier, which is sorted by IAI interval with normal attack current, is used to classify the current network stream validity and identify DDOS mode. The method is defined in real-time attack flow detection as well as attacker assessment power based on fuzzy reasoning. The process consists of two steps: (a) Statistical analysis of network traffic interval and (b) DDOS attack identification and evaluation based on intelligent fuzzy reasoning mechanism.

In Kamarudin *et al.* [9], the feature selection was performed using the Random forest genetic algorithm without an observer and a subset of the total features for each of the two datasets DARPA1999 and ISCX2012 was obtained. The use of random forest classification has been performed. In Vargas-Munoz *et al.* [10], an IDS system based on the Bayesian network is also presented. Eight features, the ICCX2012 feature set is used in the classification section.[11] Feature selection is based on Entropy. In their work, the raw features of the ISCX2012 dataset and five features based on Entropy values are considered. They classify MLP neural network, RNN (Alternative decision tree) ADT has been used.

3. PROPOSED METHOD

In this study, we present our proposed framework along with the improvements in the algorithms used to distinguish DoS/DDoS attacks in normal network mode in the ISCXIDS2012

dataset. All the steps to achieve attack detection using data mining and artificial intelligence can be summarized in the following sections.

- Data preprocessing
- Combined clustering using ant colony optimization and fuzzy clustering
- Classification and review of quantitative criteria.

After performing these steps, the constructed approach can be used. The most important feature of this study is to present a model for DDoS mode detection that improves the accuracy of the combination of the fuzzy c-means algorithm and the ant colony optimization and improves this algorithm in clustering, which improves the detection accuracy. In the following, the main steps of the proposed method are discussed.

3.1. Data Collection and Preprocessing

The study uses two datasets, NSL-KDD and ISCX datasets, which are described in the following. The number of training instances in each attack class is shown in both KDD Train (KDD cnp99) and NSL-KDD(KDD Train+) datasets in Table 2. The NSL-KDD dataset also includes two training datasets. KDD Train+ And KDD Train+_20% of which KDD Train+_20% is the improved version of KDD Train+. Test examples for the NSL-KDD dataset also include two sets Test KDD+ and KDD Test-21. KDD Test-21 has more difficulty distinguishing samples than KDD Test+ As can be seen, most of the samples removed from KDD CUP are in DOS mode with a removal rate of 82.98%. NSL-KDD is obtained by removing approximately 43.97% of the samples in the KUPD CUP99 dataset. In total, the NSL-KDD dataset has 25,192 samples and 43 features. (This dataset is comprised four sub-datasets: KDDTest+, KDDTest-21, KDDTrain+, and KDDTrain+_20Percent, although KDDTest-21 and KDDTrain+_20Percent are subsets of the KDDTrain+ and KDDTest+. From now on, KDDTrain+ will be referred to as train and KDDTest+ will be referred to as a test. The KDDTest-21 is a subset of the test, without the most difficult traffic records (Score of 21), and the KDDTrain+_20Percent is a subset of the train, whose record count makes up 20% of the entire train dataset. That being said, the traffic records that exist in the KDDTest-21 and KDDTrain+_20Percent are already in test and train, respectively, and are not new records held out of either dataset.)

TABLE 2: NSL-KDD data information

Class	KDDCUP'99 (KDD Train)	NSL-KDD (KDDTrain+)	% Reduction	% in NSL KDD
NORMAL	972,781	67,343	93.07	53.46
DOS	3,883,370	45,927	98.82	36.46
PROBE	41,102	11,656	71.64	9.25
U2R	52	52	0	0.04
R2L	1126	995	11.63	0.79
TOTAL	4,898,431	125,973	97.43	

As mentioned earlier, this article also uses the ISCX dataset [12]. The structure of the network used to generate this dataset is shown in Fig. 1. As shown in Fig. 1, the test structure consists of four separate LANs, and the fifth LAN consists of servers that provide web, email, DNS, and NAT services. All links are set on 10M bits/s.

The data began on Friday, June 11, 2020, and lasted exactly 7 days. This article examines the DDoS mode detection performed on Tuesday compared to the normal network mode (no attack) performed on Friday. Given that this dataset is available in pcap format, we use CICFlowMeter software. Together with winpcap software, we have used to extract 24 features. Then specify the data path in pcap format and the CSV data storage path to obtain user data in the preprocessing section.

4. CLUSTERING

4.1. Ant Colony Optimization Algorithm (ACOR) Algorithm

We discuss the design process of the ant colony optimization algorithm of the continuous domain for solving unconstrained optimization problems and constrained optimization problems based on the position distribution model of ant colony foraging [13].

Assuming the whole ant colony consists of m groups of substructure, each group contains n of ants. As shown in the following equation:

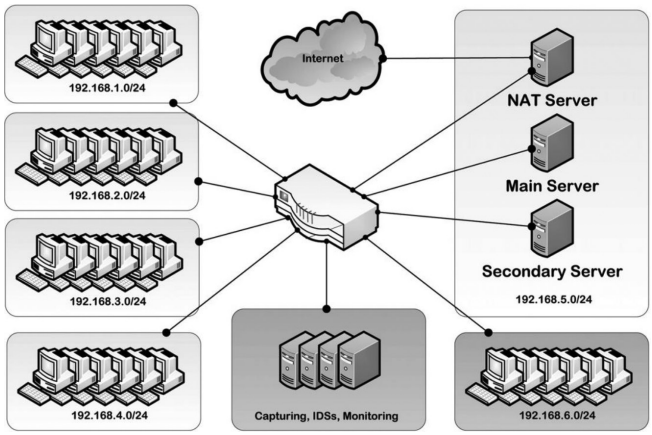

Fig. 1. Data generation network structure [12].

$$\begin{bmatrix} x_1 & x_2 & \dots & x_n \\ ant_{11} & ant_{12} & \dots & ant_{1n} \\ ant_{21} & ant_{22} & \dots & ant_{2n} \\ \vdots & \vdots & & \vdots \\ ant_{m1} & ant_{m2} & \dots & ant_{mn} \end{bmatrix} \quad (1)$$

the position ant_{ij} corresponding to the value x_j of the variable for j-ant in any sub colony i, the sub colony I of all the ants in the sequence of $\{ant_{i1}, ant_{i2},\dots, ant_{in}\}$ represents a solution of the optimization problem.

In the position distribution model of ant colony foraging, each ant releases pheromone according to the quality of a food source of their position; pheromones are dispersed in the entire space, with increasing distance of the source and the concentration decreasing. Therefore, we need to choose a probability density function as the distribution model of ant pheromone in the optimization algorithm of continuous domains. The Gaussian function is a common probability density function; we assume ants of the ant colony release pheromone externally on the function. At this point, j ant in any sub colony ant i corresponding to pheromone distribution model $\tau_{ij}(x)$ can be expressed as

$$\tau_{ij}(x) = \frac{1}{\sqrt{2\pi}\sigma_j} e^{-((x-\mu_{ij})^2/2\sigma_j^2)},$$
$$\sigma_j = \frac{(u_j - l_j)}{\Psi(1+1n(n))}, \quad (2)$$

where μ_{ij} is the position ant_{ij} of ant j in the sub colony of ants i, namely, the distribution center, $\sigma j(\sigma j>0)$ means the width of the distribution function, u_j is the maximum allowable value of the variable x_j, l_j is the minimum allowable value of the variable x_j, n is the dimension of solution for the optimization problem, ψ ($\psi > 0$) is a parameter, and σj is used to adjust the size.

Before updating the position of the ant colony, we need to choose a group as a parent from m sub colony. First, we use formula (3) to calculate each group of sub colony corresponding to the assessed value of the solution. Consider the following:

$$eval_i = \frac{1}{(1+e^{f(ant_{i1}, ant_{i2},\dots ant_{in})/T})}, \quad (3)$$

Where $f(ant_{i1}, ant_{i2},\dots, ant_{in})$ is the assessment value of the sub colony ant i; T (T>0) is the adjustment coefficient used to adjust the pressure of selection.

After the assessment value for each group of sub colony is obtained, we calculate the selected probability for each group of sub colony according to

$$pi = \frac{eval_i}{\sum_{j=1}^{m} eval_j}. \quad (4)$$

Finally, we select parent colony c according to formula (5)

$$\begin{cases} \arg\max(eval_i), & q \le q_0, \\ \quad i=1,2,\dots m & \\ C & q>q_0, \end{cases} \quad (5)$$

Where $(0\le q0\le 1)$ is a given parameter, q is a random variable is distributed in [0,1] uniformly. C is a random variable that is generated according to formula (5).

After getting the parent ant colony c, the ant pheromone distribution model function $\tau_{cj}(x)$ in the ant colony corresponding to random number generator for sampling, the k groups of children colony are generated. Then, according to the size of assessment value for each group of sub colony, we select the large assessment value of m group from (m+k) group of sub colony to achieve a position of ant colony update.

4.2. Basic Fuzzy C-means

In this section, the basic fuzzy C-means algorithm [14] will be briefly introduced. The objective function of this algorithm is defined as bellow:

$$J = \sum_{i=0}^{N} \sum_{j=1}^{C} \mu_{ij}^{m} d^2 (xi - vj) \quad (6)$$

The μ_{ij} determines the degree to which the i_th sample belongs to the center of the cluster j, and m determines the fuzzy degree. Here d^2 $(xi\text{-}vj)$ is the non-euclidean distance equal to $(xi\text{-}vj)^2$. As x_i is the i_th sample and v_j is the center of the j_th cluster. For this objective function, there are constraints $0 < \sum_{i=1}^{N} \mu_{ij} < N$ and $\mu_{ij} \varepsilon$ [0-1]. The values of the variables i and j are in the range of $1\le i\le$N and $1\le j\le$C.

Based on the objective function introduced in equation (6), equations for improving the centers and functions of the affiliation will be as follows:

$$\mu_{ij} = 1 / \sum_{l=1}^{C} \left(\frac{d^2(x_i - v_j)}{d^2(x_i - v_l)}\right)^{\frac{1}{(m-1)}}$$
$$v_j = \left(\sum_{i=1}^{N} \mu_{ij}^{m} x_i\right) \Big/ \sum_{i=1}^{N} \mu_{ij}^{m} \quad (7)$$

$$v_j = \left(\sum_{i=1}^{N} \mu_{ij}^{m} x_i\right) \Big/ \sum_{i=1}^{N} \mu_{ij}^{m} \quad (8)$$

4.3. ACOR Improvement

As mentioned before, the X matrix is the original data matrix with dimensions N*D, where N is the number of observations and D is the number of attributes. The output of the first layer is the X Matrix with dimensions N*R where R≤D is the selected properties. The second layer is responsible for clustering data in K clusters. The proposed method in this layer is to combine the fuzzy c-means clustering algorithm with the ACOR optimization algorithm. Furthermore, in the ACOR [13] algorithm and fuzzy c-means algorithm [14], changes have been applied to improve the performance of the proposed method, which is stated.

In the ACOR algorithm, to determine the weights, an exponential relationship is used, which causes a high difference for the weight of the answers with high and low rank. In other words, high-ranking answers are very important in each iteration, which reduces the breadth of finding the answer and catching the algorithm in the local minimum. For this reason, in this study, we have proposed the following weighting function to determine the weights. In this case, as the answer rank increases, the number of weight increases, but these changes occur much smoother than the case suggested in Socha and Dorigo [13]. The proposed function for means and standard deviation (σ = kq) (different) is shown in Fig. 2. As can be seen, it is an S-shape function. The cumulative distribution function (CDF) of the normal, or Gaussian, distribution with standard deviation σ and mean μ. And then, which is τ_{ij} (x) will be calculated as described below.

Instead of using formula 2 in, and from kq we can rewrite 2nd formula as formula 9. Then for a smoother objective function due to arithmetic estimation mentioned below. As in reference 15, we can compute erf(x). Its not a new formula. Its using an approximation function of the main objective function of ACOR algorithm. That gets us a smoother result in practice Now we can use this CDF estimation for formula (2) which is a PDF or Gaussian function:

$$Ae^{-B\left(\frac{x-\mu}{\sigma}\right)^2} \approx \frac{1}{2}\left(1+\operatorname{erf}\left(\frac{x-\mu}{\sigma\sqrt{2}}\right)\right) \quad (9)$$

Here, if we consider A=1/√2πσ and B=1/2, then from using formula (9) and (2) we have:,

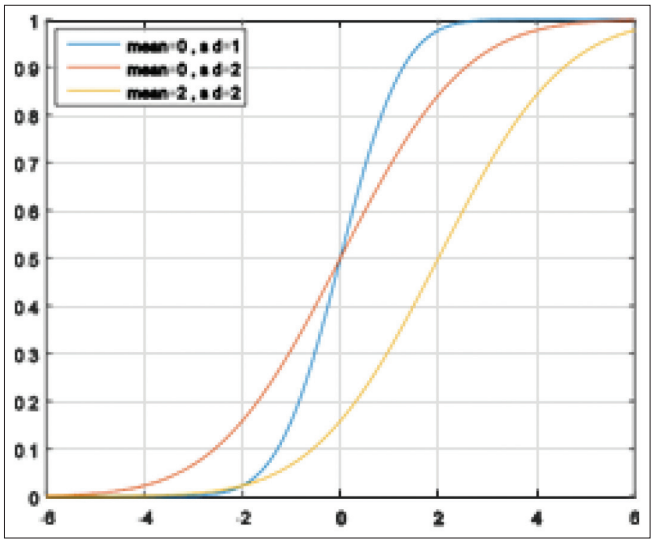

Fig. 2. Weighting functions for means and different standard deviation.

$$\tau_{ij} = \frac{1}{2} + erf\left(\frac{x-\mu_{ij}}{\sigma\sqrt{2}}\right) \quad (10\text{-a})$$

As we know then we can replace (σ = kq) to and reach

$$\tau_{ij} = \frac{1}{2} + erf\left(\frac{x-\mu_{ij}}{qk\sqrt{2}}\right) \quad (10\text{-b})$$

And we can calculate erf(x) as we have this in [15],

$$erf(x) = \frac{2}{\sqrt{\neq}}\int_0^x e^{-t^2} dt \quad (11)$$

The combination of ACOR and fuzzy c-means algorithms can be achieved in three ways. As follows:

- In the first case, the selection of cluster centers can be done by the ACOR algorithm in such a way that this algorithm calculates cluster centers based on the defined objective function; these centers are then applied to the fuzzy c-means clustering algorithm as the initial mean
- In the second case, after random generation of the first states, the fuzzy c-means algorithm of events is executed and according to the desired population in successive iterations, the centers of the clusters are obtained and then these centers are given to the ACOR algorithm as the initial population and this algorithm is based on the function. The defined goal performs the clustering operation. This will usually not produce the desired result
- In the third case, the ACOR algorithm starts clustering the data, with the difference that at the same time the

fuzzy c-means algorithm is applied and improves the location of the best available country. This method requires more time than the previous two modes.

In this paper, we used the first case, as discussed above. We use the ACOR algorithm to cluster center selection; these centers are then applied to the fuzzy c-means clustering algorithm as the initial mean. And as an improvement of ACOR weighting function, instead of formula (2) in [13] which is a PDF function, we used a smoother arithmetic estimation function or CDF function from formula (9). Then the weighting function will become as formula (10-a), by using (σ = kq), we'll get to the formula (10-b). And to calculate erf(x) we can use the formula (11). Now we have a new smoother weighting function as mentioned in formula (10-b), which can be easily calculated.

5. SIMULATION RESULTS

In this study, we used python with Pycharm IDE for implementation, and an HP laptop with 8gigabyte RAM, core i7 6600u CPU, windows 10. The first case is used and the results obtained with PSO-k-means, ICA-k-means, k-means, Fuzzy c-means ++, and DBSCAN methods and methods proposed in Kumar and Kumar [16], Kaur *et al.* [17] and Soheily-Khah *et al.* [18]. It should be noted that when using optimization methods for clustering, the defined objective function must be a function appropriate to the clustering problem. Defined for clustering, in which the objective function of the k-means problem is used, which is as follows.

$$Cost(c) = \sum_{i=1}^{N} \min\{\|xi - Cj\|\} \qquad j = 1,2,\ldots,K$$

Where the selected distance is between the sample and the center of the cluster.

To present and compare the results in this section, the training curves of hybrid algorithms along with the correct data clustering rate (D.R), Accuracy, and False Alarm have been calculated.

$$Detection\ Rate = TP/TP + FN \tag{12}$$

$$Accuracy = TP + TN/TP + TN + FP + FN \tag{13}$$

$$False\ Alarm = FP/FP + TN \tag{14}$$

The parameters used for the ACO algorithm are given in Table 3. Furthermore, the parameters of the Fuzzy c-means algorithm are obtained by the colonial competition optimization algorithm with the Davis Boldin clustering cost function.

Fig. 3 shows the training curve for the proposed method. As it turns out, the proposed method has the best performance in minimizing the cost function.

Table 4 shows all the parameters of relationships 2–3 for all methods using the ISCX dataset along with a comparison of the methods Kumar and Kumar [16] and Soheily-Khah *et al.* [18] and MBGWO [19] with the proposed method. In

TABLE 3: Parameters used in ACO Iteration algorithm (MAX)

Iteration(MAX(	Population	#of antes	α	k	Mu
1000	100	15	1	2	0.1

TABLE 4: Clustering evaluation indicators for different methods using all 24 attributes in the ISCX dataset

Clustering methods	Accuracy	False alarm rate	Detection rate
Proposed method	99.93	0.04	99.55%
ICA-Fuzzy c-means	97 %	0.06	96.2%
PSO k-means	94.6%	0.06	94.1%
Fuzzy c-means++	91.4%	0.08	91%
k-means	67.45%	0.12	67%
DBSCAN	68.67%	0.12	68%
(Kumar, 2013) method	95.2%	0.07	94.5%
(Soheily-Khah, 2018) method	99.91%	0.05	99.51%
MBGWO (M. Alzubi1 2019)	99.22	0.0064	99.10

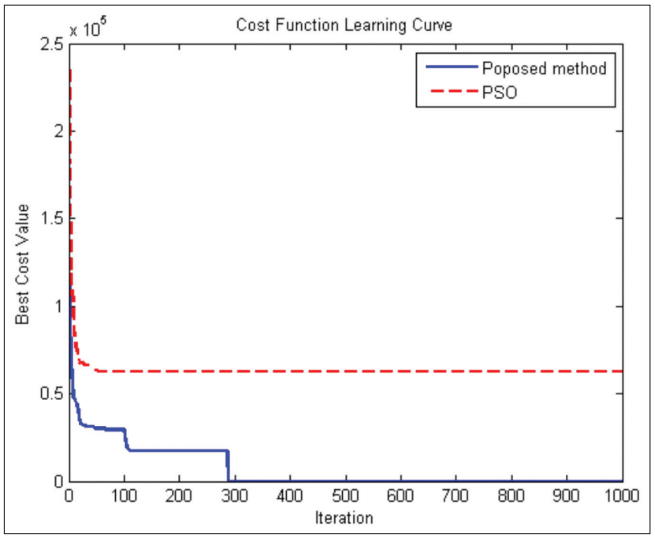

Fig. 3. Training curves for the proposed method and single-particle swarm optimization algorithm.

TABLE 5: Clustering evaluation indicators for different methods in the NSL-KDD dataset

Clustering methods	Accuracy	False alarm rate	Detection rate
Proposed method	86.98%	0.14	78%
ICA-Fuzzy c-means	84%	0.18	74.52%
PSO k-means	81.16%	0.212	70.1%
Fuzzy c-means++	74.14%	0.219	69.21%
k-means	63.35%	0.31	57%
DBSCAN	65.37%	0.31	59%
(Kaur, 2017) method	82.1%	0.211	72.57%

calculating the criteria, the total number of performances is Nt=10.

As shown in Table 4, the proposed method in clustering has shown the best performance in accuracy and detection rate indicators.

The following results are reviewed for the NSL-KDD dataset. In this dataset, the number of clusters is equal to four types of attacks and normal mode (a total of five clusters) has been considered. For this dataset, all the available features have been considered in the clustering section and no further reduction has been made. Table 5 shows all the parameters of relationships 2–3 for all methods using the NSL-KDD dataset with a comparison of the method [17] with the proposed method.

As can be seen in this case, the combined algorithm Fuzzy c-means and ACO for continuous domains give better results for three metrics met with a slight detection of the condition.

Tables 6 and 7 show the comparison of proposed algorithm to some recent deep network IDSs. Both NSL-KDD and ISCX dataset are included in the study. The algorithms are described in the table.

In Table 6, there are BGRU [20], long short-term memory (LSTM) [21], and ANN [22] accuracy calculated on NSL-KDD dataset. And two deep algorithms have better performance than proposed algorithm.

In Table 6, there are Deep CNN [23], LSTM [24], and predicting future tokens [25] accuracy calculated on NSL-KDD dataset. And two deep algorithms have better performance than proposed algorithm.

Experiments on ISCX dataset shows that the proposed method does well and is a bit better than the three deep algorithms compared.

TABLE 6: Clustering evaluation indicators for different methods in the NSL-KDD dataset

Algorithm	Accuracy	description
Jiang *et al*. (2019)	98.94	Training multiple long short term memory nets (one hidden layer) for different features extracted
Xu *et al*. (2018)	99.24	5-class classification GRU and Bidirectional GRU (BGRU) nets. Model has one layer with 128 GRU nodes, 3 feed-forward layers with 48 nodes BGRU gives best results with fast convergence
Vinayakumar *et al*. (2019)	78.5	ANN(shallow neural network) has five hidden layers with 1024, 768, 512, 256, and 128 nodes. ReLU activation
Proposed method	86.98	

TABLE 7: Clustering evaluation indicators for different methods in the ISCX dataset

Algorithm	Accuracy	Description
Zeng *et al*. (2019)	99.85	5-class classification Deep CNN (convolutional nural network): 2 1D convolutional layers, 1 fully connected layers
Chilamkurti (2018a)	99.91	Binary classification 30 embedding layers, 10 LSTM (long short term memory) layers, and sigmoid output layer
Radford *et al*. (2018)	97.01	anomaly detection by predicting future tokens (unsupervised) Token embedding layer
Proposed method	99.93	

6. CONCLUSION

In this study, the proposed framework for identifying DDoS is discussed. At first, the necessary preprocessing was performed on the data, and then the state diagnosis was performed using a combined fuzzy clustering algorithm and compared to some other methods. The results showed that, the proposed method of this study, has shown better performance in the quantitative criteria considered at most. In comparison to both classic methods and deep methods for intrusion detection, our proposed method is doing well on ISCX dataset, but for NSL-KDD dataset deep algorithms shown better performance.

As future work we should try to improve our proposed method to discover attacks, or if it will be possible, extend it to a deep method.

REFERENCES

[1] M. Mazini, B. Shirazi and I. Mahdavi. "Anomaly network-based intrusion detection system using a reliable hybrid artificial bee colony and AdaBoost algorithms". *Journal of King Saud University-Computer and Information Sciences*, vol. 31, no. 4, pp. 541-553, 2019.

[2] R. Chitrakar and H. Chuanhe. "*Anomaly Based Intrusion Detection Using Hybrid Learning Approach of Combining K-Medoids Clustering and Naïve Bayes classification*". IEEE, United States, 2012.

[3] A. Saifullah. "*Defending Against Distributed Denial-of-Service Attacks With Weight-Fair Router Throttling*", 2009. Available from: https://www.openscholarship.wustl.edu/cse_researchhttps://www.openscholarship.wustl.edu/cse_research/23. [Last accessed on 2021 May 10].

[4] I. Syarif, A. Prügel-Bennett and G. Wills. "Data mining approaches for network intrusion detection: From dimensionality reduction to misuse and anomaly detection". *Journal of Information Technology Review*, vol. 3, no.2, pp. 70-83, 2012.

[5] S. Revathi and A. Malathi. "Data Preprocessing for Intrusion Detection System using Swarm Intelligence Techniques". *International Journal of Computer Applications*, vol. 75, no. 6, pp. 22-27, 2013.

[6] K. Singh and K. Singh. "Intrusion detection and recovery of MANET by using ACO algorithm and genetic algorithm". *Advances in Intelligent Systems and Computing*, vol. 638, pp. 97-109, 2018.

[7] J. Cheng, C. Zhang, X. Tang, V. S. Sheng, Z. Dong and J. Li. "Adaptive DDoS attack detection method based on multiple-kernel learning". *Security and Communication Networks*, vol. 2018, p. 5198685, 2018.

[8] Z. Xia, S. Lu, J. Li and J. Tang. "Enhancing DDoS flood attack detection via intelligent fuzzy logic". *Informatica*, vol. 34, no. 4, pp. 497-507, 2010. Available from: http://www.informatica.si/index.php/informatica/article/view/323. [Last accessed on 2021 May 11].

[9] M. H. Kamarudin, C. Maple, T. Watson and N. S. Safa. "A new unified intrusion anomaly detection in identifying unseen web attacks". *Security and Communication Networks*, vol. 2017, p. 2539034, 2017.

[10] M. J. Vargas-Munoz, R. Martinez-Pelaez, P. Velarde-Alvarado, E. Moreno-Garcia, D. L. Torres-Roman and J. J. Ceballos-Mejia. "Classification of network anomalies in flow level network traffic using Bayesian networks". In: *2018 28th International Conference on Electronics, Communications and Computers, CONIELECOMP 2018*, vol. 2018, pp. 238-243, 2018.

[11] A. Koay, A. Chen, I. Welch and W. K. G. Seah. "A new multi classifier system using entropy-based features in DDoS attack detection". In: *International Conference on Information Networking*, vol. 2018, pp. 162-167, 2018.

[12] A. Shiravi, H. Shiravi, M. Tavallaee and A. A. Ghorbani. "Toward developing a systematic approach to generate benchmark datasets for intrusion detection". *Computers and Security*, vol. 31, no. 3, pp. 357-374, 2012.

[13] K. Socha and M. Dorigo. "Ant colony optimization for continuous domains". *European Journal of Operational Research*, vol. 185, no. 3, pp. 1155-1173, 2008.

[14] J. C. Bezdek. "*Pattern Recognition with Fuzzy Objective Function Algorithms*". Springer, United States, 1981.

[15] L. C. Andrews. "*Special Functions of Mathematics for Engineers*", 2021. Available from: https://www.books.google.nl/books?id=2caqsf-rebgc and pg=pa110 and redir_esc=y#v=onepage&q&f=false. [Last accessed on 2021 Jun 02].

[16] G. Kumar and K. Kumar. "Design of an evolutionary approach for intrusion detection". *The Scientific World Journal*, vol. 2013, p. 962185, 2013.

[17] A. Kaur, S. K. Pal and A. P. Singh. "Hybridization of K-means and firefly algorithm for intrusion detection system". *International Journal of Systems Assurance Engineering and Management*, vol. 9, no. 4, pp. 901-910, 2018.

[18] S. Soheily-Khah, P. F. Marteau and N. Bechet. "Intrusion detection in network systems through hybrid supervised and unsupervised machine learning process: A case study on the iscx dataset". In: *Proceedings-2018 1st International Conference on Data Intelligence and Security*, *ICDIS 2018*, pp. 219-226, 2018.

[19] Q. M. Alzubi, M. Anbar, Z. N. M. Alqattan, M. A. Al-Betar and R. Abdullah. "Intrusion detection system based on a modified binary grey wolf optimisation". *Neural Computing and Applications*, vol. 32, no. 10, pp. 6125-6137, 2020.

[20] C. Xu, J. Shen, X. Du and F. Zhang. "An intrusion detection system using a deep neural network with gated recurrent units". *IEEE Access*, vol. 6, pp. 48697-48707, 2018.

[21] F. Jiang, Y. Fu, B. B. Gupta, F. Lou, S. Rho, F. Meng and Z. Tian. "Deep learning based multi-channel intelligent attack detection for data security". *IEEE Transactions on Sustainable Computing*, vol. 5, no. 2, pp. 204-212, 2020.

[22] Y. Bengio, A. Courville and P. Vincent. "Representation learning: A review and new perspectives". *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 35, no. 8, pp. 1798-1828, 2013.

[23] Y. Zeng, H. Gu, W. Wei and Y. Guo. "Deep-full-range: A deep learning based network encrypted traffic classification and intrusion detection framework". *IEEE Access*, vol. 7, pp. 45182-45190, 2019.

[24] A. Diro and N. Chilamkurti. "Leveraging LSTM networks for attack detection in fog-to-things communications". *IEEE Communications Magazine*, vol. 56, no. 9, pp. 124-130, 2018.

[25] B. J. Radford, L. M. Apolonio, A. J. Trias and J. A. Simpson. "*Network Traffic Anomaly Detection Using Recurrent Neural Networks*", 2018. Available from: http://arxiv.org/abs/1803.10769. [Last accessed on 2021 Jun 08].

ORIGINAL RESEARCH ARTICLE UHD JOURNAL OF SCIENCE AND TECHNOLOGY

Comparison of Different Ensemble Methods in Credit Card Default Prediction

Azhi Abdalmohammed Faraj[1,2], Didam Ahmed Mahmud[1], Bilal Najmaddin Rashid[1]

[1]Department of Information Technology, College of Commerce, University of Sulaimani, Sulaimani, Iraq, [2]Department of Computer Engineering, College of Engineering, Dokuz Eylül Üniversitesi, İzmir, Turkey

ABSTRACT

Credit card defaults pause a business-critical threat in banking systems thus prompt detection of defaulters is a crucial and challenging research problem. Machine learning algorithms must deal with a heavily skewed dataset since the ratio of defaulters to non-defaulters is very small. The purpose of this research is to apply different ensemble methods and compare their performance in detecting the probability of defaults customer's credit card default payments in Taiwan from the UCI Machine learning repository. This is done on both the original skewed dataset and then on balanced dataset several studies have showed the superiority of neural networks as compared to traditional machine learning algorithms, the results of our study show that ensemble methods consistently outperform Neural Networks and other machine learning algorithms in terms of F1 score and area under receiver operating characteristic curve regardless of balancing the dataset or ignoring the imbalance

Index Terms: Ensemble methods, Credit card default prediction, Balanced and imbalanced dataset, Stacking and XGBoosting, Neural networks

1. INTRODUCTION

In the aftermath of the Global Financial Crisis of 2008–2009, many who took mortgages defaulted when they could not pay leading to many credit card issuers routinely encountering a credit debt crisis. Numerous occasions of over-issuing credit cards to unfit candidates have raised concerns. Concurrently a considerable percentage of cardholders regardless of their repayment capabilities heavily relied on credit cards and resulted in heavy credit debts. This has negatively affected banks and consumer confidence.

Access this article online

DOI: 10.21928/uhdjst.v5n2y2021.pp20-25

E-ISSN: 2521-4217
P-ISSN: 2521-4209

The problem of credit card defaulting is binary classification problem applicants will either default or repay their credit debts, however determining the probability of defaulting from the perspective of risk management offers more value than a result of a binary classification [1]. Improving the accuracy of fraudulent activities by only one percent can have a major impact on reducing the loss of financial institutions [2].

The aim of a credit default detection model is to solve the problem of categorizing loan customers into two groups: good customers (those who are expected to pay off their full loans in a already agreed upon time period) and bad customers (those who might default on their payments). Customers who pay their bills on time are more likely to repay their loans on time, which benefits banks. Bad customers, on the other hand, can cost you money. As a result, banks and financial institutions are increasingly focusing on the development of credit scoring models,

Corresponding author's e-mail: Azhi Abdalmohammed Faraj, Department of Information Technology, College of Commerce, University of Sulaimani, Sulaimani, Iraq/Department of Computer Engineering, College of Engineering, Dokuz Eylül Üniversitesi, İzmir, Turkey.
E-mail: azhi.faraj@univsul.edu.iq

Received: 13-03-2021 Accepted: 26-06-2021 Published: 19-07-2021

as even a 1% improvement in the quality of bad credit applicants will result in substantial potential savings for financial institutions. Therefore; organizations and scholars have conducted extensive research on credit score models, which is a significant financial management practice. Several studies have discussed the superiority of ensemble learning, as new machine learning models are proposed. Ensemble learning has been incorporated into the application of credit scoring [3].

Ensemble learning is a machine learning technique in which several machine learning algorithms are trained and combined to generate a final output that is superior to individual algorithm outputs. Ensemble learning strategies are divided into two types: Homogeneous and heterogeneous ensembles. Each base learner form is built in a different way using various machine learning techniques in the heterogeneous ensemble technique. The final forecast and the same dataset are generated by statistically combining each individual base learner prediction. Each base learner is used on different subsets of the entire training dataset in homogeneous ensemble techniques. To satisfy requirements and achieve a good ensemble, two necessary and critical conditions must be met: diversity and accuracy [4].

This research aims to answer three questions, first how well ensemble methods work on credit default predictions? Second how do they compare to NN and other traditional algorithms when used on skewed datasets? Third how does balancing the dataset affect the relative performance gain in Ensemble methods?

The ensemble techniques used in this research are Bagging, Boosting (AdaBoosting and XGBoosting), Voting, and random forests (RF).

1.1. Related Work

Advances in technology and the availability of big data have helped researcher improve results on Machine Learning in credit scoring, default prediction, and risk evaluation. Since the purpose of credit management is to improve the business performance and decrease the associated risk, rules must be established to make credit decisions. Hence, clustering algorithm is widely used in the credit default detection systems in the early stage. For instance, William and Huang combined the K-means clustering method with the supervision method for insurance risk identification [5].

Researchers in Saia *et al.* [6] performed credit scoring to detect defaults using the Wavelet transform combined with three metrics three different datasets were used in their experimentation the authors compared their results with RF and improved on RF; however, state of the art results is achieved using neural networks and to get a better perspective neural networks approach needed to be included. The work in Saia and Carta [7] transformed the canonical time domain representation to the frequency domain, by comparing differences of magnitudes after Fourier Transform conversion of time-series data. The authors in Ceronmani Sharmila *et al.* [8] applied an outlier-based score for each transaction, together with an isolation forest classifier to improve default detection. Authors of Zhang *et al.* [9] used data preprocessing and a RF optimized through a grid search step, the feature selection step while preparing the data helped to improve the accuracy of RF.

In Zhu *et al.* [10], deep learning was utilized for the 1st time by applying convolutional neural networks (CNN) approach through the transformation of features to gray scale images, their R-CNN model improved on the area under curve (AUC) of RF and logistic regression (LR) by around 10%. A thorough analysis of different neural networks, such as Multilayer Perceptron and CNNs for credit defaulting can be found in Neagoe *et al.* [11].

Ensemble learning techniques have previously been applied in different credit-related topics for example [12] used RF and majority voting to classify transactions by European cardholders in September 2013 [13], used majoring voting by combining support vector machine (SVMs) and LR, to validate a feature selection approach, called group penalty function the research mainly focuses on robustness of the models. Wang *et al.* [14] used bagging and boosting for credit scoring, Ghodselahi [15] used a hybrid SVM ensemble for binary classification of credit default predictions. The work in Zhang *et al.* [16] ensembles five classifiers (LR, SVMs, neural network, gradient boosting decision tree, and 6 RF) using a genetic algorithm and fuzzy assignment. In Feng *et al.* [17], a set of classifiers are joined in an ensemble according to their soft probabilities. In Tripathi *et al.* [18], an ensemble is used with a feature selection step based on feature clustering, and the final result is a weighted voting approach.

1.2. Overviews of Ensemble Learning

The ensemble methods seek to enhance model predictability by integrating several models to create one stable model. By training several models to train a meta-estimator, ensemble learning aims to enhance predictive efficiency. Base estimators or base learners are considered the component models of an ensemble. The strategies of the ensemble

exploit the influence of "the wisdom of crowds," which is focused on the idea that a community's collective judgment is more powerful than any person in the group. Ensemble techniques are widely used in various fields of application, including economic and business analytics, medicine and health insurance, information security, education, industrial production, predictive analytics, entertainment, and many more. Many machine-learning algorithms deal with a tradeoff of fit versus uncertainty (also known as bias-variance), which affects their ability to generalize potential knowledge accurately. To solve this tradeoff, ensemble approaches use multiple models. Two essential components are required for an effective ensemble: (1) Ensemble diversity and (2) model aggregation for the final predictions [19], [20].

1.3. Bagging

Bagging is primarily used in classification and regression, the short form for bootstrap aggregation. By utilizing decision trees, it improves the precision of models, and to a large degree decreases uncertainty. The reduction of variance increases accuracy, hence eliminating overfitting, which is a challenge to many predictive models [19]. Using bootstrapped replicas of the training data, diversity in bagging is acquired: different training data subsets are randomly drawn from the entire training data with replacement. To train a different base learner of the same type, each training data subset is used. The combination strategy of the base learners for bagging is the majority vote. Simple as it is, when combined with the basic learner generation strategies, this strategy can decrease variance. Bagging is particularly attractive when the data available is limited in size. Relatively large portions of the samples (75–100%) are drawn into each subset to ensure that there are sufficient training samples in each subset. This causes a significant overlap of individual training subsets, with many of the same instances appearing in most subsets, and some instances appearing in a given subset multiple times. A relatively unstable base learner is used to ensure diversity under this scenario, so that sufficiently different decision limits can be obtained for small disturbances in different training datasets [21].

1.4. Boosting and RF

Boosting is a form of machine-learning as well. Whereas bagging and RF use autonomous learning, sequential learning is used for boosting. In boosting method, by integrating multiple instances into a more reliable estimation, the simple concept is to improve the precision of a poor classification method [22].

RF is a decision tree-based ensemble learning algorithm. It is simple to implement and can be used for both regression and classification tasks. The bootstrap method is used by RF to collect samples from the original results. Every tree assigns a classification, and the forest selects the classification that receives the most votes among all trees. The degree of randomness is determined by the parameter m, which is the number of decision trees. The borrower is presumed to have d attributes in the RF [23].

Random fore effects of the classification produced from multiple datasets of training are organized and combined to improve the accuracy of the prediction. However, bagging uses all input variables to build each decision tree, RF uses subsets to create each decision tree that are random samplings of variables. This means that forest randomness is best adapted for high-dimensional data processing than bagging [24].

1.5. Stacking

Stacking, another tactic of the ensemble, is also known as stacked generalization. This approach works by allowing many other related learning algorithm predictions to be put together by a training algorithm. Regression, density calculations, distance learning, and classifications have been widely applied by stacking. It may also be used during bagging to calculate the error rate involved [25].

2. MATERIALS AND METHODOLOGY

2.1. The Dataset

The dataset contains information about 30,000 consumers for each consumer 23 attributes marked X1 to X23 [Table 1] are stored. The dependent variable represents whether a customer has defaulted (1) or repaid (0). All the client's data are recorded in September 2005 in Taiwan. As with all types of risk assessment datasets, the ratio of positive to negative samples causes a major imbalance in the dataset, in this dataset, only 22% of the clients have defaulted. There are no missing values in the dataset however there are 35 duplicated rows in the dataset, these have been removed.

- X1: Amount of the given credit (NT dollar): It includes both the individual consumer credit and his/her family (supplementary) credit
- X2: Gender (1 = male; 2 = female)

TABLE 1: Results of the imbalanced dataset

Ensemble methods	Accuracy	Precision	Recall	F1
Neural network	82.01	66.85	36.73	47.41
Bagging	79.43	55.45	34.62	42.62
Ada boost	81.83	68.06	33.33	44.75
XGBoosting	82.11	68.16	35.6	46.77
Voting ensemble	81.88	68.32	33.41	44.87
Stacking	81.86	65.73	37.26	47.56

- X3: Education (1 = graduate school; 2 = university; 3 = high school; 4 = others)
- X4: Marital status (1 = married; 2 = single; 3 = others)
- X5: Age (year)
- X6–X11: History of past payment. We tracked the past monthly payment records (from April to September 2005); as follows: X6 = the repayment status in September 2005 X7 = the repayment status in August 2005 X11 = the repayment status in April 2005. The measurement scale for the repayment status is: -1 = pay duly; 1 = payment delay for 1 month; 2 = payment delay for 2 months; 8 = payment delay for 8 months; 9 = payment delay for 9 months and above
- X12–X17: Amount of bill statement (NT dollar). X12 = amount of bill statement in September 2005; X13 = amount of bill statement in August 2005 X17 = amount of bill statement in April, 2005
- X18–X23: Amount of previous payment (NT dollar). X18 = amount paid in September 2005; X19 = amount paid in August 2005. X23 = amount paid in April 2005.

2.2. Evaluation Metrics

The dataset used in this research is imbalanced if this is not handled then accuracy will not provide a meaningful result because even if the model only predicts the output to be 0 it will still get 78% accuracy regardless of the dependent features. It can be presumed that those responsible for issuing these credit cards believed that every cardholder will not default otherwise it would not have been issued in the first place, thus we can conclude that the human level accuracy for this dataset is approximately 78%, This is an example of when a machine learning performs better than humans. It should be noted that misclassifying a positive example as negative will have higher cost and damage than predicting a negative class to be positive.

This means that the model with better performance on the positive cases should be preferred. Some of the common metrics for classification include accuracy, precision, recall, receiver operating characteristic (ROC), and AUC [6]. All these common metrics will be presented for each model.

In the context of credit card default recall means out of all defaulters how many did the model get correct while precision measures the correctness of the model based on its predictions. F1 score is the harmonic mean of recall and precision. In this research, all the common metrics will be presented however for the assessment of ensemble methods we will focus on the F1 score.

$$\text{Accuracy} = \frac{TP + TN}{TP + FP + TN + FN}$$

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

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

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

2.3. Methodology

We used the steps shown in Fig. 1 for each of the ensemble methods mentioned in section 1. In addition, LR and decision trees were also used however because their results were outperformed by neural networks, we opted to not include them in the results section and decided to use NN as a benchmark for performance comparison. To measure the effects of imbalance on the data all algorithms have also been tested after the down sampling of the datasets their results have included in subsequent sections.

3. RESULTS AND DISCUSSION

The results of the ensemble learning were recorded in two separate trials, first with the original imbalanced dataset and second after the imbalance aspect were eliminated.

When the ratio of positive samples to negative sample is approximately 82% accuracy cannot be used as a reliable measure and as shown in Table 1 all models retrieve an accuracy of around 80% which is equivalent to predicting the performance

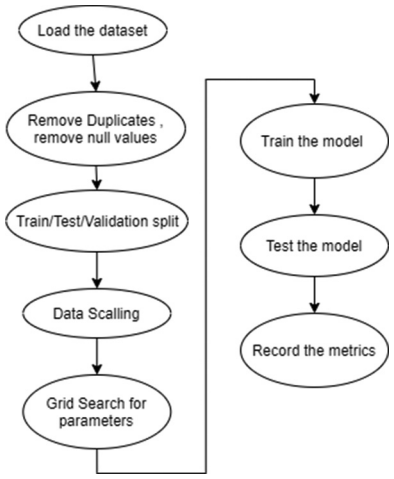

Fig. 1. Proposed method.

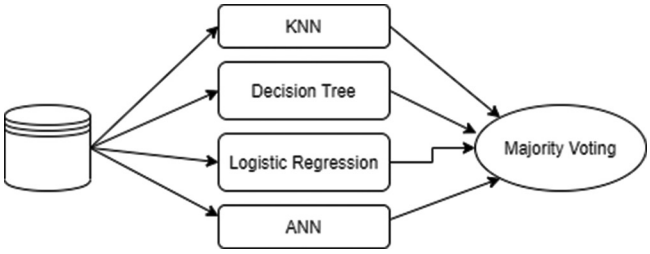

Fig. 2. Voting ensemble used in this research.

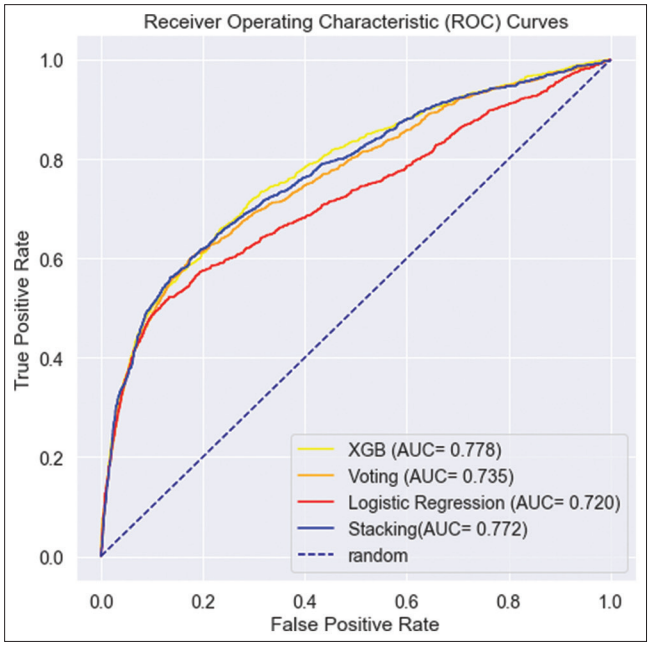

Fig. 3. Receiver operating characteristic curve for imbalanced dataset.

of ensemble methods as compared to regular prediction methods, a variety of other methods were tested such KNN, LR, decision trees, and neural networks. Neural networks performed the best as also confirmed in Cheng Yeh and Lien [1], Hand and Henley [2]. Therefore, for comparison purposes, the results of the artificial neural network are also presented with the ensemble methods for both cases. Fig. 2 show the structure of the voting ensemble used in this study, additionally, for the stacking ensemble, the same algorithms were used in the first level and later LR was applied as the final estimator. In both cases, the data were scaled using a min-max scaler.

3.1. The Imbalanced Dataset

Not default for everyone and consequently is the same as human-level error. A better metric would be the F1 score which is the harmonic mean of recall and Precision Fig. 3. Stacking produced the best result which is 47.56 marginally better than the 47.41 of neural networks. In terms of area

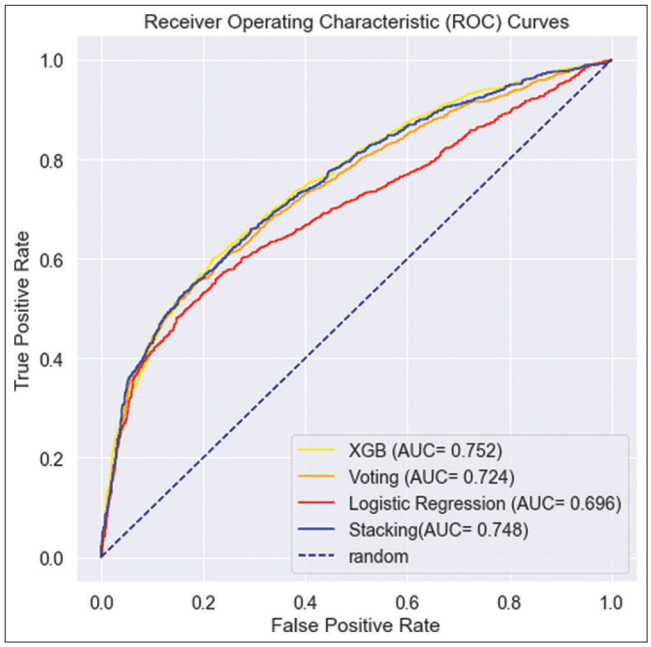

Fig. 4. Receiver operating characteristic curve for the balanced dataset.

TABLE 2: Results of the balanced dataset

Ensemble methods	Accuracy	Precision	Recall	F1
Neural network	68.53	71.9	59.86	65.33
Bagging	64.84	65.79	60.49	63.02
Ada boost	68.49	71.47	60.56	65.57
XGBoosting	68.76	71.42	61.58	66.13
Voting ensemble	67.75	72.57	56.1	63.28
Stacking	68.22	72.58	57.59	64.22

under the ROC curve Stacking and XGBoosting produced the best results.

3.2. The Balanced Dataset

For balancing the dataset down sampling was used since there are 6630 positive samples, the same number of negative samples was kept, and the rest was discarded. The samples were randomly shuffled before feeding them to ANN and Ensemble methods. Since the dataset is balanced now accuracy can also be taken into account as shown in Table 2 we can see that XGBoosting is slightly outperforming all the others in all the metrics. XGBoosting is also the fastest in terms of time consumption. Fig. 4 shows the the ROC curves for the balanced dataset, in Which XGBoosting produced the best result.

4. CONCLUSION

The Credit default prediction using ML algorithms has a crucial role in many financial situations including personal

loans, insurance policies, etc. However, establishing a model that improves the previous rule-based predictions is weakened by the data imbalance problem in datasets, where the number of unreliable cases is quite smaller than the number of reliable cases.

In this paper, we examine different ensemble methods for credit card default prediction in an imbalanced dataset and compare the results with neural networks. Most research in the literature have either focused on the balanced dataset or a skewed one however we have included both in scenarios to provide a better perspective of the performances of each used algorithm. We tested the results first without altering the imbalance aspect of the dataset in which we used AUC as a metric and ignored accuracy and later by down sampling the majority class. Our experiments show that XGBoosting performs better in both cases as compared to other ensemble methods and also better than neural networks.

REFERENCES

[1] I. Cheng Yeh and C. H. Lien. "The comparisons of data mining techniques for the predictive accuracy of probability of default of credit card clients". *Expert Systems with Applications*, vol. 36, no. 2, pp. 2473-2480, 2009.

[2] D. J. Hand and W. E. Henley. "Statistical classification methods in consumer credit scoring: A review". *Journal of the Royal Statistical Society*, vol. 160, no. 3, pp. 523-541, 1997.

[3] Y. Li and W. Chen. "A comparative performance assessment of ensemble learning for credit scoring". *Mathematics*, vol. 8, no. 10, p, 1756, 2020.

[4] M. Akour, I. Alsmadi and I. Alazzam. "Software fault proneness prediction: A comparative study between bagging, boosting, and stacking ensemble and base learner methods". *International Journal of Data Analysis Techniques and Strategies*, Vol. 9, No. 1, pp. 1-16, 2017.

[5] G. Williams and Z. Huang. "Mining the knowledge mine: The hot spots methodology for mining large real world databases". In: *Proceedings of the 10th Australian Joint Conference on Artificial Intelligence, Perth, Australia*, 1997.

[6] R. Saia, S. Carta and G. Fenu. "A wavelet-based data analysis to credit scoring". In: *ICDSP 2018: Proceedings of the 2nd International Conference on Digital Signal Processing, ACM, 2018*, pp. 176-180, 2018.

[7] R. Saia and S. Carta. "A fourier spectral pattern analysis to design credit scoring models". In: *Proceedings of the 1st International Conference on Internet of Things and Machine Learning, ACM*, p. 18, 2017.

[8] V. Ceronmani Sharmila, K. K. R., S. R., S. D. and H. R. "Credit card fraud detection using anomaly techniques". In: 2019 1st International Conference on Innovations in Information and Communication Technology (ICIICT), pp. 1-6, 2019.

[9] X. Zhang, Y. Yang and Z. Zhou. "A novel credit scoring model based on optimized random forest". In: *2018 IEEE 8th Annual Computing and Communication Workshop and Conference* (*CCWC*), pp. 60-65, 2018.

[10] B. Zhu, W. Yang, H. Wang and Y. Yuan. "A hybrid deep learning model for consumer credit scoring". In: *2018 International Conference on Artificial Intelligence and Big Data* (*ICAIBD*), pp. 205-208, 2018.

[11] V. Neagoe, A. Ciotec and G. Cucu. "Deep convolutional neural networks versus multilayer perceptron for financial prediction". In: *2018 International Conference on Communications* (*COMM*), pp. 201-206, 2018.

[12] I. Sohony, R. Pratap and U. Nambiar. "Ensemble learning for credit card fraud detection". In: *Proceedings of the ACM India Joint International Conference on Data Science and Management of Data*, 2018.

[13] J. Lpez and S. Maldonado. "Profit-based credit scoring based on robust optimization and feature selection". *Information Sciences*, vol. 500, pp. 190-202, 2019.

[14] G. Wang, J. Hao, J. Ma and H. Jiang. "A comparative assessment of ensemble learning for credit scoring". *Expert Systems with Applications*, vol. 38, no. 1, pp. 223-230, 2011.

[15] A. Ghodselahi. "A hybrid support vector machine ensemble model for credit scoring". *International Journal of Computer Applications*, vol. 17, no. 5, pp. 1-5, 2011.

[16] H. Zhang, H. He and W. Zhang. "Classifier selection and clustering with fuzzy assignment in ensemble model for credit scoring". *Neurocomputing*, vol. 316, pp. 210-221, 2018.

[17] X. Feng, Z. Xiao, B. Zhong, J. Qiu and Y. Dong. "Dynamic ensemble classification for credit scoring using soft probability". *Applied Soft Computing*, vol. 65, pp. 139-151, 2018.

[18] D. Tripathi, D. R. Edla, V. Kuppili, A. Bablani and R. Dharavath. "Credit scoring model based on weighted voting and cluster based feature selection". *Procedia Computer Science*, vol. 132, pp. 22-31, 2018.

[19] P. Bühlmann. "Bagging, boosting and ensemble methods". In: J. Gentle, W. Härdle and Y. Mori, (eds.), *Handbook of Computational Statistics*. Springer Handbooks of Computational Statistics. Springer, Berlin, Heidelberg, 2012.

[20] G. Kunapuli. "*Ensemble Methods for Machine Learning*". MEAP Publication, Shelter Island, New Work, 2020.

[21] S. Hamori, M. Kawai, T. Kume, Y. Murakami and C. Watanabe. "Ensemble learning or deep learning? Application to default risk analysis". *Journal of Risk and Financial Management*, vol. 11, p. 12, 2018.

[22] R. E. Schapire and Y. Freund. "*Boosting: Foundations and algorithms*". *Kybernetes*, vol. 42, no. 1, pp. 164-166, 2013.

[23] B. Niu, J. Ren and X. Li. "Credit scoring using machine learning by combing social network information: Evidence from peer-to-peer lending". Information, vol. 10, p. 397, 2019.

[24] A. Mayr, H. Binder, O. Gefeller and M. Schmid. "The evolution of boosting algorithms. From machine learning to statistical modeling". *Methods of Information in Medicine*, vol. 53, no. 6, pp. 419-427, 2014.

[25] R. Sikora and O. H. Al-laymoun. "A modified stacking ensemble machine learning algorithm using genetic algorithms". *Journal of International Technology and Information Management*, vol. 23, p. 1, 2014.

ORIGINAL RESEARCH ARTICLE UHD JOURNAL OF SCIENCE AND TECHNOLOGY

Face Recognition Use Local Image Dataset and Correlation Technique

Dana Faiq Abd

Department of Information Technology, College of Science and Technology, University of Human Development, Sulaimani City, Iraq

ABSTRACT

Face recognition is an extreme topic in security field which identifies humans through physiological or behavioral biometric characteristics. Face recognition can also identify the human almost in a precise detection; one of the primary problems in face recognition is the accurate recognition rate. Local datasets use for implementing this research rather than using public datasets. Midian filter uses to remove noise and identify errors, also obtains a good accuracy rate without modifying image quality. In addition, filter processing applies to modify and progress images and the discrete wavelet transforms algorithm uses as feature extraction. Many steps are applied in this approach such as image acquisition, converting images into gray scale, cropping the image, and then passing to the feature extraction. In order to get the final decision about the indicated face, some required steps are used in the comparison. The results show the accuracy of 91% of the recognition rate through the human face.

Index Terms: Face Recognition, Biometric Detection, Filter Processing, Finger Geometry, and Median Filter

1. INTRODUCTION

For many years biometric is one of the significant ways to use it to identify people. However, automatic biometric recognition appears and exists in recent years [1], [2]. Biological characteristics existing in every people however individual person is unique from these characteristics [3]. In addition, biological attributes get from the human can't be borrowed, forget, steal and replicate. Face recognition is one of the top methods in biometric identification because of its flexibility and user-friendly [4], [5]. There are some factors that directly impact faces recognition such as image illumination, quality of the image, type of camera, lens of the camera, and distance from the device to the face. These elements make to achieve accurate authentication and verification. Furthermore, face recognition is the process to identified faces through computer systems and technologies [6], [7]. It is one of the most important security methods compares to traditional security [8]. In traditional security, users log in to the system using user names and passwords that can easily steal by hackers. Moreover, attackers fulfill different procedures to compromised and break traditional security [9], [10]. Besides, face recognition consists of some steps. The first step is a sensor; digital cameras are applied to capture images and collect data, and then transformed into digital format. In the next step, the signal processing algorithm is used to identifying, adjusting, improving and controlling data. Then, information should be stored inside a database, prepare for comparison and evaluation. Furthermore, processing and analyzing the images converting them to gray scale and cropped. Later, images separately send through the filtering process using a median filter. After that, an algorithm is applying to detect and compare retained templates. Finally, the automatic decision is performed based on the similarity and dissimilarity of collecting images. A special technique

Access this article online

DOI:10.21928/uhdjst.v5n2y2021.pp26-37 E-ISSN: 2521-4217 P-ISSN: 2521-4209

Corresponding author's e-mail: Dana Faiq Abd, Assistant Lecturer, Department of Information Technology, College of Science and Technology, University of Human Development, Sulaimani City, Iraq. E-mail: dana.abd@uhd.edu.iq

Receiving: 01-05-2021 Accepting: 28-07-2021 Publishing: 05-08-2021

is implemented such as Median filter and Filter processing to prevent photos from copying [11].

This research uses two effective techniques which are Median filter and Filter processing. Therefore, the main contribution of this research is to design a hybrid system by combining the discrete wavelet transforms (DWT), filter processing, and median filter these are leads to get a high accuracy recognition rate. The median filter is a method uses to erase, decrease noise, and identify faults on images or signals. In addition, it uses to enhance the quality of the pictures and to detecting duplicated noise. The median filter is a filtering method that focuses on smoothness, sharpness, and edge upgrading images or signals. On the other hand, Filter processing is a technique that is used to change and improve images. Besides, it can focus on any part of the images to remove or cropped unnecessary portions [12], [13]. Face recognition is the subject of many algorithms; the majority of them are solely concerned with the face recognition rate. The proposed method provides an effective face recognition algorithm that achieves a high recognition rate without affecting image quality. Implementing median filter and filter processing play a significant role to get a high accuracy recognition rate. Since the median filter is removing noise without affecting image quality.

Finally, the rest of the paper is organized as follows: Section 2 presents the literature review, section 3 discusses face recognition methods (proposed approach), and section 4 presents the results in the form of tables. Finally, section 5 discusses the conclusion.

2. LITERATURE REVIEW

There are numerous kinds of biometric innovations available such as face-acknowledgment, unique finger impression acknowledgment, finger geometry, hand geometry, iris acknowledgment, vein acknowledgment, voice, and mark. The face is the usually utilized biometric attribute for individual acknowledgment. Furthermore, the most prominent ways to deal with face acknowledgment depend on the state of facial characteristics. For example, eyes, eyebrows, nose, lips, jawline, and the connections of these qualities. The most well-established ID component is recognizing people by their facial highlights (or vectors), which may be traced back to our primate ancestors. The human capacity to perceive faces is likewise imperative to the security engineer due to the across-the-board dependence set on picture IDs.

Many papers and researches have been published which are associated with face recognition in this section. We will try to describe and discuss some of these related works and approaches:

Ahmed *et al.* (2019) implemented an approach to design efficient face recognition by blending many methods to generate efficient recognition results. This technique covers five stages. Image acquisition: for this stage, ORL picture Dataset has been applied to afford the pictures with diverse positions. Preprocessing: for this stage, several tasks have been done such as cleaning, resizing, and filtering. Then, singular value decomposition applied: for this stage, the actual picture is decomposed for three orthogonal matrices. Selecting the most effective rank: For this step, the measurement of dispersion is used to present the best efficient matrix rank is achieved. In addition, the DWT applied: in this stage, the diagonal matrix produces significant characteristics by applied the DWT. Later, measures of dispersion are used to imply the division of data. According to the applied ranks, there is an acceptable reasonable rank that is important to reach via the implemented procedure. Interquartile range, mean absolute deviation, range, variance, and standard deviation are used to pick the correct rank. Rank 24, 12, and 6 acquired an outstanding 100% identification rate with data decrease up to 2: 1, 4: 1, and 8: 1 correspondingly. The regular PC and MATLAB simulator tool version 2015 used to complete the task [14]. Belhumeur *et al.* (1997), focus on developing a face recognition algorithm. Even under the substantial change in lighting and facial appearances, their projection method focuses on developing a face recognition algorithm for facial expression and light direction which is obtuse to great deviation. Consider each pixel in an image as a coordinate in a high-dimensional space when using a pattern categorization method. There are some gained advantages of the fact that images of a particular face, under fluctuating light but fixed position, reside in a 3D linear subspace of the high-dimensional image space. Supposing the face is a Lambertian (diffusely reflecting) surface with no shadows. Images will diverge from this linear subspace since faces are not fully Lambertian surfaces and create self-shadowing. They linearly project the image into a subspace which is based on Fisher's linear discriminant. Produces well-separated classes in a low-dimensional subspace [15]. Chen *et al.* (2014) Face recognition has a large dictionary and a unique value. The intra-class variation dictionary is used in the study in a subsampling state to represent the difference between samples in both the learning and examination phases [16]. Dehghan *et al.* (2015) used a portion of the training data for data modeling and used a knowledge

score to define the result, which are features of confidence representation [17]. Zheng *et al.* (2016) developing a comprehensive divided representation dictionary for facial recognition that can effectively overcome occlusion issues [18]. Gao *et al.* (2017) offered a new strategy by generating a low range of the data representation and each fault image generated by the occlusion in real-time [19]. Wu and Ding (2018) resolved the issue of face recognition, a hierarchical model based on Adaptive Sparse and Low-Rank gradients was used to solve the problem of face recognition [20]. Jing *et al.* (2019) apply a different method to find the discriminant tensor demonstration to properly apply the photo taxonomy problem [21]. Wadkar *et al.* (2012) developed a facial recognition technique based on DWT. To understand a certain range wavelet, decompositions are used in the procedure in order to improve performance. Furthermore, as a part of the approach algorithm, Haar, 9/7 wavelet filters used in the research. The work engages a variety of distance measures, including Euclidean, L1, and others. The Euclidean distance is obtained popularity [22]. Dandpat *et al.* (2013) used both conventional principal component analysis (PCA) and two-dimensional PCA (2DPCA) to increase face recognition performance. Their method shows that eigenvectors with limited nonzero eigenvalues are related. Different weights should be given to non-trivial fundamental parts for each category. Finally, the k-nearest-distance method is applied to determine facial recognition [23]. The proposed method is based on the MATLAB simulation tool and applied DWT. Furthermore, low pass band and high pass band applied on both rows and columns, respectively, using a median filter. Moreover, remarkable biometric distinguishing obtains by applying multiple biometric techniques like face, voice, or Iris, face. These highlights have their particular favorable circumstances and burdens with respect to the applications and necessities [24]. One acknowledgment framework that functions admirably for a specific prerequisite may not be appropriate to other people. Among these component acknowledgment frameworks, face acknowledgment likewise finds significance from the previous year's spreading over various fields. Face acknowledgment is vital on the grounds that it finds a few down-to-earth applications. Also, it is a key human conduct for communicating feelings [25]. The utilization of face as a component for acknowledgment framework additionally has preferences in law implementation. As it doesn't require dynamic support from the individual as enormous information test assortment is reachable in a helpful manner [26]. In the end, median filter has some advantages and disadvantages. The advantages median filter is one of the most important filters because it works well with noise types including "Gaussian," "random," and "salt and pepper." Noise pixels are distinguished from the median. This type of noise problem can be eliminated by using this approach median filter. When use low pass filter in median filter only the noise reduces contrary all other type of filter reduces the image quality [27]. When the noise amplitude probability distribution contains big tails and periodic patterns, median filters are beneficial for decreasing random noise. In addition, when apply median filter there is no need to create a new pixel value and it is simple to put into practice. The downsides of such filters are that they tend to break up image edges and produce misleading noise edges when signal-to-noise ratios are low, and they cannot suppress medium-tailed (Gaussian) noise distributions. Finally, the median filter causes some delay in processing time [28], [29].

3. THE PROPOSED APPROACH

3.1. Local Image Dataset

For data collection and image acquisition, ten volunteers are prepared in a suitable area with a natural environment without external lighting factors. Besides, all volunteers are between 20 and 30 years old, take pictures using an ordinary camera for each volunteer. Four images of head captured that include all front hair, forehead, eyebrow, eyelash, eye, ears, nose, cheek, mouth, lip, and chin.

HUAWEI Y6II mobile phone camera, with resolution 4160 × 3120 and 13 mega pixels' technology used in the procedure with a distance of 2-m in between mobile phone camera and volunteer.

3.2. DWT

As shown in Fig. 1, the DWT distributes an image into four groups such as approximation band, vertical band, horizontal band, and diagonal band. The maximum number of levels equal to N × N image size. Both the row and columns of images traversing across the low path and high pass filter to create low and high-frequency approximation and vertical bands consequently. Besides, the column crossing through low pass and high pass filter with results of row high pass filter to make horizontal and diagonal bands. The LL band includes important information about an original image. The LH, HL, and HH bands have vertical, horizontal, and diagonal information of an original image. The original image recreated by deliberately only the LL band image and eliminating other unimportant information from other bands [30], [31].

A dataset with forty images obtained (four images for each volunteer), for preparation of processing on the face only, as shown in Fig. 2.

3.3. Implemented Approach

Image acquisition, converting images from color images to gray scale images, cropping image by a rectangle with 1000 × 1000 pixels for face images as demonstrate in Fig. 3. Images composed of eyes, nose, mouth, and cheeks are ongoing steps implemented in the proposed approach for face recognition. Each rectangle image separately passed via filtering process using median filter (which is a non-linear digital filtering method used in a stage of preprocessing to remove noise from an image). A median filter has very robust efficiency which concurrently can decrease noise and retain edges. The median filter makes evaluation and computation based on adjacent pixels which select the middle value of a dataset listed from smallest to the largest. It examines the entire original image from top to bottom, left to the right, and makes a particular image that is based on median values.

Fig. 4 is an example of a median filter which is by default it collects the adjacency pixels around the target pixel, and categorizing them from small value to high value, and picks the median value from the arranged list. As demonstrated in Fig. 4 from the output window, there are different values and from sorter out the values are sorting from smallest to the highest value, and then from output pixel, the median value is selected.

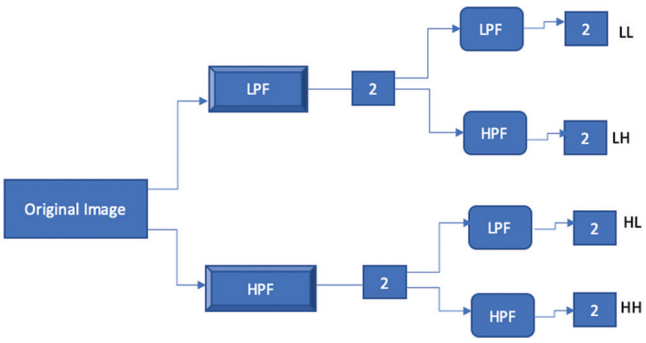

Fig. 1. Decomposition of discrete wavelet transforms.

According to the research procedure of data collection, there are 40 images collected for the face. There are 40 volunteers, for each volunteer four different images uses. The value of correlation between each pair of images recorded to each volunteer. Finally, the correlation value calculated for each face image as shown in Figs. 4 and 5. Because for each volunteer four images captured then perform processing and applying the medium filter for each image. The correlation value between images is calculated. As discussed in (3.4 Discussion).

Fig. 4 is illustrating an image face for a volunteer the first one is the image face before filtering and the second one is the image face after the filtering process.

4. DISCUSSION

Although comparisons have been done all images, respectively, the variance between both images is small or big depending on the details of the image, in this situation; the median filter plays an important role in the identification process. When the result value (RV) obtained for each pair of images, consequently added to a table to show the degree of similarity among all pictures taken for each volunteer as shown in Table 1.

- N^{th}: Number of volunteers (we have 10 volunteers).
- 1^{st} pic: Volunteer first picture.

Fig. 2. A face image dataset.

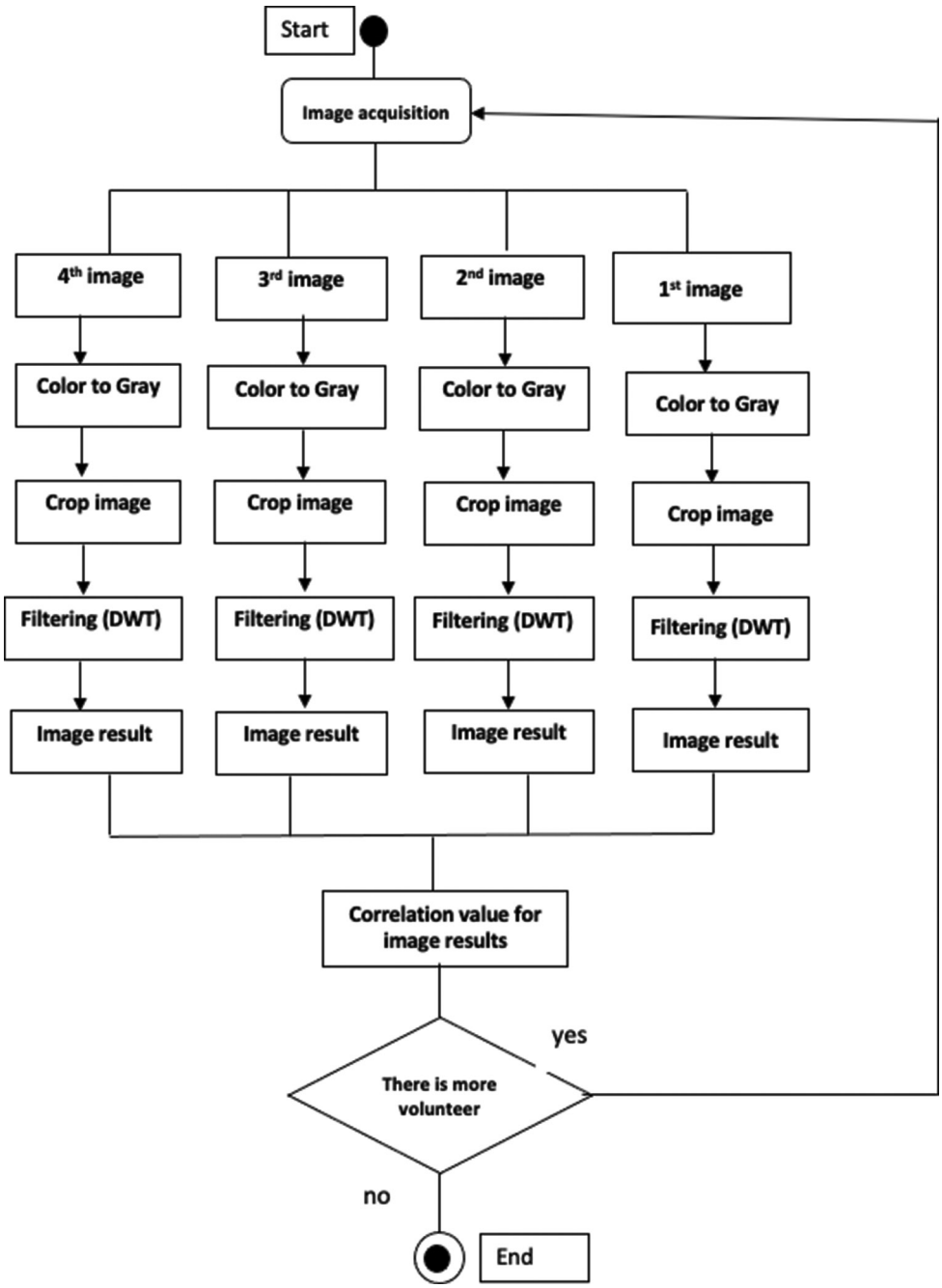

Fig. 3. Main steps of methodology.

TABLE 1: The result values of correlation

		1st pic	2nd pic	3rd pic	4th pic
Number of volunteer Face Images	1st pic	Result value 1	Result value 2	Result value 3	Result value 4
	2nd pic	Result value 1	Result value 2	Result value 3	Result value 4
	3rd pic	Result value 1	Result value 2	Result value 3	Result value 4
	4th pic	Result value 1	Result value 2	Result value 3	Result value 4

- 2nd pic: Volunteer second picture.
- 3rd pic: Volunteer third picture.
- 4th pic: Volunteer fourth picture.
- RV: Correlation value between each image in a row with its corresponding image in a column.

To be more familiar with understanding the results, there is a chart graph for each table of result …… to show similarity between images as shown in Fig. 7.

- Blue column represents Correlation RV between the volunteer First picture with other pictures in the first column of the table.
- Brown column represents Correlation RV between volunteer Second picture with other pictures in Second column of the table.
- Gray column represents Correlation RV between volunteer Third picture with other pictures in the third column of the table.
- Yellow column represents Correlation RV between volunteer Fourth picture with other pictures in Second column of the table.
- Each aggregated 4 column in the X-axis represents a unique row of the table.

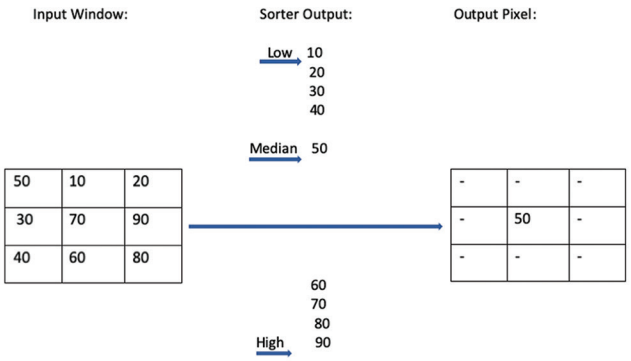

Fig. 4. The median filtering algorithm.

- The degrees located at Y-axis represent the value result.

5. RESULTS

According to the implementation of the proposed system, many results obtained according to the tested samples. These results are the correlation between different images, similarities, and differences to achieve the identification process. Considering the range of (1.0–0.5) enumerate as a positive correlation, which indicated the best identification result between different tested images. Then considering the range from 0.5 to the minimum value, this means a negative correlation that has no relation between the tested images. According to the obtained results from ten samples of face images, the result is 91% for all correlation RVs for face processed with a median filter. Good results of identification and differentiation in biometry process obtained.

The blue column represents the correlation RV between the volunteer's first pictures with other pictures in the first column of the table. The Brown column represents the correlation RV between the volunteer second pictures with other pictures in the second column of the table. The Gray column represents the correlation RV between the volunteer third picture with other pictures in the third column of the table. The yellow column represents the correlation RV between the volunteer fourth pictures with other pictures in the second column of the table. The black line represents a total average correlation. The degrees located at Y-axis represent the value result.

As shown from Table 2 and Fig. 8 below, Correlation average for 1st volunteer Face images shows RV of 0.9 for each image with its corresponding person images, which is

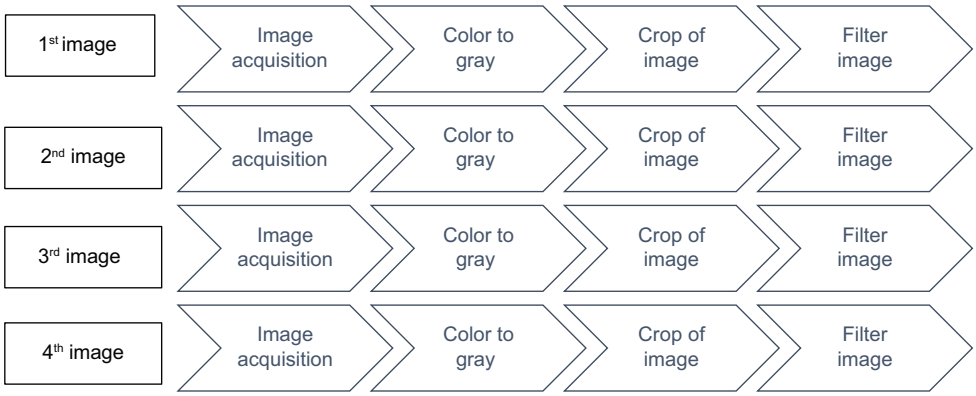

Fig. 5. Implemented approach.

a positive result according to the range mentioned above. Blue column represents correlation RV between volunteer first pictures with other pictures in first column of the table. Brown column represents correlation RV between volunteer second picture with other pictures in second column of the table. Gray column represents correlation RV between volunteer third picture with other pictures in third column of the table. Yellow column represents correlation RV between volunteer fourth picture with other pictures in second column of the table. The black line represents total of average correlation. The degrees located at Y-axis represent value result.

As shown from Table 3 and Fig. 9 below, Correlation average for 2nd volunteer Face images shows RV of 0.9 for each image with its corresponding person images, which is a positive result according to the range mentioned above. Blue column represents correlation RV between volunteer first picture with other pictures in first column of the table. Brown column represents correlation RV between volunteer second picture with other pictures in second column of the table. Gray column represents correlation RV between volunteer third picture with other pictures in third column of the table. Yellow column represents correlation RV between volunteer fourth picture with other pictures in second column of the table. The black line represents total of average correlation. The degrees located at Y-axis represent value result.

As shown from Table 4 and Fig. 10 below, Correlation average for 3rd volunteer Face images shows RV of 1.0 for each image with its corresponding person images, which is a positive result according to the range mentioned above. Blue column represents correlation RV between volunteer first picture with other pictures in first column of the table. Brown column represents correlation RV between

TABLE 2: Correlation value for 1st volunteer FACE images

		1st pic	2nd pic	3rd pic	4th pic
1st Volunteer FACE image	1st pic	1	0.8765	0.8574	0.8961
	2nd pic	0.8765	1	0.7032	0.9882
	3rd pic	0.8574	0.7032	1	0.7348
	4th pic	0.8961	0.9882	0.7348	1

TABLE 3: Correlation value for 2nd volunteer FACE images

		1st pic	2nd pic	3rd pic	4th pic
2nd Volunteer FACE image	1st pic	1	0.8862	0.8761	0.8633
	2nd pic	0.8862	1	0.9452	0.9359
	3rd pic	0.8761	0.9452	1	0.9911
	4th pic	0.8633	0.9359	0.9911	1

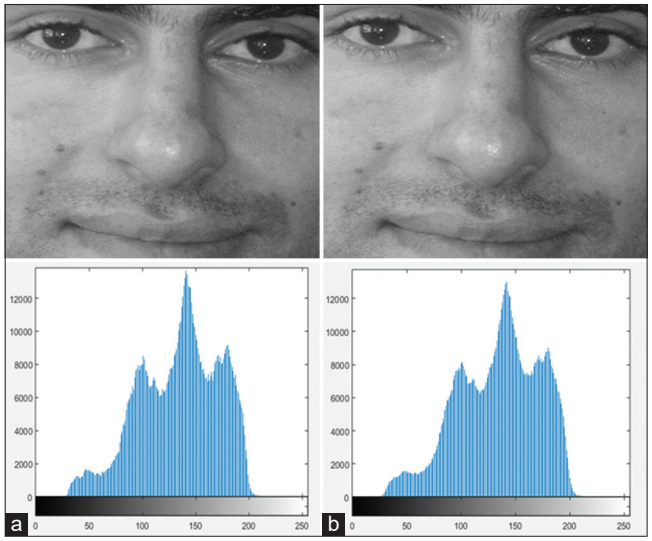

Fig. 6. (a and b) Faces image before and after filtering.

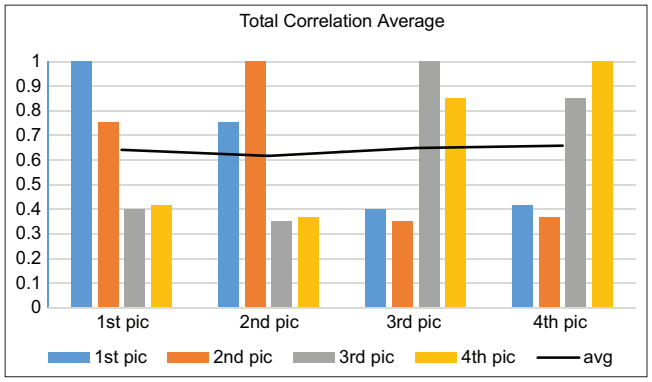

Fig. 7. The result average value of the correlation.

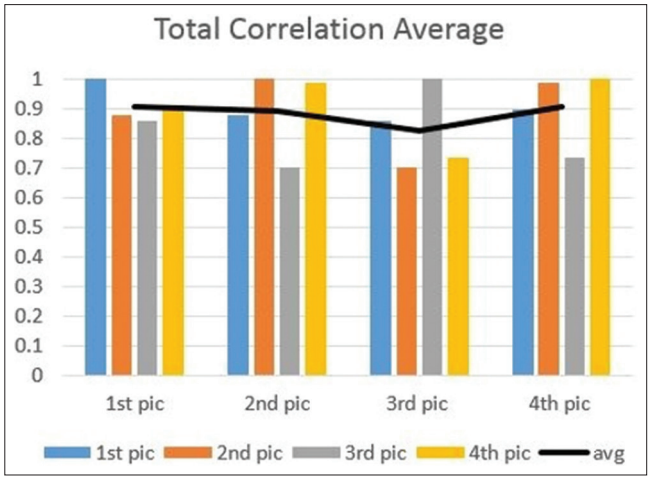

Fig. 8. Average correlation value for 1st volunteer FACE images.

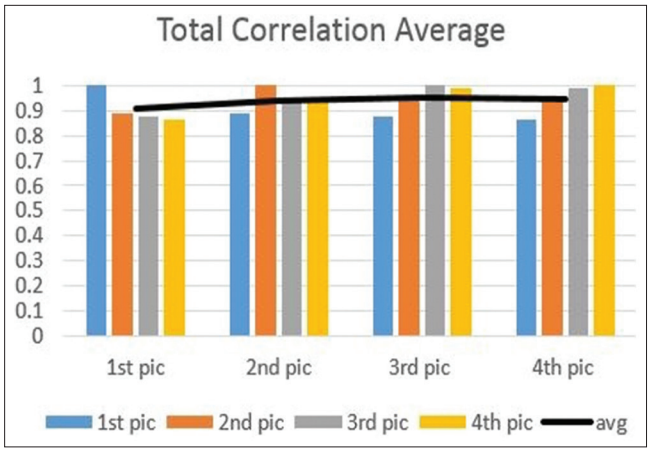

Fig. 9. Average Correlation value for 2nd volunteer FACE images.

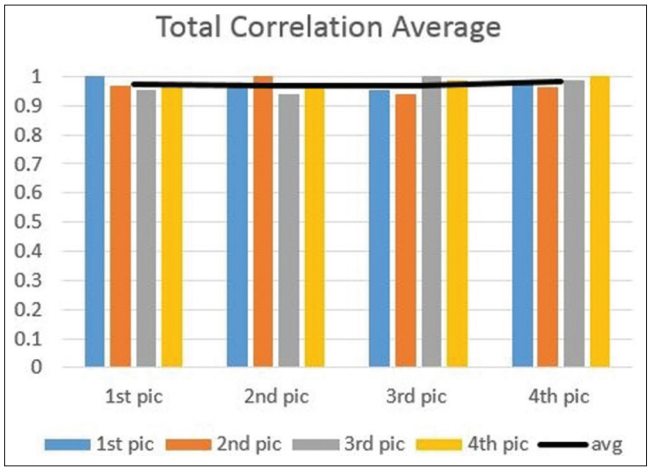

Fig. 10. Average correlation value for 3rd volunteer FACE images.

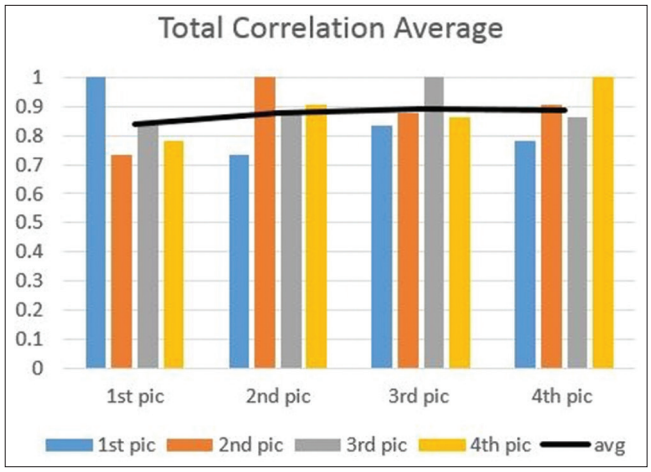

Fig. 11. Average correlation value for 4th volunteer FACE images.

TABLE 4: Correlation value for 3rd volunteer FACE images

		1st pic	2nd pic	3rd pic	4th pic
3rd Volunteer FACE image	1st pic	1	0.9668	0.9548	0.9763
	2nd pic	0.9668	1	0.938	0.9631
	3rd pic	0.9548	0.938	1	0.9892
	4th pic	0.9763	0.9631	0.9892	1

TABLE 5: Correlation value for 4th volunteer FACE images

		1st pic	2nd pic	3rd pic	4th pic
4th Volunteer FACE image	1st pic	1	0.7339	0.8339	0.7822
	2nd pic	0.7339	1	0.8757	0.9069
	3rd pic	0.8339	0.8757	1	0.8611
	4th pic	0.7822	0.9069	0.8611	1

volunteer second picture with other pictures in second column of the table. Gray column represents correlation RV between volunteer third picture with other pictures in third column of the table. Yellow column represents correlation RV between volunteer fourth picture with other pictures in second column of the table. The black line represents total of average correlation. The degrees located at Y-axis represent value result.

As shown from Table 5 and Fig. 11 below, Correlation average for 4th volunteer Face images shows RV of 0.9 for each image with its corresponding person images, which is a positive result according to the range mentioned above. Blue column represents correlation RV between volunteer first picture with other pictures in first column of the table. Brown column represents correlation RV between volunteer second picture with other pictures in second column of the table. Gray column represents correlation RV between volunteer third picture with other pictures in third column of the table. Yellow column represents correlation RV between volunteer fourth picture with other pictures in second column of the table. The black line represents total of average correlation. The degrees located at Y-axis represent value result.

As shown from Table 6 and Fig. 12 below, Correlation average for 5th volunteer Face images shows RV of 0.9 for each image with its corresponding person images, which is a positive result according to the range mentioned above. Blue column represents correlation RV between volunteer first picture with other pictures in first column of the table. Brown column represents correlation RV between volunteer second picture with other pictures in second column of the table. Gray column represents correlation RV between volunteer third picture with other pictures in third column of the table. Yellow column represents correlation RV between volunteer

fourth picture with other pictures in second column of the table. The black line represents total of average correlation. The degrees located at Y-axis represent value result.

As shown from Table 7 and Fig. 13 below, Correlation average for 6th volunteer Face images shows RV of 1.0 for each image with its corresponding person images, which is a positive result according to the range mentioned above. Blue column represents correlation RV between volunteer first picture with other pictures in first column of the table. Brown column represents correlation RV between volunteer second picture with other pictures in second column of the table. Gray column represents correlation RV between volunteer third picture with other pictures in third column of the table. Yellow column represents correlation RV between volunteer fourth picture with other pictures in second column of the table. The black line represents total of average correlation. The degrees located at Y-axis represent value result.

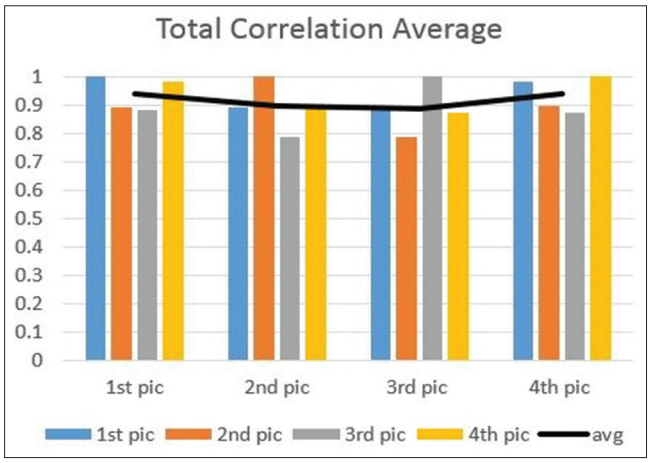

Fig. 12. Average correlation value for 5th volunteer FACE images.

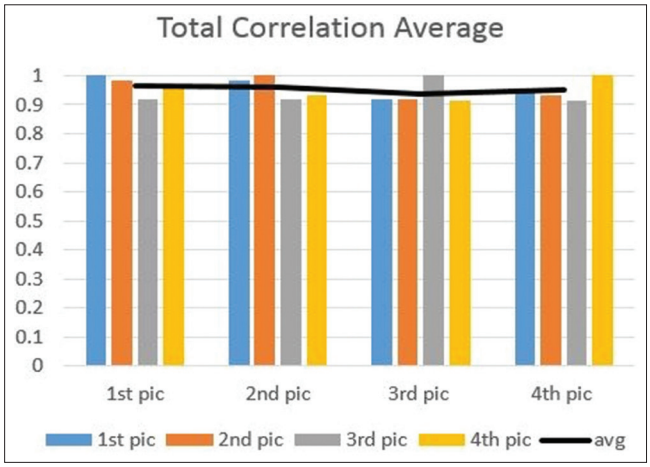

Fig. 13. Average correlation value for 6th volunteer FACE images.

As shown from Table 8 and Fig. 14 below, Correlation average for 7th volunteer Face images shows RV of 0.8 for each image with its corresponding person images, which is a positive result according to the range mentioned above. Blue column represents correlation RV between volunteer first picture with other pictures in first column of the table. Brown column represents correlation RV between volunteer second picture with other pictures in second column of the table. Gray column represents correlation RV between volunteer third picture with other pictures in third column of the table. Yellow column represents correlation RV between volunteer fourth picture with other pictures in second column of the table. The black line represents total of average correlation. The degrees located at Y-axis represent value result.

As shown from Table 9 and Fig. 15 below, Correlation average for 8th volunteer Face images shows RV of 0.9 for each image with its corresponding person images, which is a positive result according to the range mentioned above.

TABLE 6: Correlation value for 5th volunteer FACE images

		1st pic	2nd pic	3rd pic	4th pic
5th Volunteer FACE image	1st pic	1	0.8907	0.8833	0.9845
	2nd pic	0.8907	1	0.789	0.8994
	3rd pic	0.8833	0.789	1	0.8723
	4th pic	0.9845	0.8994	0.8723	1

TABLE 7: Correlation value for 6th volunteer FACE images

		1st pic	2nd pic	3rd pic	4th pic
6th Volunteer FACE image	1st pic	1	0.9811	0.918	0.9543
	2nd pic	0.9811	1	0.9197	0.9309
	3rd pic	0.918	0.9197	1	0.9119
	4th pic	0.9543	0.9309	0.9119	1

TABLE 8: Correlation value for 7th volunteer FACE images

		1st pic	2nd pic	3rd pic	4th pic
7th Volunteer FACE image	1st pic	1	0.867	0.9224	0.6389
	2nd pic	0.9224	1	0.7905	0.5321
	3rd pic	0.9224	0.7905	1	0.7209
	4th pic	0.6389	0.5321	0.7209	1

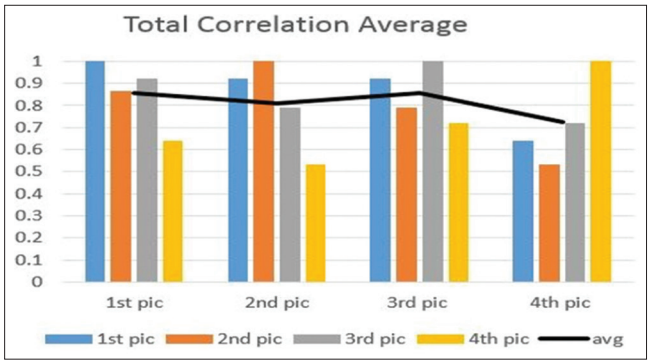

Fig. 14. Average correlation value for 7th volunteer FACE images.

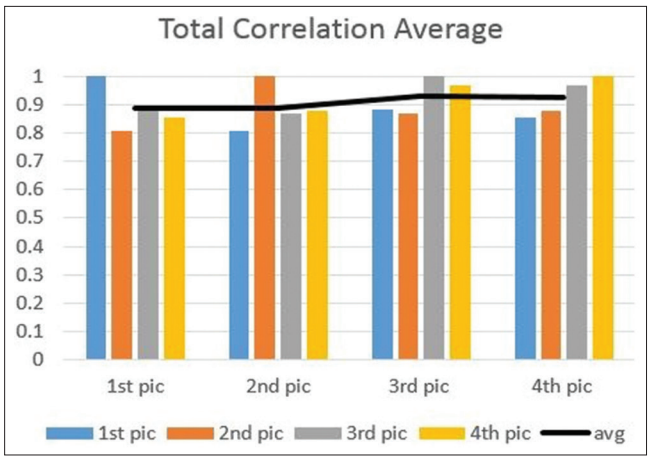

Fig. 15. Average correlation value for 8th volunteer FACE images.

Blue column represents correlation RV between volunteer first picture with other pictures in first column of the table. Brown column represents correlation RV between volunteer second picture with other pictures in second column of the table. Gray column represents correlation RV between volunteer third picture with other pictures in third column of the table. Yellow column represents correlation RV between volunteer fourth picture with other pictures in second column of the table. The black line represents total of average correlation. The degrees located at Y-axis represent value result.

As shown from Table 10 and Fig. 16 below, Correlation average for 9th volunteer Face images shows RV of 0.9 for each image with its corresponding person images, which is a positive result according to the range mentioned above. Blue column represents correlation RV between volunteer first picture with other pictures in first column of the table. Brown column represents correlation RV between volunteer second picture with other pictures in second column of the table. Gray column represents correlation RV

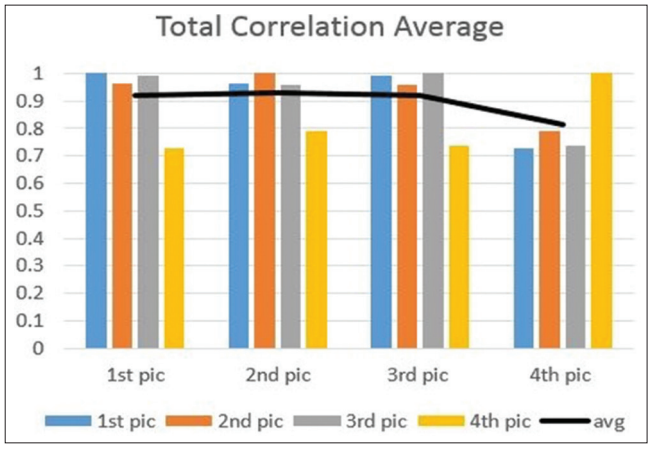

Fig. 16. Average correlation value for 9th volunteer FACE images.

TABLE 9: Correlation value for 8th volunteer FACE images

		1st pic	2nd pic	3rd pic	4th pic
8th Volunteer FACE image	1st pic	1	0.8087	0.8857	0.8572
	2nd pic	0.8087	1	0.869	0.8778
	3rd pic	0.8857	0.869	1	0.9703
	4th pic	0.8572	0.8778	0.9703	1

TABLE 10: Correlation value for 9th volunteer FACE images

		1st pic	2nd pic	3rd pic	4th pic
9th Volunteer FACE image	1st pic	1	0.9634	0.991	0.7296
	2nd pic	0.9634	1	0.9576	0.7892
	3rd pic	0.991	0.9576	1	0.7373
	4th pic	0.7296	0.7892	0.7373	1

between volunteer third picture with other pictures in third column of the table. Yellow column represents correlation RV between volunteer fourth picture with other pictures in second column of the table. The black line represents total of average correlation. The degrees located at Y-axis represent value result.

As shown from Table 11 and Fig. 17 below, Correlation average for 10th volunteer Face images shows RV of 0.9 for each image with its corresponding person images, which is a positive result according to the range mentioned above. Blue column represents correlation RV between volunteer first picture with other pictures in first column of the table. Brown column represents correlation RV between volunteer second picture with other pictures in second column of the table. Gray column represents correlation RV between volunteer third picture with other pictures in third column of the table. Yellow column represents correlation RV between volunteer

fourth picture with other pictures in second column of the table. The black line represents total of average correlation. The degrees located at Y-axis represent value result.

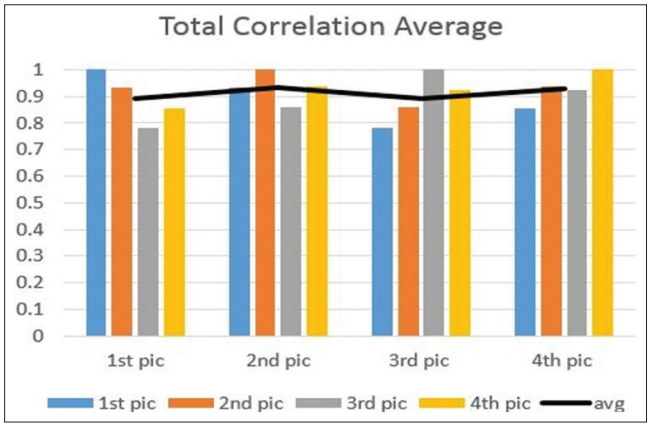

Fig. 17. Average CORRELATION value for 10th volunteer FACE images.

TABLE 11 :Correlation value for 10th volunteer FACE images

		1st pic	2nd pic	3rd pic	4th pic
10th Volunteer FACE image	1st pic	1	0.9344	0.7794	0.8566
	2nd pic	0.9344	1	0.8583	0.9354
	3rd pic	0.7794	0.8583	1	0.9224
	4th pic	0.8566	0.9354	0.9224	1

TABLE 12: Correlation value result for all volunteers

Volunteers	v1	v2	v3	v4	v5	v6	v7	v8	v9	v10
Correlation Result	0.9	0.9	1	0.9	0.9	1	0.8	0.9	0.9	0.9

TABLE 13: Comparison some related works with our proposed algorithm

No	Authors	Techniques	% Accuracy
1	Swarup Kumar Dandpat and SukadevMeher [28]	PCA and 2DPCA	90.5
2	Pallavi D. Wadakar and MeghaWankhade [27]	DWT	90
3	Chun-Hou Zheng, Yi-Fu Hou, and Jun Zhang [18]	ERSC_LR algorithm SVM (linear) NN FDDL SVM (RBF)	97.97 94.37 90.28 97.01 97.72
4	Omed Hassan Ahmed, Joan Lu, Qiang Xu, and Muzhir Shaban Al-Ani [25]	SVD + DWT	100
5	Proposed Method	DWT and Median Filter	91

Table 12 explains total accuracy rate for all 10 volunteers. For the (volunteer 1, volunteer 2, volunteer 4, volunteer 8, volunteer 9, and volunteer 10). the accuracy is 90%. In addition, the accuracy is 100% for volunteer 3, volunteer 6, and for volunteer 7 is 80%).

Total Accuracy rate = total result with error/10 = 9.1/10 = (0.91*100) = 91%

Table 13 demonstrates the comparison works from other researchers and our proposed system. Kumar *et al.* apply (PCA and 2DPCA) techniques the accuracy is 90.5% [28]. Wadakar *et al.* apply DWT technique the accuracy is 90% [27]. Zheng *et al.* use (ERSC_LR algorithm, SVM (linear), NN, FDDL, SVM (RBF)) techniques the accuracy is 97.97, 94.93, 90.28, 90.01, and 97.72%, respectively [18]. Ahmed. *et al.* apply (SVD and DWT) the accuracy is 1000% [25]. Finally, (DWT and Median Filter) techniques apply in our proposed system the accuracy is 91%.

6. CONCLUSION

Biometry is an effective method for identification and differentiation among humans, it has done either on physiological properties such as iris, face, finger…. or on behavior properties such as voice, signature…. The obtained identification results for face biometrics are 91% of accuracy (based on Table 12 average value) by applying DWT and Median filter approach. To compare our accuracy result with other mentioned results a good accuracy score has been achieved. Moreover, all images are captured within the same environment and specification without additional change all images have the same angle and background.

REFERENCES

[1] R. Malathi and R. R. R. Jeberson. "An integrated approach of physical biometric authentication system" *Procedia Computer Science*, vol. 85, pp. 820-826, 2016.

[2] F. Battaglia, G. Iannizzotto and L. Lo Bello. "A person authentication system based on RFID tags and a cascade of face recognition algorithms". *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 27, no. 8, pp. 1676-1690, 2017.

[3] A. James, I. Fedorova, T. Ibrayev and D. Kudithipudi. "HTM spatial pooler with memristor crossbar circuits for sparse biometric recognition". *IEEE Transactions on Biomedical Circuits and Systems*, vol. 11, no. 3, pp. 640-651, 2017.

[4] L. Best-Rowden and A. Jain. "Longitudinal study of automatic face recognition". *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 40, no. 1, pp. 148-162, 2018.

[5] L. Zhang, P. Dou and I. Kakadiaris. "Patch-based face recognition using a hierarchical multi-label matcher". *Image and Vision*

Computing, vol. 73, pp. 28-39, 2018.

[6] A. Punnappurath, A. Rajagopalan, S. Taheri, R. Chellappa and G. Seetharaman. "Face recognition across non-uniform motion blur, illumination, and pose". *IEEE Transactions on Image Processing*, vol. 24, no. 7, pp. 2067-2082, 2015.

[7] Z. A. Kakarash, D. F. Abd, M. Al-Ani, G. Abubakr and K. M. Omar. "Biometric Iris recognition approach based on filtering techniques". *Journal of the University of Garmian*, vol. 6, no. 2, p. 34243, 2019.

[8] M. Sharif, S. Bhagavatula, L. Bauer and M. Reiter. "*Accessorize to a Crime*". Proceedings of the 2016 ACM SIGSAC Conference on Computer and Communications Security-CCS'16, United States, 2016.

[9] A. Aboshosha, K. A. El Dahshan, E. A. Karam and E. A. Ebeid. "Score level fusion for fingerprint, Iris and face biometrics". *International Journal of Computer Applications*, vol. 111, no. 4, pp. 47-55, 2015.

[10] Y. Xu, Z. Zhang, G. Lu and J. Yang. "Approximately symmetrical face images for image preprocessing in face recognition and sparse representation based classification". *Pattern Recognition*, vol. 54, pp. 68-82, 2016.

[11] S. Karahan, M. Kilinc Yildirum, K. Kirtac, F. Rende, G. Butun and H. *Ekenel. How Image Degradations Affect Deep CNN-based Face Recognition*?" 2016 International Conference of the Biometrics Special Interest Group, Darmstadt, Germany, 2016.

[12] R. V. Chawngsangpuii and K. S. Yumnam. "Different approaches to face recognition". *International Journal of Engineering Research and Technology*, vol. 4, no. 9, pp. 71-75, 2015.

[13] E. Hjelmås. "*Biometric Systems: A Face Recognition Approach*". Department of Informatics University of Oslo, Oslo, Norway.

[14] O. H. Ahmed, J. Lu, Q. Xu and M. S. Al-Ani. "Face recognition based rank reduction SVD approach". *The ISC International Journal of Information Security*, vol. 11, no. 3, p. 6, 2019.

[15] P. N. Belhumeur, J. P. Hespanha and D. J. Kriegman. "Eigenfaces vs. Fisherfaces: Recognition using class specific linear projection". *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 19, p.7, 1997.

[16] S. B. Chen, C. H. Q. Ding and B. Luo. "Extended linear regression for undersampled face recognition". *Journal of Visual Communication and Image Representation*, vol. 25, no. 7, pp. 1800-1809, 2014.

[17] A. Dehghan, O. Oreifej and M. Shah. "Complex event recognition using con-strained low-rank representation". *Image and Vision Computing*, vol. 42, pp. 13-21, 2015.

[18] C. H. Zheng, Y. F. Hou and J. Zhang. "Improved sparse representation with low-rank representation for robust face recognition". *Neurocomputing*, vol. 198, pp. 114-124, 2016.

[19] G. Gao, J. Yang, X. Y. Jing, F. Shen, W. Yang and D. Yue. "Learning robust and discriminative low-rank rep-resentations for face recognition with occlusion". *Pattern Recognition*, vol. 66, pp. 129-143, 2017.

[20] C. Y. Wu and J. J. Ding. "Occluded face recognition using low-rank regression with gen- eralized gradient direction". *Pattern Recognition*, vol. 80, pp. 256-268, 2018.

[21] P. Jing, Y. Su, Z. Li, J. Liu and L. Nie. "Low-rank regularized tensor discriminant representation for image set classification". *Signal Processing*, vol. 156, pp. 62-70, 2019.

[22] P. D. Wadkar and M. Wankhade. "Face recognition using discrete wavelet transforms". *International Journal of Advanced Engineering Technology*, vol. 3, pp. 239-242, 2012.

[23] S. K. Dandapat and S. Meher. "Performance improvement for face recognition using PCA and two-dimensional PCA". *IEEE International Conference on Computer Communication and Informatics*, vol. 2013, pp. 1-5, 2013.

[24] R. Chellappa, C. L. Wilson, and S. Sirohey. "Human and machine recognition of faces: A survey". *Proceedings of the IEEE*, vol. 83, no. 5, pp. 405-741, 1995.

[25] A. S. Tolba, A. H. El-Baz and A. A. El-Harby. "Face recognition: A literature review". *World Academy of Science*, *Engineering and Technology*, vol. 2, pp. 7-21, 2008.

[26] Z. A. Kakarash, S. H. T. Karim and Mohammadi, M. "Fall detection using neural network based on internet of things streaming data". *UHD Journal of Science and Technology*, vol. 4, no. 2, pp. 91-98, 2020.

[27] D. Leonidas. "*Emerging Trends in Image Processing*, *Computer Vision and Pattern Recognition*. Elsevier, Huntsville, AL, USA, pp. 183-199, 2015.

[28] L. Cornelius. "*Multidimensional Systems: Signal Processing and Modeling Techniques*". Elsevier, Los Angeles, USA, 1995.

[29] I. Masi, F. Chang, J. Choi, S. Harel, J. Kim, K. Kim, J. Leksut, S. Rawls, Y. Wu, T. Hassner, W. AbdAlmageed, G. Medioni, L. Morency, P. Natarajan and R. Nevatia. "Learning pose-aware models for pose-invariant face recognition in the wild". *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 41, no. 2, pp. 1-1, 2018.

[30] G. V. Sagar, S. Y. Barker, K. B. Raja, K. S. Babu and K. R. Venugopal. "*Convolution Based Face Recognition Using DWT and Feature Vector Compression*". Institute of Electrical and Electronics Engineers Conference Paper, United States, 2015.

[31] C. Kyong, K. W. Bowyer and S. Sarkar. "Comparison and combination of ear and face images in appearance-based biometrics". *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 25, no. 9, pp. 1160-1165, 2013.

ORIGINAL RESEARCH ARTICLE UHD JOURNAL OF SCIENCE AND TECHNOLOGY

A Review of Database Security Concepts, Risks, and Problems

Ramyar Abdulrahman Teimoor*

Department of Computer, College of Science, University of Sulaimani, Sulaymaniyah, Iraq

ABSTRACT

Currently, data production is as quick as possible; however, databases are collections of well-organized data that can be accessed, maintained, and updated quickly. Database systems are critical to your company because they convey data about sales transactions, product inventories, customer profiles, and marketing activities. To accomplish data manipulation and maintenance activities the Database Management System considered. Databases differ because their conclusions based on countless rules about what an invulnerable database constitutes. As a result, database protection seekers encounter difficulties in terms of a fantastic figure selection to maintain their database security. The main goal of this study is to identify the risk and how we can secure databases, encrypt sensitive data, modify system databases, and update database systems, as well as to evaluate some of the methods to handle these problems in security databases. However, because information plays such an important role in any organization, understanding the security risk and preventing it from occurring in any database system require a high level of knowledge. As a result, through this paper, all necessary information for any organization has been explained; in addition, also a new technological tool that plays an essential role in database security was discussed.

Index Terms: Database security, Attack, Threats, Protection, Encryption, Database vulnerability

1. INTRODUCTION

Databases and database systems are an indispensable part of contemporary life; most of us engage in at least one database-related activity each day [1]. Simply, everyone can save data and information into a database in order to keep business apparatuses safe and protected. In the case of an emergency, technology has dramatically increased our odds of survival. In reality, technology has enhanced how we live, travel, communicate, study, and be treated medically, as well as how we conduct our lives. Technology employed in essential infrastructure that sustains our daily lives is becoming a necessity, and life would be unthinkable without it [2].

Today's databases and whole systems are often subjected to a variety of security risks. Many of these risks are prevalent in small businesses, but in large businesses and institutions, vulnerability is critical since they contain sensitive information that is utilized by many individuals and departments [3].

It is concerned with protecting databases against some form of unwanted access or danger at any stage. Server protection entails allowing or disallowing user behavior on the database and its properties. The security of their database has been sought by well-functioning organizations. They do not let the unlicensed user admittance their files or documents. They also state that their information is safe from any deceptive or unintended variations. The security priority is on data protection and privacy [4].

Access this article online

DOI: 10.21928/uhdjst.v5n2y2021.pp38-46

E-ISSN: 2521-4217
P-ISSN: 2521-4209

Corresponding author's e-mail: Ramyar Abdulrahman Teimoor, Department of Computer, College of Science, University of Sulaimani, Sulaymaniyah, Iraq. ramyar.teimoor@univsul.edu.iq

Received: 17-07-2021 Accepted: 30-09-2021 Published: 10-10-2021

This study concentrates on the database security risks that the database forensics can mitigate, as it is becoming an increasingly important topic for investigation. The study aims at assisting organization to protect data by introducing the highest nine vulnerabilities found in database. Context information, being Up-to-date risk information for each threat, protection data and information user database safety gateway protections, and some other method has also been investigated [5].

What is every organization next issue, are data using database protected? Nowadays, security is one of the most critical and difficult tasks people encounter. It is difficult to maintain databases. Practitioners of database protection do not comprehend the assaults as well as associated to database protection issues. companies are unaware of the sensitive data contained within databases, tables, and columns, as per IT experts and Database Administrator (Admin), since either they are managing inherited presentations or taking no records or maintain the data model documentations. If you know the database properties, the databases are more difficult to be secured due to their specific implementation and procedures. We can describe the database protection as the tool for implementing a wide scale controlling data security, protecting databases internally and externally, as well as compromising database privacy, truthfulness, and accessibility, such as technological, managerial, and bodily controls, are used to ensure security [6].

The following is a breakdown of how this study is structured. In Section 2 a further overview of related works is presented. Section 3 describes type of attack in any system. In section 4, threat and prevention that may be used against any database system has been explained. In section 5 describe some methods for protecting information in the database, as well as several brand of new technologies that have a positive impact on database security, have been introduced. Finally, the study's conclusion has been described in section 6.

2. LITERATURE REVIEW

In this section, significant amount of work presented. It enabled us to check and use the following sources accordingly.

2.1. Thilina [7]

This review has focused on the utilization of virtual resources in storing data for database users. It also includes information on the strategies used to address database security problems, as well as database attack and privacy risk mitigation measures. Organizations may store data and information in databases using an innovative business model that requires no initial investment. It also includes information on database security needs and assets. This review paper covers database security breaching risks, malware activities on such data, and how to address or mitigate those issues, as well as Oracle security database implementation.

2.2. Sharma [8]

This article discusses the security of relational database protection and security frameworks as an example of how internet application security for explicit database authorization may be designed and implemented. Because Relational Database securities are the most popular target for attackers, protection associations and substance are regarded as significant company resources that must be meticulously protected. This research was conducted to identify the problems and risks associated with relational database security, as well as the requirements for relational database set security and how Database Relations are used at different levels to provide security.

2.3. Albalawi [9]

They propose an intelligent system for hiding sensitive data when statistical searches are combined. To begin with, the framework is helpful for defining sensitive information in order for the Admin to make decisions and establish regulations. Second, in the event of rule discrimination based on attribute-orientation, the framework investigates the connection between sensitive and other characteristics, allowing for the selection of attributes that may be used to drive private data.

2.4. Juma and Makupi [10]

In their view, databases are the heart of Information Systems (IS), therefore it is critical to maintain database quality to ensure IS quality. Recently, determining what constitutes a good database model or architecture has proven to be difficult. As a result, they measured certain characteristics and aspects in a database implementation in their discussion. A measure of evaluation is created using the many elements and qualities inherent in a database.

2.5. Odirichukwu and Asagba [11]

They believe that the number of businesses putting their data online is growing every day, enabling people to engage with and manipulate data all around the world. More information on the internet.

As the number of websites on the internet grows, so does the number of database security risks.

Ensuring security in developed applications, owing to a combination of factors a lack of security incentives, a tight timeline, and a web deficit Training on application security testing a review of the literature is presented in this article on twenty database security risks that affect web applications. Control actions that might be taken to prevent this attacks were investigated to raise awareness about online security the general population, as well as application developers. The work expresses an opinion that developers should make every effort to incorporate all of the required features while developing apps, take security precautions. Involvement is also important. All developers should get security testing training. The task at hand despite ensuring sufficient security, the author finds that Admin should develop a method of maintaining a continuous backup of their database apps available online.

2.6. Paul and Aithal [3]

This article discusses the fundamentals of databases, such as their meaning, features, and roles, with an emphasis on various database security issues. Furthermore, this article emphasizes the fundamentals of security management, as well as relevant technologies. As a result, various aspects of database security have been briefly discussed in this article.

2.7. Mousa *et al*. [12]

According to the authors of this study, assaulters would rather attack the database because of the data sensitivity and value.

Databases compromised in many different ways. The database should be secured against different forms of attacks and threats. Most of the assaults identified in this study can be solved. Some of the assaults are real and some are not. In this article, they discuss various types of assaults.

2.8. Singh and Rai [13]

They concluded that databases are the foundation of modern applications. For businesses, they are the primary storage option. As a result, database attacks are on the rise, but crucially threatening. They give the intruder (InT) access to sensitive information. This study discusses a variety of database assaults. This study also includes a review of relevant database security strategies as well as potential study in the field of database protection. This study will result in a more concrete approach to the database security issue.

2.9. Tabrizchi and Rafsanjani [14]

The goal of this project is to examine the many components of cloud computing as well as the current security and privacy issues that these systems confront. Furthermore, this work introduces a new classification system for recent security solutions in this field. This study also addressed outstanding problems and suggested future approaches, as well as introducing different kinds of security risks that are affecting cloud computing services. This article will concentrate on and investigate the security issues that cloud organizations, such as cloud service providers, data owners, and cloud users, confront.

3. TYPE OF ATTACK

In a database, there are several protection layers. An InT will compromise protection at all of these levels, which include the database Admin, server Admin, security officer, developers, and employees [5].

Three types of attackers can be found [15]:

A. Intruder (InT)
   InT is an unwanted user who attempts to obtain useful information from a computer device by manipulating it excessively.
B. Insider (InS)
   InS is one of the members of trustworthy users who violates his or her permission and attempts to obtain knowledge outside his or her own allocation.
C. Administrator (Admin)
   Admin is a user with authority to operate a computer system who, in violation of the organization's security policies, abuses his or her management rights by spying on database management systems (DBMS) activities and obtaining sensitive data.

When an attacker breaks into the system, the two of the following attacked can be conducted [1]:

3.1.1. Direct Attacks

It refers to targeting the goal data first. These attacks are only possible and effective if the database has no security mechanism in place. If this attack is unsuccessful, the InT will move on to the next.

3.1.2. Indirect Attacks

It does not explicitly attack the goal, nonetheless data from or around the goal can be obtained by other in-between items, as the name suggests. Many of the variations of various questions are used and try to get through the authentication mechanism. It is difficult to keep track of these threats.

In general, database attacks are composed of two types [5] which are:

3.2.1. Passive Attack

In this case, the InT just inspects the data in the database and makes no changes. The following are few examples of passive attacks:

1. Static leakage: This attack obtains data about database plaintext content by analyzing a database snap taken at a given period.
2. Outflow of information: In this case, data about plaintext values can be accessed by connecting database values to the index location of mentioned values.
3. Dynamic leakage: Modifications made to a database over time may be detected and evaluated, as well as facts about plain text values.

3.2.2. Active Attack: Real database values are changed during an aggressive attack. These are more dangerous than passive attacks because they can lead to consumer confusion. For instance, a user can incorrectly capture information as a result of a query [5]. There are many methods for carrying out such an attack, which are mentioned below:

1. Spoofing – In this attack, a produced value is substituted for the cipher text value.
2. Splicing – This involves replacing a cipher text value with a new cipher text value.
3. Replay – This is an attack in which the cipher text value is replaced with an older version that has been changed or removed previously.

Because of the data they carry and their size, databases are the most popular target for cybercriminals [1]. A variety of database security risks and issues are addressed in this article.

4. THREAT AND PREVENTION

In this part, we'll go through nine of the most dangerous threats that may be utilized against databases, as well as how to avoid them form happening.

4.1. First Threat- Excessive Privilege Abuse

As soon as database access privileges given to users that go beyond what is required by their procedure, those privileges can be exploited for malicious purposes. A university Admin with the ability to alter student contact details can also use unnecessary database updating privileges to change grades, which is built-in.

Since Admin cannot identify and replace granularly, databases get admission to privilege management processes for each user, and a user eventually ends up with unnecessary privileges. Consequently, user(s) are given ordinary nonpayment access to privileges that go beyond the requirements of certain tasks [16].

Prevented by: Query-Level Access Control – Excessive Privilege Abuse Prevention

Accessing the question-level for regulation can be the solution of disproportionate rights. A process known as question-degree gets admission to control limits database rights to the bare minimum of SQL operations (select, update, and so on) and facts. The granularity of accessible data manipulations should develop outside the table to include the table rows and columns. A granular query-stage mechanism of accessible control could permit the previously mentioned college Admin updating the contact records while raising certain alarm if trying to change grades. Accessible control of query-stage can be valuable for detecting malicious workers who misuse their privileges but also for detecting unnecessary privilege abuse, in general, as well as for preventing the maximum assaults identified.

Implementation of the database software applications include a certain level of question-diploma management (triggers, row-stage protection, so on so forth), but the directed design of those "built-in" features make them unreasonable for anything but the built-in integrated deployments, the process of manual determination of the question-level access control policy for all database customers. The rows, columns, and operations take much time, to make matters worse, as user functions change over time, we should update query policies to represent the changes! Maximum database Admin will struggle to define a useful question policy for a few customers at a single time, let alone a smaller group of users over time. Consequently, most organizations provide unlimited rights of access to users with special collection of the paintings for a wider range of users. Automated gears are necessary for real-time question-degree access management to become a reality [16].

4.2. Second Threat - Authentic Privilege Abuse

Users to perform unauthorized tasks can use valid database privileges. Consider a hypothetical villain healthcare worker who has access to patient details through an application as the custom web. Internet application architecture usually limits users from accessing the medical history of a single patient. It is not possible to display several facts at the same time; meanwhile, electronic copies are not permitted. The villain worker, on the other hand, can get around those obstacles using a different client, such as MS-Excel connecting the database. The worker can also retrieve and buy all patient records using MS-Excel and the correct login credentials.

Personal copies of medical record files are unlikely to adhere to any healthcare record security policies of organization's patient. We have to be aware of two risks.

1) The villain employee swap personal information for make cash.
2) A careless employee retrieves and saves significant quantities of data to their client computer for authentic work purposes. Once information stored on an endpoint device, it would expose for Trojan virus, PC theft, and other assaults.

Prevented by: Legitimate Privilege Abuse Prevention

Database access management is the solution for legitimate privilege misuse that applies to queries but also to the context nearby database access. One can probably identify users abusing legitimate database access rights by enforcing a policy of patron packages, place, and time.

4.3. Third Threat - Privilege Elevation

Attackers of database platform software to adjust a regular user access privileges to those of an Admin can also use vulnerabilities. Stored procedures integration, built-in capabilities, implementations of protocols, and square statements may all be vulnerable. For instance, a financial institution software developer may consider taking advantage of a prone feature acquiring database administrative privilege.

The villain developer can disable audit mechanisms; build fictitious versions, transfer funds, and more administrative privileges [13].

Prevented By: Privilege Elevation Preventive – Institution Prevention Systems (IPS) and Query Level Access Control (QLAC)

Combination usage of traditional IPS and QLAC to manipulate, privilege elevate exploitation can be avoided (see excessive privileges above). IPS examines database site users for patterns that may lead to identified weaknesses. Once a characteristic identified as a prone, for instance, IPS most probably blocks the entire access to the prone method or, if likely, blocks the most successful processes with embedded attacks.

4.4. Fourth Threat - Platform Vulnerabilities

Unauthorized entry, data corruption, or service denial can result from flaws in underlying working frameworks (Windows 2000, UNIX, and so on so forth) and extra services installed on a database server. For instance, the blaster computer virus exploited a weakness in Windows 2000 causing a situation of Denial of Service (DoS) [13]. With the advancement of technology, security has improved as well, and as a result, many vulnerabilities have been solved in later versions of Windows or other platforms.

Prevented By: Avoiding Assaults, Updating the Software, and Preventing InTs

Protecting database property requires a mixture of protection programs and the IPS network security. Over-time, provisions of updates by the seller mitigated vulnerabilities discovered in the database platform. Unfortunately, businesses provide and enforce software upgrades regularly. Databases are not covered during the replacement periods. Furthermore, compatibility problems can often prevent software upgrades from happening. IPS must be introduced to address these problems. As mentioned before, IPS examines database visitors and detects assaults aiming recognized defenselessness.

4.5. Fifth Threat - SQL Injection

In certain cases, SQL injection assault, the perpetrator introduction (or "injection") unauthorized database reports into an inclined SQL channel. Saved approaches and web utility enter parameters are typical instances of oriented record channels. Then these injected reports sent to the database, where they are completed. Using SQL injections permits attackers may acquire unlimited access to all the database [13].

Prevented By: SQL Injection Prevention

IPS query-level gets proper access to governing (see disproportionate Privilege Abuse), and event correlation is three strategies that can be mixed to effectively fight rectangular SQL injection such as:

1. Input validation and Parametrized queries
2. Avoiding administrative privileges
3. Using Web application firewall

IPS can detect SQL injection strings or save strategies that are vulnerable to attacks. However, we believe that IPS is unreliable because square inoculation threads are disposed to incorrect positivity. Those managers of protection who exclusively consider IPS application are inundated by viable warnings of SQL injection. Nevertheless, considering the correlation of the SQL inoculation mark along any other breach, including enquiry-stage get-in-to-manipulate violation, it is possible to manipulate the violation, and a real assault can be pinpointed with extreme precision. During normal business operations, SQL inoculation mark as well as any infringement does not probably occur similarly in the submission.

4.6. Sixth Threat - Weak Audit Trail

The inspiration behind the implementation of any database must involve the automated documentation of all sensitive and/or irregular database transactions. In several ways, a shaky database audit policy poses a serious threat to the company.

- Regulatory danger
- Deterrence
- Detection and Recovery
- Lack of User Accountability
- Performance Degradation
- Separation of Duties
- Limited Granularity
- Proprietary.

Prevented by: Preventing Weak Audit

The majority of the vulnerabilities associated with local audit equipment are resolved by high-quality network-based audit home equipment.

- High performance
- Separation of Duties
- Cross-Platform Auditing.

4.7. Seven Threat - DoS

Another common form of cyber-attack is the DoS in which demonstrative consumers are denied access to network applications or data. Common techniques may be used to establish DoS conditions, in which many of them can be linked to the mentioned vulnerabilities.

DoS can be motivated by different factors. Ransom scams are often associated with DoS attacks linked to computer, in which a remote attacker constantly crashes computers before depositing money into a global bank account by the victim. A bug infection instead maybe blamed for DoS. Regardless of availability, the seriousness of the threat for most companies maybe posed by DoS [17].

Prevented By: DoS Preventive

DoS preventive necessitates several layers of protection. Protections at the network, software, and database levels are all critical. This study focuses on database-specific security. Recommendation focuses on link charge manipulations, query access control, IPS, and reaction timing controls the database-specific contexts.

4.8. Eighth Threat - Weak Authentication

By stealing or otherwise obtaining login credentials, attackers can predict the identity of legitimate database customers using vulnerable authentication schemes. To acquire credentials, an attacker can use various number of methods available [18].

- Cryptanalytic attack
- Social Engineering
- Direct Credential Theft.

Prevented By: Preventing Authentication Attacks

1- Strong authentication

It is crucial to use the most advanced realistic authentication technologies and rules. Where possible, two-factor (tokens, certificates, biometrics, etc.) authentication is preferred. Unfortunately, cost and ease of use, frequently furnish authentication impracticality. These situations necessitate the implementation of strict username and password policies (least possible duration, gender selection, as well as obscurity).

2- Directory integration

However, incorporation of strong authentication mechanisms with business enterprises catalogs the substructure for scalability and simplicity of use. The directory structure, among others, allows the user to consider using a particular login detail for numerous database and program. However, it increases the usefulness of double-factor authentication system. Moreover, it is easier for consumers remembering alternative passphrases on a regular basis.

4.9. Ninth Threat - Backup Data Exposure

Database backup in certain cases, storage media is completely unregulated. As a result, stealing backup disks and hard drives have been the focus of many high-profile security breaches.

Prevented by: Preventing Backup Data Exposure

Both backups of databases require encryption application. Indeed, certain carriers stated that the potential products of DBMS not necessarily support the unencrypted backup's usage. However, online produce of database statistics encryption advised frequently. Nevertheless, key management issues of presentation and cryptographic are frequently impractical, and granular privilege controls described above are in general considered as a weak substitute.

5. METHOD FOR PROTECTING DATABASE SYSTEM

In this section, two types of method have been explained, the first one is to eliminate security risks, any company must have a security policy in place that must be followed.

Authentication is crucial in security policy since proper authentication reduces the probability of attacks. On different database objects, different users have different access rights. The management of access rights is the responsibility of access control systems.

To safeguarding database statistic substances, besides the majority DBMS supports it is the greatest elementary practices [5]. The control methods concerning database protection depicted (See Fig. 1).

5.1. Access Controller

Is the most basic service, which any DBMS can offer? It has safeguarded data from unauthorized reads and writes. All contact to the database and objects of other systems must adhere to the policies defined by access control. Errors may be serious enough to cause issues in a company's operations. Admission rights controlling can aid mitigating dangers, which have a direct effect on the database protecting the main server. The access control is able to prevent the deletion or changing of a table made by accident. The access control can rollback, and prevent the deletion of particular files. access control systems consist of:

- File permissions to create, read, edit, or delete files on the server.
- Program permissions, the rights of executing an application program on the server.
- Data rights, the rights of retrieving, or updating data in a database.

5.2. Inference Strategy

Data protection at a particular level is critical. It is used once the processing of specific data in the form of facts should stop at a maximum level of protection. It helps the determination of how to keep knowledge from being posted. The inference control aims to prevent information from being revealed indirectly. Unauthorized data disclosure can ocur in one of three ways:

- Correlated data - a popular channel when visible data X and invisible data Y are semantically linked.
- Missing data - NULL values in the query masks a sensitive data. That way, existed data could be detected.
- Statistical inference - this is common in database, which contain a numerical data in regards to individuals.

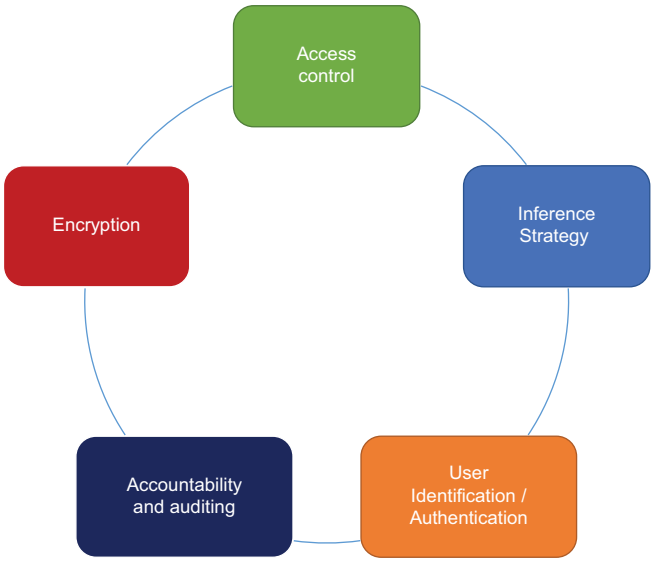

Fig. 1. Control methods for protecting database system.

5.3. Identification or Authentication of the User

It is better to know your users as a basic security requirement. After you've classified people, you'll need to determine what privileges and access permissions they have, as well as verifying their data that must use.

Until a user is allowed to construct a database, they should be authenticated in several ways. User identification and authentication are a part of database authentication, the OS or network service can perform an external authentication process. To establish user authentication Secure Sockets Layer (SSL), business parts, and middle-tier server authentication, also known as proxy authentication, can all be used. it is the most basic prerequisite for ensuring protection when considering the identification process which identifies a collection of people who are permitted to access the data. To ensure confidentiality, the authentication of identity initiates preventing unauthorized users from modifying sensitive data. Attackers make use of various methods such as bypass authentication, default password, privilege escalation, brute-force password guessing, and rainbow attack when attempting to breach a user identity and authentication [16].

5.4. Audit and Accountability

Database/non-database users audit and monitor a configured database behavior. Accountability refers to the method of keeping track of user activities on a device. To ensure the physical integrity of the data, auditing checks and accountability are required which necessitates a specific database access carried out with auditing and maintaining the resiliency of the data. If users' authentication accesses a resource successfully, the system will track all successful and unsuccessful attempts, and attempted accesses and their statuses will show in the audit trail files [16].

5.5. Encryption

It is a method of translating information into cipher or code that only those who have access to the cipher text key can make it ready. Encrypted data is the referral to cipher or encoded text. In a database, there are two states for data security. Data is in two statuses: at rest and in motion – data stored in a database, on a backup disk, or a hard drive. Once transiting through the network, it necessitates the use of various encryption solutions. Any of the problems of data

at rest can be solved by encrypting it. Utilize solutions such as SSL/Transport Layer Security for Data in Transit [16].

In the second method any organization may make advantage of a using new technology tool that has a significant effect on database security such as:

1. Database Firewalls: Are a kind of Web Application Firewall that monitor databases to detect and defend against database-specific attacks, which are usually aimed at gaining access to sensitive data contained in the databases. Database Firewalls also allow you to monitor and audit every database access via the logs they keep. Specific compliance reports for laws like as PCI, SOX, and others may be generated by a Database Firewall [19].
   Herse some tool:
   - Cloudflare
   - Site Lock
   - Tufin Secure Track.
   - ManageEngine Firewall Analyzer.
   - FireMon.
   - AlgoSec.
2. Real Time Data Monitoring (RTDM): An Admin may examine, analyze, and change the addition, deletion, modification, and usage of data on software, a database, or a system using RTDM. Through graphical charts and bars on a single interface/dashboard, data managers may examine the general operations and functions done on the data in real time, or as they happen [20].
   Herse some tool:
   - Real-time Database profiler tool
   - Firebase console
   - Cloud Monitoring
3. Multi-factor database Authentication: Is a technique and technology for confirming a user's identification that requires two or more credential category kinds for the user to log into a system or complete a transaction. This technique requires the effective Answering of at least two separate credentials such as: Entering password, email verification, phone verification, or answering security question [21].
   Herse some tool:
   - LastPass
   - Duo Security
   - Ping Identity
   - RSA SecurID Access

6. CONCLUSION

The database security problems and research into various issues affecting the industry have frequently been listed in this survey. Organizations are now dependent on documents to make decisions about different business processes that will improve their bottom line. As a result, it is a smart idea to keep confidential details secure from prying eyes. Server security research papers have attempted to investigate the issues of potential assaults to database systems such as loss of confidentiality and honesty. Because of the knowledge and volume contained in databases, they are the most common and simple targets for attackers. There are many options to accommodate a database. Today, there are several forms of attacks and threats against which a database should be secured. This paper discusses the decisions that must be made in order to protect personal data from attackers. It also goes into depth about how a loss of privacy can lead to extortion and humiliation in the workplace. This survey also looked at strategies for dealing with any form of hazards. Views and authentication should be used in this case. Another method is to use an encryption strategy, which means the information is secured so that if an InT finds it, he or she is unable to use it, and the criteria for a reliable DBMS were also discussed.

REFERENCES

[1] M. Malik and T. Patel. "Database security attacks and control methods". *International Journal of Information Technology*, vol. 6, no. 1/2, pp. 175-183, 2016.

[2] I. Ghafir, J. Saleem, M. Hammoudeh, H. Faour, V. Prenosil, S. Jaf, S. Jabbar and T. Baker. "Security threats to critical infrastructure: The human factor". *The Journal of Supercomputing*, vol. 74, no. 10, pp. 4986-5002, 2018.

[3] P. K. Paul and P. S. Aithal. "Database security: An overview and analysis of current trend". *International Journal in Management and Social Science*, vol. 4, no. 2, pp. 53-58, 2019.

[4] S. B. Sadkhan. "Related Papers". *Over Rim*, pp. 191-199, 2017.

[5] H. Kothari, A. Suwalka and S. Kumar. "Various database attacks, approaches and countermeasures to database security". *International Journal of Advance Research in Computer Science and Management Studies*, vol. 5, no. 5, pp. 357-362, 2019.

[6] J. C. Ogbonna, F. O. Nwokoma and A. Ejem. "Database security issues: A review". *International Journal of Engineering Inventions*, vol. 6, no. 8, pp. 1812-1816, 2017.

[7] T. Dharmakeerthi. "A Study on Security Concerns and Resolutions". *Researchgate. Net*, No. May, 2020.

[8] E. Technology and V. Sharma. "An analytical disparity of harbor tools erection for database system". *International Research Journal of Modernization in Engineering Technology and Science*, vol. 3, no. 2, pp. 501-510, 2021.

[9] U. Albalawi. "Countermeasure of Statistical Inference in Database Security". *Proceeding 2018 IEEE International Conference Big Data, Big Data 2018*, pp. 2044-2047, 2019.

[10] J. Juma and D. Makupi. "*Understanding Database Security Metrics: A Review*". Vol. 1. Mara International Journal of Social Sciences Research Publications, pp. 40-47, 2017.

[11] J. C. Odirichukwu and P. O. Asagba. "Security Concept in Web Database Development and Administration a Review Perspective. *2017 IEEE 3rd International Conference Electro-Technology National Development NIGERCON 2017*, vol. 2018-Janua, pp. 383-391, 2018.

[12] A. Mousa, M. Karabatak and T. Mustafa. "Database Security Threats and Challenges". *8th International Symposium Digital Forensics Secur. ISDFS 2020*, vol. 3, no. 5, pp. 810-813, 2020.

[13] S. Singh and R. K. Rai. "A Review Report on Security Threats on Database". *International Journal of Computer Science and Information Technologies*, vol. 5, no. 3, pp. 3215-3219, 2014.

[14] H. Tabrizchi and M. K. Rafsanjani. "*A Survey on Security Challenges in Cloud Computing: Issues, Threats, and Solutions*". Vol. 76. Springer, United States, 2020.

[15] P. Sharma. "Database Security: Attacks and Techniques". *International Journal of Scientific and Engineering Research*, vol. 7, no. 12, pp. 313-319, 2016.

[16] S. S. Sarmah. "Database Security Threats and Prevention". *International Journal of Computer Trends and Technology*, vol. 67, no. 5, pp. 46-53, 2019.

[17] T. Mahjabin, Y. Xiao, G. Sun and W. Jiang. "A survey of distributed denial-of-service attack, prevention, and mitigation techniques". *International Journal of Distributed Sensor Networks*, vol. 13, no. 12, 2017.

[18] H. B. Hashim. "Challenges and security vulnerabilities to impact on database systems". *Al-Mustansiriyah Journal of Science*, vol. 29, no. 2, p. 117, 2018.

[19] W. Lee. "*Lecture Notes in Electrical Engineering 461 Proceedings of the 7th International Conference on Emerging Databases*", 2019.

[20] I. Kotsiuba, M. Nesterov, Y. Yanovich, I. Skarga-Bandurova, T. Biloborodova and V. Zhygulin. "Multi-Database Monitoring Tool for the E-Health Services. *Proceeding 2018 IEEE International Conference Big Data, Big Data 2018*, pp. 2442-2448, 2019.

[21] C. Hamilton and A. Olmstead. "Database Multi-factor Authentication Via Pluggable Authentication Modules". *2017 12th International Conference Internet Internet Technology and Secured Transactions ICITST 2017*, pp. 367-368, 2018.

ORIGINAL RESEARCH ARTICLE UHD JOURNAL OF SCIENCE AND TECHNOLOGY

Knowledge Management Functions Applied in Jordanian Industrial Companies: Study the Impact of Regulatory Overload

Muzhir Shaban Al-Ani[1], Shawqi N. Jawad[2], Suha Abdelal[2]

[1]Department of Information Technology, University of Human Development, College of Science and Technology, Sulaymaniyah, KRG, Iraq, [2]Department of Management, Amman Arab University, College of Business, Amman, Jordan

ABSTRACT

This research aims to study the impact of electronic information overload on knowledge management functions in Jordanian industrial companies. The research population included all Jordanian industrial companies listed on the Amman Stock Exchange. A simple random sample of 30% of the research population of 1242 seniors and middle managers in the research population was done to 373 individuals. 206 questionnaires are successfully retrieved to be analyzed. Descriptive and heuristic statistical methods such as simple and multiple regression analysis were applied using SPSS.16 program. The obtained result indicated that there is a statistically significant impact of the electronic information overload (organizational overload) on the knowledge management functions (acquisition, generation, transmission, sharing, and application of knowledge) in Jordanian industrial companies. In the scope of the results, this work made a number of recommendations, including: Adopting an organizational aspect that suits the nature of the tasks that the industrial companies operate in Jordan, in addition to providing technical capabilities to reduce the electronic information overload faced by the industrial companies in Jordan while practicing their tasks.

Index Terms: Knowledge Management, Organizational Overload, Statistical Analysis, Jordanian Industrial Companies

1. INTRODUCTION

The last decades and the present century have witnessed an acceleration in the pace of change toward the knowledge economy, as the production and organization of knowledge have become a top priority for business organizations [1]. Knowledge is an essential ingredient for driving economic growth in countries [1]. Knowledge has already become an intangible asset of the organization, prompting organizations to rearrange their priorities (National Information Technology Council, 2004) [2]. Many technological applications have been developed that have enhanced organizational capabilities and have created a huge influence of information and their use in organizations [3].

Business organizations now face a clear challenge as a result of the knowledge and technical revolution in all areas of knowledge [4]. Effective decision for enabling senior management to enhance its role in investing in technical and knowledge developments is very important to face the turbulent environment and its requirements [5]. It is necessary to identify theoretical foundations and theoretical structures that are capable of achieving the goals of the organization [6]. Business organizations in general and industrial companies in particular are affected by the dramatic change in the business environment and its drive toward the use of information technology in which information has become a key resource

Access this article online

DOI:10.21928/uhdjst.v5n2y2021.pp47-56 E-ISSN: 2521-4217 P-ISSN: 2521-4209

Corresponding author's e-mail: Muzhir Shaban Al-Ani, Department of Information Technology, University of Human Development, College of Science and Technology, Sulaymaniyah, KRG, Iraq. muzhir.al-ani@uhd.edu.iq

Received: 14-09-2021 Accepted: 28-10-2021 Published: 04-11-2021

for the growth and progress of these organizations [7]. The information becomes the most important in terms of its accessibility and storage in electronic databases and then re-employment in which generated the overload of electronic information [8].

Knowledge management is one of the most recent topics in the world of management and it is of great interest to stakeholders in business organizations Sekaran, [9]. In addition, this increased of interest in the rush of organizations of various types toward the possibility of applying knowledge [10]. Knowledge management is gateway to the development of contemporary organizations to enable them to meet future challenges [10]. The significance of knowledge management in business organizations is not in the knowledge itself, but in the value added to these organizations, in addition to the role it plays in transferring organizations to the knowledge economy that emphasizes investment in knowledge capital [11].

The business environment of organizations is characterized by rapid change, dominated by the information and communication revolution [12]. Knowledge is the weapon adopted by organizations to ensure their growth and sustainability [13]. Participatory knowledge is widespread and increased by practice and use. Knowledge is an important resource that contributes to the success of various organizations [14].

Modern business organizations constantly strive to adapt in every stage of development in the knowledge economy and keep pace with the requirements of the era [12]. Electronic information systems have become the basis for management and productivity processes in business organizations of all kinds (Bawden and Robinson, 2008) [13]. These systems are playing a clear role in the processes related to the objectives, business, marketing, and productivity of the organizations [14].

This study is characterized by the fact that it verified the regulatory overload in electronic information within the Jordanian industrial companies. This was done through a survey questionnaire to study the reality of these companies within this new environment. This study is the first of its kind in this field and within the Jordanian industrial sector.

2. STATEMENT OF THE PROBLEM

The organization and its staff face with the phenomenon of information overload that requires attention, study, and treatment. Since the companies in general and the Jordanian industrial companies in particular deal with the large amount of information that is available as electronic data. This weakens their position in making various decisions and makes mistakes due to the excessive overload in the aspects of knowledge information. This requires companies to find a new mechanism to enable them to meet these overloads with the importance of finding a form of control over the application of this mechanism to determine the prospects for dealing with their data.

Therefore, the research seeks to measure the "Knowledge Management Functions Applied in Jordanian Industrial Companies: Study the Impact of Organizational Overload."

3. RESEARCH QUESTIONS

The following questions are achieved to perform the research:

- What is the level of managers' perceptions of the regulatory overload in two dimensions (channels of communication and regulatory environment) in Jordanian industrial companies?
- What is the level of managers' perceptions of the regulatory overload in two dimensions (communication channels and regulatory environment) in the Jordanian industrial companies?
- What is the level of perceptions of managers in the impact of the regulatory overload on the functions of knowledge management dimensions (acquisition, generation, transport, sharing, and application of knowledge) in the Jordanian industrial companies?

4. RESEARCH OBJECTIVES

The research hopes to achieve the following objectives:

- Measuring the impact of variables related to the electronic information (regulatory overload) on knowledge management functions in the researched companies
- Identify the positive aspects that help to improve knowledge management functions and the negative aspects that limit the effectiveness of these functions
- Measuring the level of application of knowledge management functions by industrial companies in Jordan; to reach appropriate recommendations that can be made to deal with electronic information (regulatory overload).

5. RESEARCH HYPOTHESIS

There is no statistically significant effect at the level ($\alpha \leq 0.05$) of the regulatory overload (channels of communication and regulatory environment) in the knowledge management functions (acquisition, generation, transmission, sharing, and application of knowledge) in Jordanian industrial companies.

6. RESEARCH MODEL

The study model is designed with its variables regarding to the problem of the study and its hypothesis and to achieve its purpose and reach its specific objective (Fig. 1).

7. ELECTRONIC INFORMATION OVERLOAD

The overload of electronic information in businesses companies is very important and requires a series of creative measures to be achieved.

The use of electronic data warehouse and the application of electronic knowledge in the organization are constantly re-using knowledge within the organization [16]. Noted that technology caused the explosion of information due to the lower costs of multimedia technology, which simplified access to information and helped in their publication. As Whelan and Teigland, 2010, [16] explained, the information overload is a problem facing contemporary organizations.

According to Himma, 2007, [17], there is a difference in meaning between these two terms, contrary to what some think that the overload means an increase. This was made clear when the excess quantity was seen as a precautionary measure. Therefore, this quantity had no negative effect and could be dealt without incurring high costs. If this increase becomes negative for individuals and becomes problematic, in which there is an overload.

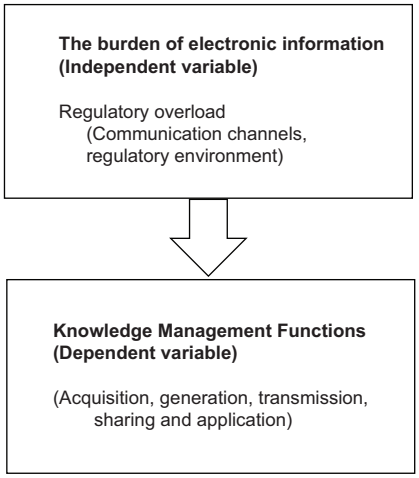

Fig. 1. Research model.

Eppler and Mengis, 2003, [15] indicated that the overload of information appears on the receipt of a large amount of information beyond the capacity of individuals to deal with a process, which reflects negatively on the quality of the decision. Grise and Gallupe [18] defined the overload of information as having a lot of information with the inability of the individuals concerned with that vast amount of information. Mulder *et al.*, 2006, [19] added that the information overload increases the sense of tension when the volume of information exceeds the capacity to be processed. Kim *et al.*, 2007, [20] considered that the information overload meant confusion in the information received that impeded learning and impaired individuals' decision-making capacity.

Farhoomand and Drury, 2002, [21] have warned that the information overload may be created by situations such as the Web and the Internet, which are the main reasons for the overloads because of the abundant information you provide from external resources. The complexity of organizational tasks and the lack of sufficient number of people to complete the tasks, which all lead to an information burden, as well as the huge amount of information coming to the offices of managers daily. Not to mention the availability of a lot of information that is not understood by individuals or not knowing whether this information serves their orientation or not (Dubosson and Fragniere, 2009) [22].

Lesa, 2009, [23] reported that the regulatory overload is endemic in today's fast-track environment. Inadequate regulatory environment for organizational learning impedes the flow of information, ideas, and knowledge into the organization, resulting in an information overload. Lesa, 2009, [23] mentioned that there are a number of strategies for dealing with the information overload at the organizational level, including: Establishing task-specific task forces, building informal relationships across the organizational structure and applying modern information management systems. The organizational overload (Wilson, 2001) [24] is also a situation in which the information overload flowing widely from individuals across the organizational structure is reflected in a reduction in the overall effectiveness of the organization's operations management. Filippov and Iastrebova, 2010, [25] explained that the regulatory overload implies an imbalance between the requirements for processing organizational information and the ability to process that information

within the organization. The organizational structure helps facilitate the collection, processing, and dissemination of information and the protection of individuals from the burden of information.

Below some concepts that clarify knowledge will be addressed:

7.1. Knowledge

Information combined with experience and intuition (Yan, 2009) [26]. That is, knowledge is information that has been processed, organized, and structured to be applicable (Hester, 2009) [27]. Knowledge consists of a combination of values, contextual information, and expertise, as well as new information and expertise that exists in knowledgeable minds, organizational routines, documents, rules, processes, and practices in organizations (Haytham, 2005) [28]. Knowledge is also seen as an intellectual capital and a critical component of today's organizations and is growing with increasing practice and learning (Najm, 2005) [29]. As knowledge is a combination of experience, practice, judgments, and values of both the individual and the organization that are reflected in the work of employing knowledge for the desired goals (Salwa, 2008) [30].

There are two types of knowledge: Tacit knowledge and explicit knowledge. Tacit knowledge is the knowledge stored in human minds and behavior, and what is generated by learning from past experiences that are difficult to document and transfer to others. Explicit knowledge is knowledge that can be shared among individuals, groups, and organizations and can be documented, stored electronically, transmitted and used through various means (Jazar and Talaat, 2005) [31].

7.2. Knowledge Management

Fernandez *et al.*, 2004, [32] defined knowledge management as doing what is needed to maximize the benefit of knowledge resources. Since knowledge management is the gateway to adding and generating value by mixing knowledge elements to create better knowledge combinations, this will change the role of data, information and knowledge to flow individually (Najm, 2005) [29]. Mohammed and Ziad, 2010, [33] described knowledge management as the organization's knowledge resources and assets, adaptability and learning, increasing the creative process, sharing and optimal use of these assets. Ashoc, 2004, [34] also noted that knowledge management is effective learning processes associated with the exploration, exploitation, and sharing of human knowledge (explicit and Tacit), which applies appropriate civilization, technology, and culture to extract performance and intellectual capital.

Regarding to the organizational implications of knowledge management, it was pointed out that knowledge management contributes to the generation of knowledge that seeks to improve the performance of organizations through four dimensions (Fernandez *et al.*, 2004) [32]. These dimensions are influencing individuals, influencing processes, influencing product, and influencing organizational performance.

There are five knowledge management functions as follows:

- Knowledge acquisition: Knowledge acquisition is a function that seeks to gain knowledge and obtain it from a variety of documented sources, as well as the acquisition of undocumented sources that stored in the minds of individuals and issued through their behavior. Knowledge can be gained from experts and stakeholders and information technology plays an important role in supporting knowledge acquisition through its role in data capture, classification, processing, and harnessing to build the competitive advantage of an organization (Kamel, 1999) [35]. The function of knowledge acquisition goes through four stages (Kamel, 1999) [35]: Collection, interpretation, analysis, and design
- Knowledge generation: Knowledge generation comes from a variety of sources and channels to expand the repositories of organizational memory and enable the organization to creatively solve solutions to its problems leading to innovation. It is individuals who generate knowledge within the organization. This will be done through four processes of knowledge transfer: Social participation, embodied external knowledge, integrated internal knowledge, and synthetic knowledge
- Knowledge transfer: Knowledge transfer depends on several factors that need to be considered: leadership, support for organizational structures, absorptive capacity, degree of privacy, degree of complexity, and dependability of knowledge vocabulary. Knowledge is transferred through the use of management information systems, training and e-learning systems using the internet (Ashoc, 2004; Alavi and Leidner, 2001) [10], [34]
- Knowledge sharing: Knowledge sharing is an important element in production, responding to environmental changes, promoting opportunities, outperforming competitors, and maintaining the effectiveness of modern organizations. The knowledge base of the organization is increasingly shared by individuals with their knowledge and experience formally through regular formal meetings. The sharing of knowledge with individuals prevents the loss, fading and erosion of that knowledge over time. In addition, sharing knowledge between organization and other organizations

enhances the knowledge storage that will be available in organization repositories from those of other organizations [36]

- Knowledge application: It is a knowledge management purpose where modern organizations apply knowledge using web-based technology systems and knowledge retrieval techniques. Furthermore, these systems provide the ability to access, transfer and use information in a timely and appropriate manner and to communicate with the right person [36].

Zainab, 2009, [37] presented a case study of King Abdul Aziz University, which measured the readiness of organizations to apply knowledge management through the four dimensions of knowledge management (human dimension, technological dimension, strategic dimension, and operations dimension). He concluded that KAU has a readiness to manage knowledge with a medium degree. Carlevale, 2010, [38] concluded that technology causing the information overload caused by the huge amount of e-mail that managers are exposed to every day in which generating more pressure. That e-mail is the biggest cause of the burden of information on these managers on a daily basis, hinders the decision-making process and that incoming emails need to be filtered and managed.

Hodge, 2010, [39] noted that there is a positive correlation between knowledge management processes (capturing, storing, classifying, and applying) and knowledge management capabilities (lessons learned, experiences, and knowledge documents). Salwa, 2008, [30] emphasized the role of knowledge management and information technology in achieving competitive advantages that concluded the banks in question apply the knowledge management technology system in all units and departments within banks, although there is no organizational unit or special department for knowledge management and information technology. Inside any bank (Ismail and Yusof, 2010) [40] showed that there is a positive relationship between individual factors (awareness, confidence, and personality), and the quality of knowledge sharing. Personal style (extrovert and introvert) is the most important for the quality of knowledge sharing followed by trust and awareness.

8. METHODOLOGY

8.1. Research: Questionnaire Design

The overload of electronic information in businesses companies and the relevant factors and tasks that have been explained in the previous sections. This section explains how the questionnaire is designed and what are their paragraphs. The questionnaire is designed based on the Likert scale of five fields: Strongly agree, agree, neutral, disagree, and strongly disagree. Regarding to the research model the designed questionnaire is divided into: Independent variable (regulatory overload) and dependent variable (knowledge management functions).

- Independent variable (Regulatory Overload) (Fig. 2): This field is divided into two parts: Communication channels and regulatory environment. Communication channels related to loss of boundaries between roles, tasks and work duties, which affects the movement and exchange of information within the departments of a single organization due to the inadequacy or lack of clarity of the organizational structure. Regulatory environment related to the availability of work requirements in the work environment through which the organization can control the variables of its environment of humans, devices and the administrative decisions
- Dependent variable (knowledge management functions) (Fig. 3): This field is divided into five parts: knowledge acquisition, knowledge generation, knowledge transfer, knowledge sharing, and knowledge application. Knowledge acquisition according to obtain knowledge from various internal sources (such as collaboration, learning, feedback from staff, workshops, training programs, and databases within which knowledge is stored) and external (such as competitors, customers, consultants, attract experienced and competent personnel, and establish relationships with partners and allies). Knowledge generation regarding to derive and create new creative knowledge from existing knowledge within the organization to secure various types of knowledge for the benefit of future decisions which are concerned with equipping workers in the knowledge field with graphics and analysis, and this is done through teaching, learning, research, and development. Knowledge transfer related to communicate the right knowledge to the right person in the appropriate manner (communications, bulletins, reports, staff movements, and use of technological means to facilitate knowledge transfer), and unintentional (informal meetings of individuals) at the right cost and at the right time. Knowledge sharing regarding to circulate and exchange of various types of knowledge among individuals. Interacting with others' dialogues inside and outside the organization, securing collective cooperation among them, reaching out and working simultaneously on the same document and from different locations to form new creative mental ideas. Knowledge application

#	Description	Strongly agree	Agree	Neutral	Disagree	Strongly disagree
1	The technological burden affects the company's departments and internal departments' interaction with the communication network for information exchange.					
2	The administrative organization of the company is suitable for business flow.					
3	The key roles of working in the company, are clearly defined.					
4	The exchange of information electronically between company divisions is an additional burden for employees.					
5	The technological burden complicates the processing of data.					
6	Administrative organization helps in achieving efficient and effective business performance in the company.					

#	Description	Strongly agree	Agree	Neutral	Disagree	Strongly disagree
7	The company is slow to provide business requirements in the electronic environment when the technological burden increases.					
8	The working environment gets worse as the technological burden increases.					
9	The company's rush to have advanced management technology systems, which makes it unable to operate efficiently.					
10	The technological burden is reflected in the company's ability to control business variables in the electronic environment.					
11	The worsening of the technological burden shows the problems of providing adequate personnel.					
12	The large amount of information received, reduces the effectiveness of the company's operations.					

Fig. 2. Regulatory overload questioner.

#	Description	Strongly agree	Agree	Neutral	Disagree	Strongly disagree
1	The company receives feedback from employees on a permanent basis.					
2	The company uses training programs and workshops as a way to equip employees with the necessary knowledge.					
3	The company acquires knowledge from its partners or allies by establishing relationships with them.					
4	The company is keen to provide employees with information consistent with its products					
5	The company recruits experienced and competent staff to work for it to enhance its knowledge.					

#	Description	Strongly agree	Agree	Neutral	Disagree	Strongly disagree
6	The company always synthesizes information collected from multiple sources, in order to generate new knowledge.					
7	The company adopts advanced R&D policies to generate new knowledge.					
8	The company provides incentives for new innovations and knowledge.					
9	The company encourages brainstorming among employees to generate new ideas.					
10	The company seeks to meet its knowledge needs by bridging the knowledge gap.					

#	Description	Strongly agree	Agree	Neutral	Disagree	Strongly disagree
11	Company information flows smoothly across functional boundaries.					
12	The company contributes in sending scholarships for specializations in order to transfer knowledge.					
13	IT helps bring people in need of knowledge closer to those who have it.					
14	The company encourages dialogue between employees to impart knowledge.					
15	The company makes periodic transfers between departments and departments as a means of knowledge sharing.					

#	Description	Strongly agree	Agree	Neutral	Disagree	Strongly disagree
16	The company employs informal meetings and dialogues for the purposes of expanding sharing of knowledge.					
17	The company has an atmosphere of mutual cooperation to support knowledge sharing.					
18	The company fosters a culture of knowledge sharing among employees.					
19	The company provides multiple channels for knowledge sharing (Internet, Extranet and Intranet).					
20	The company holds meetings to discuss its annual reports to get feedback.					

#	Description	Strongly agree	Agree	Neutral	Disagree	Strongly disagree
21	The organization holds training courses on how to use and apply the knowledge gained to achieve specific objectives.					
22	Directors recognize that the Organization has a non-invested knowledge balance.					
23	The company uses modern technologies to apply knowledge and invest its returns.					
24	The management of the company is keen to use the new knowledge generated by the company.					
25	The company is keen to ensure that employees are aware of the methods of applying the acquired knowledge.					

Fig. 3. Knowledge management functions questioner.

dealing with the utilizing of knowledge to support innovation, development of people and resource, business improvement, using specific technology systems and knowledge dissemination channels at all organizational levels.

8.2. Research: Population and Sample

The study population consisted of the industrial companies listed in the Amman Stock Exchange, and the Inspection and Analysis Unit included all directors working in the higher managements (general managers, their assistants, or their representatives), as well as managers working in the middle managements (managers of the main departments and heads of departments). A relatively random sample of (1242) members of the sampling and analysis unit, (30%) was selected to become the sample (373). The number of questionnaires valid for statistical analysis (206) questionnaires (55%) of the total questionnaires distributed. Regarding to data sources to achieve the objective of the study, secondary sources were adopted, which include those data and information published in various library sources for review of previous literatures. The primary sources of data were the questionnaire built for this purpose which was aimed at obtaining the raw data to complete the applied aspect of the study in terms of handling the study questions and testing the hypotheses.

9. RESULTS AND DISCUSSION

Descriptive statistical analysis of study variables is applied on the obtained data. A low (less than 2.33), middle (2.33–3.66), and high (3.77 and above) was used to determine the relative importance of the respondents' perceptions of the study questions based on the Likert-5 scale, and the arithmetic and standard deviations are shown in Fig. 4.

Question 1: What is the level of perceptions of managers working for the regulatory overload (channels of communication and regulatory environment) in Jordanian industrial companies?

The results of Fig. 1 show that the level of perceptions of the respondents regarding the regulatory overload was high and the researchers attribute the result to the fact that the administrative organization in the industrial companies serves the nature of the work adopted in the presence of electronic systems. Regarding the regulatory environment, the result indicates that there is an average level of influence of the regulatory environment as one dimension of the regulatory overload on knowledge management where the general arithmetic mean of the type of technology was (3.35) and the standard deviation (0.67).

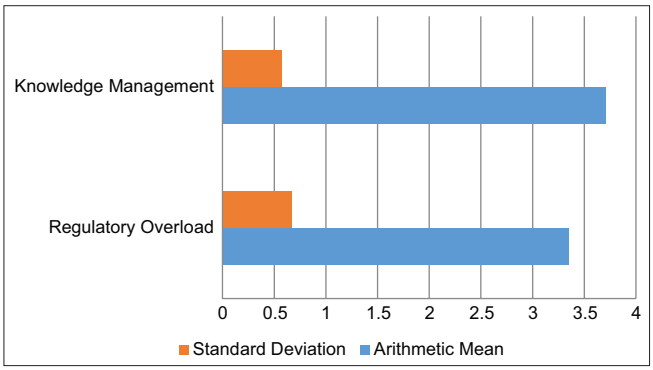

Fig. 4. Arithmetic mean and standard deviation (regulatory overload).

Question 2: What is the level of perceptions of managers working for knowledge management functions (acquisition, generation, transmission, sharing, and application of knowledge) in Jordanian industrial companies?

The results of Fig. 1 indicate that the respondents' level of response was high where arithmetic mean is 3.71 and the standard deviation is 0.57. The roles of the work they do and this knowledge are done in accordance with the surrounding environment, as they derive from the external environment that reflects the relations of companies with customers, as well as the relationships between companies at the level of industry.

Multiple regression and the accompanying tests are used to verify the validity of the hypotheses. In addition, F-test for the regression model significance, *t*-test for the significance of the effect, and the value of the coefficient of determination R^2 are used to determine the interpreted percentage of independent variables in the dependent variable, depending on the statistical significance values extracted under the statistical software.

In this study, there is no statistically significant effect at $\alpha \leq 0.05$ level of the regulatory overload (communication channels and regulatory environment) in the knowledge management functions (acquisition, generation, transmission, sharing, and application of knowledge) in Jordanian industrial companies.

TABLE 1: The impact of regulatory overload on knowledge management functions

Dependent variable	Coefficient of determination	Standard deviation	F-value	Degree of freedom	Regression coefficients				
					Independent variable	β	Standard error	T-test	Significance level
Acquisition of knowledge	0.318	0.101	22.966	(1,204)	Regulatory overload	0.373	0.078	4.792	0.000
Generation of knowledge	0.339	0.115	26.416	(1,204)	Regulatory overload	0.419	0.081	5.140	0.000
Transmission of knowledge	0.422	0.178	44.278	(1,204)	Regulatory overload	0.499	0.075	6.654	0.000
Sharing of knowledge	0.337	0.113	26.091	(1,204)	Regulatory overload	0.422	0.083	5.108	0.000
Application of knowledge	0.311	0.097	21.813	(1,204)	Regulatory overload	0.337	0.072	4.670	0.000
Functions of knowledge management	0.398	0.158	38.396	(1,204)	Regulatory overload	0.410	0.066	6.196	0.000

TABLE 2: Multiple regression analysis to test the effect of organizational burden dimensions on the dimensions of knowledge management functions

Dependent variable	Coefficient of determination	Standard deviation	F-value	Degree of freedom	Significance level	Regression coefficients				
						Independent variable	β	Standard error	t-test	Significance level
Acquisition of knowledge	0.493	0.243	32.556	(203,2)	0.000	Communication Channels	0.631	0.081	7.835	0.000
						Regulatory Environment	0.090	0.057	1.568	0.118
Generation of knowledge	0.492	0.242	32.456	(203,2)	0.000	Communication Channels	0.655	0.085	7.701	0.000
						Regulatory Environment	0.067	0.061	1.114	0.267
Transmission of knowledge	0.521	0.271	37.776	(203,2)	0.000	Communication Channels	0.612	0.080	7.689	0.000
						Regulatory Environment	0.024	0.057	0.421	0.674
Sharing of knowledge	0.485	0.235	31.232	(203,2)	0.000	Communication Channels	0.652	0.087	7.534	0.000
						Regulatory Environment	0.063	0.062	1.024	0.307
Application of knowledge	0.457	0.209	26.741	(203,2)	0.000	Communication Channels	0.534	0.076	7.011	0.000
						Regulatory Environment	0.059	0.054	1.083	0.280
Functions of knowledge management	0.562	0.316	46.863	(203,2)	0.000	Communication Channels	0.617	0.067	9.170	0.000
						Regulatory Environment	0.051	0.048	7.835	0.288

Table 1 shows that the simple regression model is applied to measure the impact of the organizational burden on the dimensions of knowledge management functions, in which have significant impact. This effect is significant based on the test value ($t = 6.196$) when compared with the value of the significance level (sig = $0.000 \leq 0.05$).

Table 2 indicates that multiple regression model to measure the effect of both dimensions (communication channels and organizational environment) in knowledge management functions is significant, where the value of ($F = 46.863$) at the level of significance (sig = 0.000). Together, the two variables explain that R2 = 31.6% of the differences in the values of knowledge management functions are reinforced by this result ($t = 9.170$).

10. CONCLUSIONS

Focusing of the previous discussions, the study reached a number of conclusions as below:

- The results of the study pointed to the relative importance of the regulatory overload in relation to the communication channels. This reflects that the roles of individuals are not clearly defined. The results of the study concur with (Lesa, 2009) [23] in the regulatory section, considering that regulatory factors are the primary cause of the information burden phenomenon according to (Lesa, 2009) [23]. The study also agreed with (Raoufi, 2003) [41] in terms of organizational factors, especially on the leadership side and their impact on the information overload created especially with those working in the field of knowledge
- The results in the level of importance of the regulatory overload in relation to the channels of communication in the Jordanian industrial companies coincided with the results of the analysis of the regulatory environment. (Manovas, 2004) [42], particularly in the field of knowledge transfer, learning culture, sharing, and incentive systems as elements of infrastructure in the regulatory environment
- The results of the study show that there is a high level of interest in knowledge generation due to the ability of managers to diversity knowledge sources and their focus on research and development and bridging knowledge gaps as a result of developments in the work environment and attention to the internal organizational dimension in the generation of knowledge through

brainstorming processes. This result is consistent with the findings of the Zakia, 2009, [43] that studied on knowledge sources, acquisition and transmission. The results were also consistent with the Zainab, 2009, [37] which examined the dimensions and processes of knowledge management (acquisition, generation, transmission, distribution, and application) at King Abdulaziz University

- The results of the study demonstrated the impact of the regulatory overload (channels of communication and regulatory environment) on knowledge management functions (acquisition, generation, transmission, sharing, and application of knowledge) in Jordanian industrial companies from the point of view of managers working in Jordanian industrial companies. The results of the current study are consistent with the (Carlevale, 2010) [25] study, as reliance on communications technology creates a burden, especially e-mail, which creates a burden for managers and that e-mail needs to be filtered. This was also agreed with (Dubosson and Fragniere, 2009) [27] and (Lesa, 2009) [23].

11. RECOMMENDATIONS

Regarding to the results reached through the research, the researchers provide a number of recommendations to adopt them by the Jordanian industrial companies in the course of the research, so as to adapt them in reducing the burden of electronic information in term of regulatory overload and these recommendations as follows:

- Adopting the type of technology appropriate to the environment in which Jordanian industrial companies operate in such a way as to reduce the burden of information in term of regulatory overload that may be exposed in carrying out their decision-making tasks
- The Jordanian industrial companies should conduct a strategic analysis of the strengths and weaknesses that are reflected in the performance of the company's departments and departments and determine their impact on increasing or decreasing the regulatory overload in the departments
- Developing companies in their regulatory environment to achieve effective communication systems based on the concept of reducing the regulatory overload in an attempt to restructure their systems to achieve their effectiveness
- Companies continue to filter and exclude unnecessary information in a way that reduces the large amount of information that restricts the capabilities of the employees of the initiative and this does not negatively affect the capabilities of the public in providing initiatives in the field of electronic work, and not affected by the capabilities of employees in solving electronic problems.

REFERENCES

[1] S. N. Jawad, A. A. M. Shaban, H. H. Ali and I. Husen. "*Small Business Management, a Technology Entrepreneurial Perspective*". SAFA Publishing House, Amman, Jordan, 2010.

[2] National Information Technology Council (NITC). "*Malaysia*, (*K-Economy-Introduction and Background*)". 2004. Available from: http://www.nitc.org. [Last accessed on 2017 Dec 15].

[3] M. Song, H. Bij and M. Weggeman. "Factors for improving the level of knowledge generation in new product". *R & D Management*, vol. 36, no. 2, pp. 173-187, 2006. Available from: http://www.ssrn.com. [Last accessed on 2017 Dec 15].

[4] T. Asmahan and M. Ibrahim. "*Requirements for Sharing Knowledge and Obstacles Facing its Application in Jordanian Telecommunication Companies, Presented to the Scientific Conference*". Applied Science University, Amman, Jordan, 2007.

[5] A. A. M. Shaban and J. S. Naji. "*Management Process and Information Technology*". Al-Ethaa Publishing House, Amman, Jordan, 2008.

[6] A. A. M. Shaban and J. S. Naji. "*Business Intelligence and Information Technology*". Amman, Jordan, Safa Publishing House, 2012.

[7]. A. M. Sami. "*Measuring the Impact of Organizational Culture Factors on the Implementation of Knowledge Management in the Jordan Telecom Group (Orange): Case Study, Unpublished Master Thesis, Graduate School of Administrative and Financial Studies*". Amman, Jordan, Amman Arab University for Graduate Studies, 2008.

[8] N. Abboud. "*Knowledge Management Concepts, Strategies and Operations*". Dar Al Warraq, Amman, Jordan, 2005.

[9] U. Sekaran. "*Research Methods for Business*". 4th ed. John Wiley & Sons, Ltd., New York, United States, 2003.

[10] M. Alavi and D. Leidner. "Review: Knowledge management and knowledge management systems: Conceptual foundation and research issues". *MIS Quarterly*, vol. 25, no. 1, p. 107-136, 2001. Available from: http://www.ebsco.host.com. [Last accessed on 2017 Dec 15].

[11] Z. Mohammed. "*Contemporary Trends in Knowledge Management*". SAFAA Publishing and Distribution House, Amman, Jordan, 2008.

[12] M. Abbas. "Knowledge management and its effect on organizational innovation". *Journal of Arts Kufa*, vol. 1, p. 257, 2008.

[13] D. Bawden and L. Robinson. "The dark side of information: Overload, anxiety, and other paradoxes and pathologies". *Journal of Information Science*, vol. 35, no. 2, pp. 180-191, 2008.

[14] L, Ruff. "*Information Overload: Causes, Symptoms and Solution, Harvard Graduate School of Education's Learning Innovations Laboratory*, (*LILA*)". 2002. Available from: http://www.lila.pz.harvard.edu/_upload/lib/infooverloadbrief.pdf. [Last accessed on 2017 Dec 20].

[15] M. Eppler and J. Mengis. "*A Framework for Information Overload Research in Organizations: Insights from Organization Science, Accounting, Marketing, MIS, and Related Disciplines, ICA Working Paper*". University of Lugano, Lugano, 2003. Available from: http://www.bul.unisi.ch/cerca/bul/pubblicazioni/com/pdf/wpca0301.pdf.

[Last accessed on 2018 Jan 10].

[16] E. Whelan and R. Teigland. "*Managing Information Overload: Examining the Role of the Human Filter*". 2010. Available from: http://www.ssrn.com. [Last accessed on 2018 Jan 10].

[17] K. E. Himma. "*A preliminary step in understanding the nature of a harmful information-related condition: An analysis of the concept of information overload*". *Ethics and Information Technology*, vol. 9, no. 4, pp. 259-272, 2007.

[18] M. L. Grise and B. Gallupe. "Information overload: Addressing the productivity paradox in face-to-face electronic meetings". *Journal of Management Information Systems*, vol. 16, no. 3, pp. 157-186, 2000.

[19] I. Mulder, H. de Poot, C. Verwij, R. Janssen and M. Bijlsma. "*An information Overload Study: Using Design Methods for Understanding, Conference on Computer-Human Interaction: Design: Activities, Artefacts and Environments, Sydney, Australia*". pp. 245-252, 2006.

[20] K. Kim, M. Lustria and D. Burke. "*Predictors of Cancer Information Overload: Findings from a National Survey*". 2007. Available from: http://www.informationr.net/ir/12-4/paper326.html#mil56. [Last accessed on 2018 Jan 10].

[21] A. Farhoomand and D. Drury. "Managerial information overload". *Communications of the ACM*, vol. 45, no. 10, pp. 127-131, 2002.

[22] M. Dubosson and E. Fragniere. "The consequences of information overload in knowledge based service economies: An empirical research conducted in Geneva". *Service Science*, vol. 1, no. 1, pp. 56-62, 2009.

[23] B. Lesa. "*The Impact of Organizational Information Overload on Leaders: Making Knowledge Work Productive in the 21st Century*, *Doctoral Dissertation*". University of Idaho, United States, 2009.

[24] T. Wilson. "Information overload: Implications for healthcare services". *Health Informatics Journal*, vol. 7, no. 2, pp. 112- 117, 2001.

[25] S. Filippov and K. Iastrebova. "Managing information overload: Organizational perspective". *Journal on Innovation and Sustainability*, vol. 1, no. 1, pp. 1-17, 2010. Available from: http://www.revistas.pucsp.br/index.php/risus/article/view/4260. [Last accessed on 2018 Jan 10].

[26] X. Yan. "*An Empirical Analysis of the Antecedents of Knowledge Management Strategies, Doctoral Dissertation*". Nova Southeastern University, United States, 2009.

[27] A. Hester. "*Analysis of Factors Influencing Adoption and Usage of Knowledge Management Systems and Investigation of Wiki Technology as an Innovative Alternative to Traditional Systems, Doctoral Dissertation*". University of Colorado Denver, United States, 2009.

[28] H. Haytham. "*Measuring the Impact of Knowledge Management Perception on Employment in Jordanian Organizations: A Comparative Analytical Study between the Public and Private Sectors towards Building a Model for Knowledge Management Employment*, *Unpublished Doctoral Thesis, Faculty of Administrative and Financial Studies*". Amman Arab University for Graduate Studies, Amman, Jordan, 2005.

[29] H. Najm. "*Management Information Systems: Contemporary Entrance*". Wael Publishing House, Amman, Jordan, 2005.

[30] A. S. Salwa. "*The Role of Knowledge Management and Information Technology in Achieving Competitive Advantages in Banks Operating in Gaza Strip*". Master of Business Administration, Islamic University, Gaza, 2008.

[31] A. Jazar and A. Talaat. "*Proposed Project for Knowledge Management in Jordanian Public Universities*, *Unpublished Doctoral Thesis, Faculty of Higher Education Studies*". Amman Arab University for Graduate Studies, Amman, Jordan, 2005.

[32] I. Fernandez, A. Gonzalez and R. Sabherwal. "*Knowledge Management, Challenges, Solution, and Technologies*". 1st ed. Pearson Prentice Hall, London, United Kingdom, 2004.

[33] B. Mohammed and M. Ziad. "*Knowledge Management between Theory and Practice*". Jalis Al-Zaman Publishing House, Amman, 2010.

[34] J. Ashoc. "*Knowledge Management an Integrated Approach*". Pearson Education, Prentice-Hall, London, United Kingdom, 2004.

[35] M. Kamel. "*Knowledge Acquisition, Wiley Encyclopedia of Electrical and Electronics Engineering*". John Wiley and Sons, Inc., New York, United States, 1999.

[36] X. Zhang. "*Understanding Conceptual Framework of Knowledge Management in Government* (*Condensed Version*)*, Presentation on UN Capacity-Building Workshop on Back Office Management for e/m-Government in Asia and the Pacific Region, Shanghai, China*", 2008.

[37] S. Zainab. "*The Readiness of Public Organizations for Knowledge Management: A Case Study of King Abdul Aziz University in Jeddah, An Introduction to the International Conference on Administrative Development: Towards Distinguished Performance of the Government Sector, Riyadh*", 2009.

[38] E. Carlevale. "*Exploring the Influence of Information Overload on Middle Management Decision Making in Organizations, Doctoral Dissertation*". University of Phoenix, United States, 2010.

[39] J. Hodge. "*Examining Knowledge Management Capability: Verifying Knowledge Process Factors and Areas in an Educational Organization, Doctoral Dissertation*". Northcentral University, United States, 2010.

[40] M. Ismail and Z. Yusof. "The impact of individual factors on knowledge sharing quality". *Journal of Organizational Knowledge Management*, vol. 2010, p. 327569, 2010.

[41] M. Raoufi. "*Avoiding Information Overload-A Study on Individual's Use of Communication Tools, Proceeding of the 36th Hawaii International Conference on System Sciences*". 2003.

[42] M. Manovas. "*Investigating the Relationship between Knowledge Management Capability and Knowledge Transfer Success, Mastery Degree*". Concordia University, Canada, 2004.

[43] T. Zakia. "*Knowledge Management: The Importance and Extent of Application of its Operations from the Point of View of the Supervisors and Administrator's Departments of the Department of Education in Makkah and Jeddah, Master Thesis, Umm Al-Qura University*". 2009.

ORIGINAL RESEARCH ARTICLE UHD JOURNAL OF SCIENCE AND TECHNOLOGY

Characterization of European Medieval Silver Bars Using Micro X-ray Fluorescence, Conductivity Meter and Scanning Electron Microscopy

Ahmad Nizamedien Barzingi

Department of Chemistry, College of Education, University of Garmian, University in Kalar, Iraq

ABSTRACT

The objective of this paper is to use μ-X-ray fluorescence (XRF) analysis to evaluate the fineness and components of European Medieval Silver Bars samples. Conductivity measurements were used to assess the fineness and localization of the faults found in the samples. Because unevenness causes a change in conductivity, the tests were performed on the flattest areas of the Bars. Some rods, such as B3 and B9, have greater conductivity than others. All bars were subjected to the segregation test. In the instance of certain bars, it was not always practicable to categorically state that segregation had happened. There is no diminishing conductivity curve as one moves away from the zero height, as there is for bars B1, B8, and B9. As a result, there may be no solidification on these bars from Obverse to Reverse. A scanning electron microscope was used to record the following bars at various positions on the bars, and quantitative determinations were achieved using energy-dispersed XRF analysis through intensity measurements of the element-specific wavelength.

Index Terms: Silver bars, X-ray fluorescence, Micro-X-ray fluorescence, Conductivity Meter, Scanning electron microscope

1. INTRODUCTION

Money has been used in Europe for over 2600 years to denote coined metal, particularly in the form of coins. When the term money (derived from the Middle High German "Geld") is reduced to its most fundamental meanings (replacement, compensation, value, price, retribution), all that remains is a widely accepted standard of billing, value storage, and pricing that has developed from barter. This standard is usually based on a certain amount of a coveted and durable material, often a metal, the most important raw material of antiquity. The importance of metals is demonstrated by the fact that entire periods of history are called after them: the Copper Age (end of the Neolithic to before the 3rd century BC), the Bronze Age (before 2200 BC in Europe), and the Iron Age (before 1200 BC in Europe) (in Central Europe before 1200 BC to after 500 AD) [1]. Hence, it is not unexpected that a metal bar has been used for almost 4000 years, since the Bronze Age, as one of the original and natural forms of money (in addition to non-metallic types of money) (Fig. 1) [2].

The oldest bronze casting sites in Central Europe and Germany that have been verified to make bar bars in stone molds are in Saxony-Rotta Anhalt's and Schackstedt (evidence is available in the State Museum of Saxony-Anhalt in Halle). Since the early Bronze Age, metal bars of varying forms and weights have been used in this manner to preserve payment and value (at least verifiable from approx. 2700 BC) [4] and are also used today in many international transactions (e.g. the

Access this article online

DOI:10.21928/uhdjst.v5n2y2021.pp57-65 | E-ISSN: 2521-4217 P-ISSN: 2521-4209

Corresponding author's e-mail: Ahmad Nizamedien Barzingi, Department of Chemistry, College of Education, University of Garmian, University in Kalar, Iraq. E-mail: Ahmad.barzingi@garmian.edu.krd

Received: **23-09-2021 Accepted: 20-11-2021 Published: 25-11-2021**

method of choice for repaying national debts by switching gold bars in the Federal Reserve Bank of New York [5].

Bullion's journey can be traced throughout human history and to every corner of the globe. The ancient Egyptians utilized bar money, which was made out of precious metal rings and pieces imprinted with the Pharaoh's name. Special officials were assigned to oversee the quality and weight of silver bars in Mesopotamia during the period of Babylonian King Hammurabi (1792–1750 BC) [6]. Since silver has been in Mesopotamia since 2100 BC was the measure of value for goods, wages and prices are already in the 18th century BC a standard house for silver as a means of payment in Babylon, probably the earliest documented forerunner of DIN, known [7].

For commercial reasons, the Celts and Teutons in Northern Europe utilized standardized bronze rings and high-quality iron bars [4].

The Romans, together with the Greeks, are regarded as the forefathers of European civilization, and they utilized metal ingots as currency until approximately 200 BC [8]. The actual currency was not used until the 7th century BC, when it was developed (Fig. 2), most likely by Lydian king Alyattes II, the father of the legendary Croesus [9].

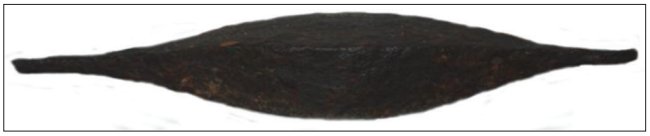

Fig. 1. Celtic pointed bars made of iron, approx. 5 kg, around 100 BC. Chr. To 100 AD. Such high-quality iron bars from the Celtic and Germanic tribes were a sought-after commodity and were mainly exported to Rome. Among other things, high-quality Roman swords were forged from it [3].

Fig. 2. Electron trite (third stater) of the Lydian king Alyattes II (approx. 613-556 BC), 4.71 g, 12 mm, enlarged illustration [3]. Alyattes was the father of Croesus and is considered the inventor of minted money [9].

Small electron bars (from Latin "electrum" = amber, signifying a natural mixture of gold and silver) were used to create the currency[10]. A seal (lion's head as the Mermnaden dynasty's coat of arms) was imprinted on the first coin's hour of birth to indicate its origin and ensure its value]. Coincidentally, about the same period, the first counterfeit coins appeared. These are lead bars that have had an electron coat applied to them (Herodotus 5th century BC). To confirm the coins' validity, they had to be sliced apart or given a deep notch, a procedure that left traces on several ancient coins that may still be seen today.

Gold and silver bars were used to replace money in Rome until the fall of the Western Roman Empire (AD 474), despite the highly established Roman currency system (Figs 4-6).

From approximately 300 AD, the Roman emperor gave legionaries a 5-year present (donative) consisting of five gold coins (aurei or solidi, Fig. 3) and a silver bar weighing one Roman pound (327, 63 g) [11].

A silver bar of this size was valued at around 200 L of wine, two annual rations of food, or about a third of the typical

Fig. 3. Solidus of the Roman Emperor Constantinus (306–337 AD), 4.56 g, 20 mm, minted around 310–313 in Trier, Fig. Enlarged (Lehmann 2010b) [3]. Flavius Valerius Constantinus, also known as Constantine the Great, introduced the solidus in 309 AD, which replaced the aureus that ran before it as a Roman imperial gold coin [10].

Fig. 4. Roman gold bar from the Munich State Coin Collection. This bar probably only represents a form of transport of the gold to the mint. The gold bar was once owned by the Swiss Federal Bank, which determines the fineness and - regardless of the art-historical value of the bar - had this stamped on the bar next to an inventory number. The fineness is 99.15%. The bar shape has been a popular shape for non-ferrous and precious metals since the Bronze Age because it made it easier to cut the bar (Photo: Lehmann).

Fig. 5. Overview of silver bars from different epochs. On the left is a Roman silver bar (target weight: 327.63 g), 7 × 10 cm), in the middle a medieval German bar (one silver mark nominally 233.856 g, 6–7 cm), on the right modern silver bar (1 kg). While you could still buy a slave for around 3 Roman bars and a small piece of land for the medieval bar, at the end of 2010 300 g of fine silver cost just under 200 euros [12].

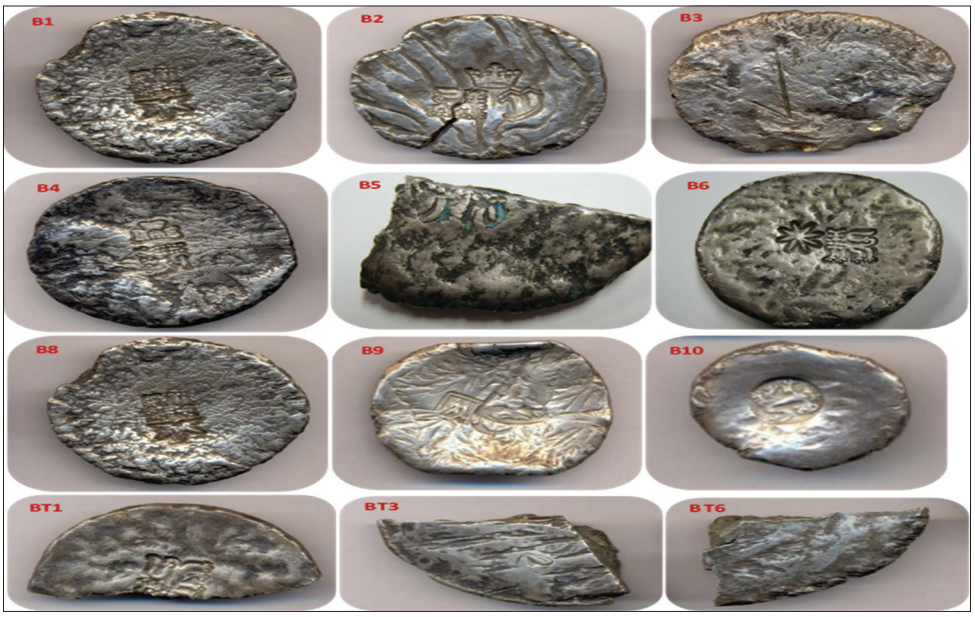

Fig. 6. Silver bars samples.

slave price [13]. Antiquity and the Middle Ages silver bars had a radically different form and shape from today's bar silver, as seen in (Fig. 5). The double ax or skin form is seen on the Roman silver bars illustrated here [8], although most items from the German Middle Ages were shaped like a hemisphere or dome [14].

In general, X-ray fluorescence (XRF) is a non-destructive technique for the qualitative and quantitative evaluation of inorganic solids, liquids, and burned organic compounds that are based on the interaction of X-rays with matter. Cement, glass, and ceramic industries consume rocks and slag. It has become an established routine method for archaeometric investigations on precious metal objects (coins, jewelery, etc.) and other historical objects for several years, especially since the development of transportable devices in archaeology, where it is used to quickly classify found objects and has been an established routine method for archaeometric investigations on precious metal objects (coins, jewelery, etc.). This approach is also appropriate for multi-elemental analysis of inorganic components in an organic matrix, for example to identify inorganic impurities in biological samples, only elements with an atomic number $Z > 13$ (carbon) may typically be examined [15]. More recent developments also allow the analysis of aerosols and gases.

The X-ray radiation used in XRF is electromagnetic radiation with a wavelength in the range of 0.01–10 nm and, together with the γ-rays, forms the short-wave limit range of the electromagnetic spectrum [16].

Electron microscopy enlarges pictures of things using electron beams. It is utilized for micrometer-scale surface analysis. When compared to a light microscope, accelerated electrons may attain a resolution a 1000 times >3 nm, allowing finer features to be distinguished. The device is composed of four assemblies: an electron gun, electron optics, a sample holder, and an electron detector.

When an electron beam collides with a sample surface, two types of interactions occur elastic interactions, which change the direction of the electrons without altering their energy, and inelastic interactions, which transfer the electron energies partially or entirely to the sample atoms.

Secondary electrons, Auger electrons, X-rays, heat, and light are then emitted by the excited material (electromagnetic radiation over many frequencies). Secondary electrons can be utilized to generate images [17].

The collision of an electron with an atom changes the direction of the electron but the speed and therefore the kinetic energy stays unchanged in elastic scattering. Some electrons lose their energy due to inelastic collisions and remain in the material after numerous collisions.

The majority of the energy is emitted from the surface as backscattered electrons. Secondary electrons, which are particularly crucial for image production in the scanning electron microscope (SEM), are produced in the condition of inelastic scattering. These are produced by the interaction of high-energy electrons in the beam with weakly bound electrons in the solid's conduction band, which results in the release of conduction electrons. Secondary electrons are created across the beam's interaction region with the material, but because of their low energy, they are quickly absorbed again. Only secondary electrons produced near the sample's surface can escape. Small elevations on the sample surface result in a shorter electron route length than level regions, allowing more secondary electrons to escape.

Electrical conductivity is a critical material characteristic that not only tells us how effectively a metal conducts electrical current, but also allows us to make judgments about its composition, microstructure, and mechanical capabilities.

The measurement of conductivity is based on the principle of measuring resistance. An electric current is produced in the material to be studied using a probe for this purpose, and the electrical resistance of the substance is measured. The reciprocal value of the electrical resistance determines conductivity.

$$\gamma = \frac{1}{\rho} = \frac{G}{s} = \frac{1}{R \cdot s}$$

γ: Electrical conductivity/conductivity (Sm^{-1})
ρ: The specific electrical resistance (Ω m)
G: The electrical conductance, measured in Siemens (S)
s: S: The length of Specimen measured in meters (m),
R: The resistance, measured in ohms (Ω)

2. MATERIALS AND METHODS

2.1. Conductivity Meter

The conductivity measurements were performed using a "Hocking" "AutoSigma 2000" instrument. The probe has an interior diameter of 11.5 mm and an exterior diameter of 12.5 mm, and it operates at a frequency of 60 Hz.

Conductivity measurements were used to attempt to identify the fineness and position of the flaws on the sample. The stability of the device display was verified before each measurement. A copper standard was measured 3 times before and after point measurements, and a copper standard was measured 3 times beneath plastic film, to evaluate device drift (Table 1).

2.2. SEM

XL 30, JEOL JSM-6700F Manufacturer: Philips Electronics was used for the recordings, and this has an acceleration voltage of 0.5–30 kV. The device has a secondary

TABLE 1: Stability of the measuring device

Bar money	Conductivity S/m*10⁶ Average 1	Conductivity S/m*10⁶ Average 2	Conductivity S/m*10⁶ Average - Foil 1	Conductivity S/m*10⁶ Average -Foil 2
B8 (Obvers)	58.6	58.567	58.3	58.4
B8 (Revers)	58.63	58.63	58.3	58.367
B2 (Obvers)	58.7	58.71	58.367	58.367
B2 (Reverse)	58.667	58.7	58.33	58.366

electron detector, backscattered electron detector, Si (Li) semiconductor detector as a detector.

2.3. XRF Spectroscopy

The measurements with μ-XRF analysis aimed to determine the fineness and the components of the bars B1, B2, B3, B4, B6, B8, B9, B10, BT1, BT3, BT6 (Table 2).

The "Eagle μ-Probe II" device from EDAX was used for the measurement. The device has an X-ray tube with a beryllium window, rhodium target, an acceleration voltage of 10–40 keV, a cathode current of 40–1000 μA and a Si (Li) detector was used as a detector.

3. RESULTS AND DISCUSSION

3.1. Conductivity Measurement

The conductivity of the following bars was measured: B1, B2, B3, B4, B5, B6, B8, B9, B10, BT1, BT3, BT6. (Fig. 6). Because unevenness affects conductivity, the tests were taken on the flattest areas of the bars. On each sample, 10 points on the front and 10 points on the reverse were chosen. Because of unbalancing, several samples could not be measured to 10 points. The conductivity was measured on all 10 silver bars. The detailed measurement results can be found in Table 3. The values fluctuate within the error limits specified by the device manufacturer of ± 0.1 MS/m at room temperature.

Fig. 7 shows wave-like fluctuations in the conductivity of the bars. There are some bars such as B3 and B9; these bars have a higher conductivity than the rest. This is perhaps due to the silver content of the bars.

The average conductivity of the obverse is 22.05 ± 1.70 and the reverse 16.07 ± 2.47 MS/m. As can be seen from the results, avers has a higher conductivity than the reverse. That means the obverse has a higher fineness than the reverse. This is proof that the bars slowly solidified from the obverse so that segregation has occurred (Fig. 8).

There is a 5.98 MS/m average conductivity difference between the front and back. This discrepancy might be

TABLE 2: Silver bars description in term of weight and dimension

No.	Weight (g)	Dimension (cm)
B1	191.4	5.9×1.2
B2	187.5	6×1.1
B3	246.3150	6.3×1.09
B4	249.149	6.7×1.2
B5	187.232	
B6	197.2456	5.8×1.3
B8	211.790	6.1×1.1
B9	132.0670	6.4×1.2
B10	128.876	5.3×1.08
BT1	71.165	
BT3	64.317	
BT6	62.835	

TABLE 3: Conductivity of the bars with standard deviation (σ)

Conductivity (S/m *10⁶)					
No.	**Bars**	**Obvers**	**Revers**	**σ (Obvers)**	**σ (Revers)**
1	B1	16.05	10.41	2.45	1.98
2	B2	13.16	10.07	2.85	3.15
3	B3	40.51	33.95	3.25	1.9
4	B4	16.14	11.73	1.47	1.67
5	B5	20.03	12.69	0.75	1.84
6	B6	20.34	14.63	1.27	2.45
7	B8	15.46	12.5	1.06	3.39
8	B9	23.87	18.98	1.13	2.87
9	B10	36.96	24.76	5.19	0.40
10	BT1	20.40	13.89	0.53	2.23
11	BT3	20.06	11.27	0.29	2.37
12	BT6	21.65	18.06	0.19	5.41

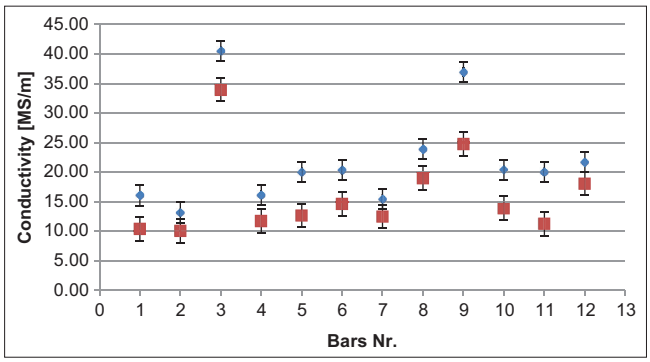

Fig. 7. Average conductivities on the obverse and lapel. ◆ Blue conductivity of the obverse (MS/m), ■ Red Conductivity of the revers (MS/m).

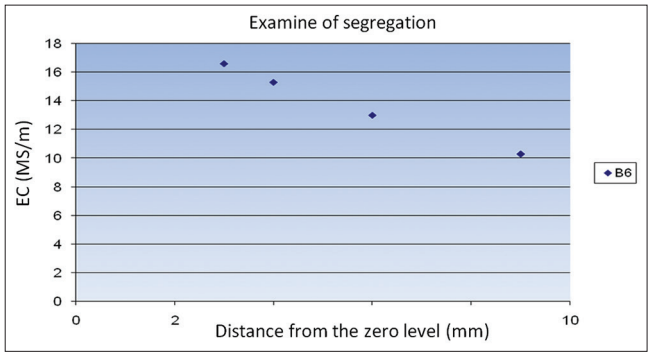

Fig. 8. Conductivity curve on the bar 6.

produced by a variety of factors, including device limitations, corrosion layers, and fineness changes from front to back, such as more silver on the front than the opposite.

The segregation of the bars was investigated. The technique of separating distinct items in an observation area is referred to as segregation. A zero level was established on the bars for this reason. The transition from obverse to reverse was specified at the zero levels. Segregation on the bars can be noticed if the conductivity diminishes proportionately to the distance from the zero levels.

All bars were subjected to the segregation test. It was not feasible to prove that segregation had occurred in the case of certain bears. There is no decreasing conductivity curve as one moves away from the zero height, as there is for bars B1, B8, and B9. This might result in no solidification on these bars from the Avers to the Revers.

Influence of the stamps on the conductivity (EC);

The influence of the stamps on the conductivity was investigated. According to the theory, the stamping should result in lower conductivity. The results are shown in Table 4.

According to theory, stamping should result in lower EC, but this difference does not seem to be significant, that is, on the Revers, segregation is probably the main cause of the various EC.

3.2. Measurements by the SEM

The goal is to learn about the equipment and how to utilize SEM, as well as to learn about the surface characteristics of the bars. A SEM was used to record the following bars at various positions on the bars, and quantitative determinations were achieved using energy-dispersed X-ray fluorescence analysis (EDX) through intensity measurements element-specific wavelength (Fig. 9).

For image B3-05, EDX was made on a wise matrix, on points and dark points (Fig. 10).

The results are given in the table below.

In Table 5, it is clear that the dark points are predominantly made of carbon. The Cl content determined whether the salts have formed on the surface of the bar, but it cannot be clearly stated that the salts have formed because the Cl content is too small.

With the SEM-EDX, light elements such as carbon and oxygen were detected compared to the μ-XRF

Image B5-01 was taken to determine whether the drawings can be recognized as writing. The drawings suggest the scriptures to some extent (Fig. 11).

An attempt was made in image B5-08 to clarify whether or not it is a "M" letter. However, it is not as obvious (Fig. 12).

3.3. XRF Analysis

The measurements with μ-XRF analysis aimed to determine the fineness and the components of the bars B1, B2, B3, B4, B6, B8, B9, B10, BT1, BT3, BT6.

The "Eagle μ-Probe II" device from EDAX was used for the measurement. The device has an X-ray tube with a beryllium window, rhodium target, an acceleration voltage of 10 to 40 keV, a cathode current of 40 to 1000 μA and a Si (Li) detector was used as a detector.

For each bar, 20 points were placed on the obverse and 20 points on the reverse, resulting in a generally circular form. Table 6 shows the corrected findings after adding the standards to the basic parameter correction. Silver

TABLE 4: Conductivity values with and without stamps, with the picture punch element-specific wavelength. B3 bar with different photos

Bars	EC without stamps	EC with stamp1	EC with stamp 2	EC with stamp 3	Picture punch 1	Picture punch 2	Picture punch 1
B6	20.8	21.8	19.9		Star	Lion	
B8	13.7	16.3	16.9		Head	Lion	
B3	36.4	32.3			Sign		
B2	14.1	16.4	14.4	7.9	Crown	D.	Lion (crack)
B9	24.8	24.1	24.6	24.6	Flower	W.	?
B1	14.3	17.6	18.8		Lion	Star	
B4	16.4	18.6	18.5		Lion	Flowe	
BT1	20.6	19.9			Lion		
B10	33.8	34.8			Lion		
B5	21.4	18.8			?		
BT3	20.3	19.3			Pikas		

TABLE 5: The chemical composition of B3-05

Elements	Wise Matrix	Point	Dark structure
C %	6,5	44	38
O %	1.2	8	8
Cl %	0.6	0.1	2
Ag %	88.6	4	45
Cu %	0.7	36	1
Pb %	2.3	6	5

was found as the major component in all twelve bars, with copper and lead as minor components. The bars contain a silver composition of 6.49 4.07 percent by weight copper and 6.38 3.28 percent by weight lead, with an average of 87.63 5.86 percent by weight. There are extremely high standard variations with an average copper and lead concentration. This demonstrates how the copper and lead contents of each ingot vary greatly. The silver content has

TABLE 6: Measuring the weight percentage of silver bar samples by μ-XRF of both sides

Sample ID	Side	Obverse	AgK	AgL	Cu	Pb
B1	Obv.	Average	85.944	87.013	7.779	6.278
		STD	3.179	3.057	2.451	1.981
	Rev.	Average	77.239	78.268	10.103	12.659
		STD.	2.269	6.135	1.471	3.091
B2	Obv.	Average	81.678	89.595	13.085	5.237
		STD	6.739	2.613	4.287	4.218
	Rev.	Average	89.922	88.340	4.295	5.737
		STD	3.154	3.997	1.824	2.135
B3	Obv.	Average	89.786	94.151	5.637	3.635
		STD	4.201	2.342	2.509	2.482
	Rev.	Average	96.593	96.264	2.014	1.395
		STD	1.829	2.042	0.639	1.742
B4	Obv.	Average	82.717	89.601	8.261	9.024
		STD	4.854	3.715	1.481	5.440
	Rev.	Average	85.966	85.995	4.746	9.289
		STD	4.507	3.457	1.430	4.324
B5	Obv.	Average	87.367	88.984	4.289	5.345
		STD	3.321	3.324	1.547	2.768
	Rev.	Average	88.462	89.664	5.915	5.622
		STD	3.718	3.177	1.585	2.853
B6	Obv.	Average	81.312	88.461	9.256	9.331
		STD	4.501	3.226	1.987	3.286
	Rev.	Average	91.031	92.098	5.110	3.853
		STD	4.520	3.465	1.537	3.085
B8	Obv.	Average	87.359	90.729	6.040	6.601
		STD	6.388	3.862	3.734	4.317
	Rev.	Average	91.333	90.635	2.875	5.793
		STD	2.747	4.778	0.936	2.864
B9	Obv.	Average	94.835	95.275	1.867	3.297
		STD	1.210	0.765	0.352	1.163
	Rev.	Average	93.847	93.689	2.809	3.316
		STD	3.089	3.438	0.379	3.056
B10	Obv.	Average	96.265	91.511	1.557	2.179
		STD	1.657	2.534	0.555	1.195
	Rev.	Average	94.539	94.794	2.404	3.057
		STD	3.129	4.005	0.665	3.452
BT1	Obv.	Average	90.301	89.729	5.359	5.495
		STD	4.167	4.914	3.538	3.356
	Rev.	Average	90.430	82.873	4.869	5.273
		STD	4.351	28.147	1.510	3.958
BT3	Obv.	Average	75.150	77.330	13.250	11.610
		STD	8.160	5.540	7.690	3.590
	Rev.	Average	69.948	83.247	15.648	13.300
		STD	4.708	3.152	3.581	3.668
BT6	Obv.	Average	89.701	89.433	4.892	5.408
		STD	4.233	4.187	2.112	2.533
	Rev.	Average	75.963	88.397	13.644	10.393
		STD	10.026	3.254	4.820	7.771

the highest B9 bar, with an approximate value of the copper and lead content have the maximum BT3 bar, with around 95.14 percent and 12 percent, respectively.

The iron and gold composition of the bar could be determined. That was not accomplished in our effort. Table 7 shows the ratio of the intensities of the generated Ag-K and Ag-L signals to determine the thickness of the corrosion layer.

According to the AgK/AgL ratios, the obverse of all Roman bars has a slightly thicker corrosion layer than the reverse. This might be owing to a site-specific storage facility in the earth or other unique features.

4. CONCLUSION

The conductivity of the bars is shown by wave-like variations. The conductivity of the B3 and B9 is greater than that of the others. This is because the bars contain silver.

The average conductivity of the obverse is 22.05 ± 1.70 and the reverse 16.07 ± 2.47 MS/m. Obvers has a higher conductivity than the reverse. That means obverse has a higher fineness than revers. This is proof that the bars have slowly solidified from the obverse so that segregation has occurred.

TABLE 7: AgK/AgL ratio of difference silver bars

	*AgK/AgL Ratio					
	B1	**B2**	**B3**	**B4**	**B5**	**B6**
Obvers	0.088	0.081	0.075	0.092	0.098	0.091
Revers	0.098	0.090	0.088	0.097	0.099	0.094
	B8	**B9**	**B10**	**BT1**	**BT3**	**BT6**
Obvers	0.076	0.074	0.085	0.096	0.077	0.071
Reverse	0.094	0.079	0.099	0.010	0.084	0.085

The peak intensities of AgK and AgL line are measured as cps/eV[18]

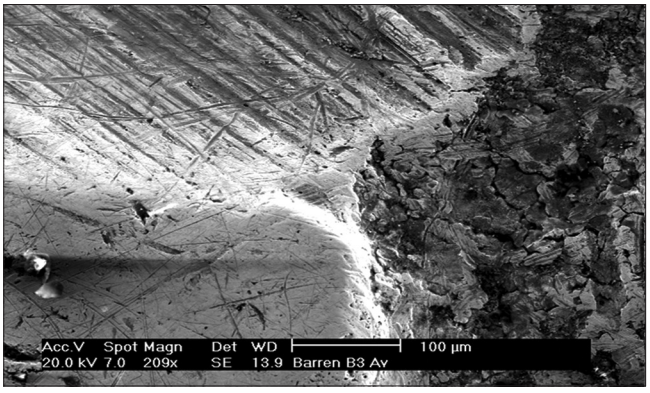

Fig. 9. Image B3-01: Photo at the point where is the casting-demolition transition.

All bars were subjected to the segregation test. In the instance of certain bars, it was not always feasible to state unequivocally that segregation had happened. There is no diminishing conductivity curve as one moves away from the zero height, as there is for bars B1, B8, and B9. As a result, there may be no solidification on these bars from obvers to reveres.

A SEM was used to record the following bars at various positions on the bars, and quantitative determinations were achieved using EDX through intensity measurements of the element-specific wavelength.

The measurements with μ-XRF analysis aimed to determine the fineness and the components of the bars B1, B2, B3, B4, B6, B8, B9, B10, BT1, BT3, BT6.

For each bar, 20 points were placed on the obverse and 20 points on the reveres, resulting in a generally circular form. In all twelve bars, silver was identified as the main component and copper and lead as the minor component. With an average of 87.63 ± 5.86% by weight, the bars have a silver content of 6.49 ± 4.07% by weight copper and 6.38 ± 3.28% by weight lead. With an average copper and lead content, there are very large standard deviations.

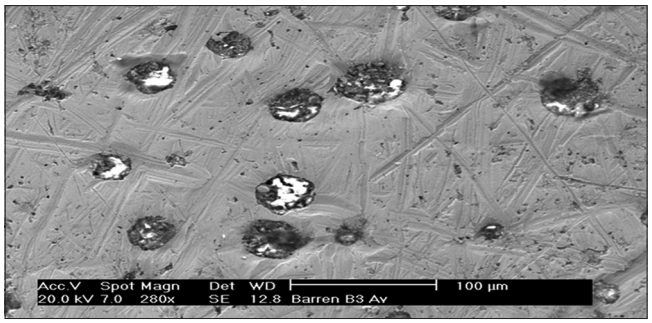

Fig. 10. Image B3-05: The obverse image of round inclusions [18].

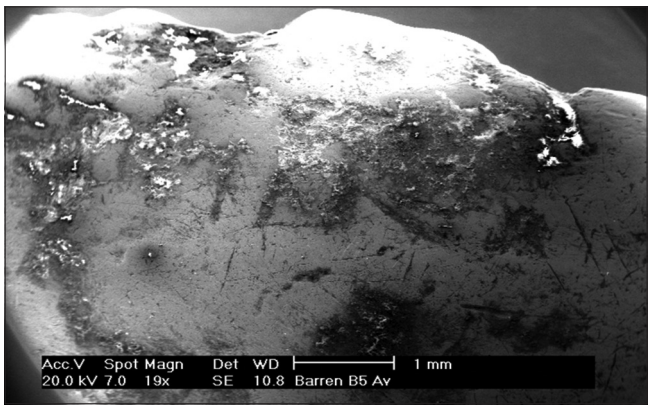

Fig. 11. Image B5-01: Photo of bar B5.

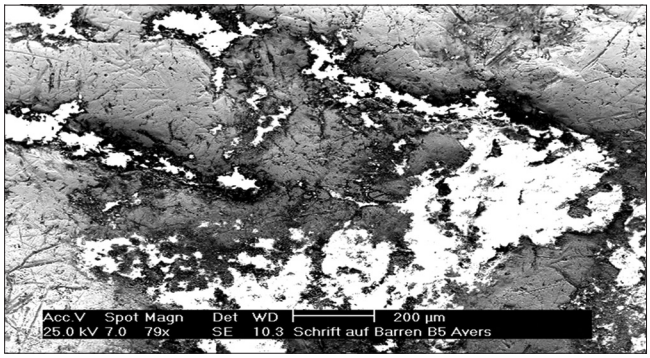

Fig. 12. Image B5-08: Writing on obverse B5 bar.

This shows that the copper and lead contents of each ingot vary widely. The silver content has the highest B9 bar with approx. 95%, the copper and lead content have the highest BT3 bar with approx. 14% and 12%. The iron and gold composition of the bar may be analyzed. That was not done in our work. According to the AgK/AgL ratios, the reverse of all silver bars (B1, B2, B3, B4, B5, and B6) has a somewhat thicker corrosion layer than the face. This might be owing to a location-specific storage facility in the earth or other particular properties.

REFERENCES

[1] H. J. Eggers. "Zur Absoluten Chronologie der römischen Kaiserzeit im Freien Germanien". In: *Aufstieg u. Niedergang röm. Welt II*, pp. 3-6, 1976.

[2] M. Pfefferkorn. "*Seit 4000 Jahren Zahlungsmittel, Vorstufen des Münzgeldes-Vom Bronzebarren Zum Goldenen Oban, Moneytrend*", vol. 9, pp. 118-122, 2003.

[3] R. Lehmann. *"Barren versus Münze, Teil I: Barrengeld Damals Und Heute Und Die Besondere Bedeutung im Mittelalter, Moneytrend*", vol. 10, pp. 134-139, 2010b.

[4] T. Lautz. *"Barren als Zahlungsmittel. Von der Bronzezeit Bis Ins 20. Jahrhundert, Das Fenster in der Kreissparkasse Köln, Thema 163, Köln*", 2003.

[5] A. Anikin. *"Gold, Verlag Die Wirtschaft, Berlin*", pp. 77-80, 1982.

[6] K. Reiter. *"Die Metalle im Alten Orient unter besonderer Berücksichtigung Altbabylonischer Quellen, Ugarit-Verlag, Münster*", pp. 43, 47, 83-103, 113, 289, 299-305, 401, 405-415, 425, 435-437, 452, 455, 1997.

[7] K. Volke. *"Zu den Anfängen der Analytischen Chemie: Wider Fälscher und Betrüger, Chemie in unserer Zeit*". 4th ed. WILEY-VCH Verlag, Weinheim, pp. 268-275, 2004.

[8] R. Wiegels. *"Silberbarren der Römischen Kaiserzeit, Katalog und Versuch Einer Deutung, Freiberger Beiträge zur Archäologie und Geschichte des Ersten Jahrtausends, Rahden/Westf*", 2004.

[9] L. Biosas. *"Griechische Münzen, Faszination und Geschichte, Numismatischer Verlag Fritz Rudolf Künker, Osnabrück*", pp. 493-504, 2005.

[10] F. Schrötter. "*Wörterbuch der Münzkunde, de Gruyter Verlag, Berlin*", 1930.

[11] B. Kluge. "*Die Monetarisierung Europas in Staufischer Zeit, Numismatisches Nachrichtenblatt (NNB), 09/2010, 14150*", pp. 325-331, 2010.

[12] R. Lehmann and C. Vogt, C. "*Wer Den Pfennig nicht Ehrt Ist Die Mark Nicht Wert?*" 2010a.

[13] "*W. Szaivert-R. Wolters, Löhne, Preise*", 2005. Available from: https://www.w.pd.file:///e:/research2021/xrf&sem/referances/"no title. [Last accessed on 2021 Nov 12].

[14] A. Loehr. "*Probleme der Silberbarren, Numismatische Zeitschrift, Wien*", vol. 24, pp. 101-109, 1931.

[15] H. Ehrhardt. "*Röntgenfluoreszenzanalyse-Anwendung in Betriebslaboratorien*". 2nd ed. Springer-Verlag, Berlin, 1989.

[16] C. Vogt. "*Skript zur Vorlesung, Grundlagen der Analytik I, Röntgenspektroskopie, Universität Hannover*", 2006.

[17] A.Feldhoff,2007.Availablefrom:https://www.notitlefile:///f:/localdisk/analytikvorlesungungen.vogt/fortgeschrittenematerialanalytik/materialanalytik-i-2007-feldhoff.pdf. [Last accessed on 2021 Oct 11].

[18] S. M. M. Hrnjić, G. A. Hagen-Peter,T. Birch, G. H. Barfod, S. Sindbæk and C. E. Lesher. "*No Title Non-Destructive Identification of Surface Enrichment and Trace Element Fractionation 4 in Ancient Silver Coins*", 2020.

ORIGINAL RESEARCH ARTICLE UHD JOURNAL OF SCIENCE AND TECHNOLOGY

Comparative Study of Supervised Machine Learning Algorithms on Thoracic Surgery Patients based on Ranker Feature Algorithms

Hezha M.Tareq Abdulhadi[1], Hardi Sabah Talabani[2]

[1]Department of Information Technology, National Institute of Technology (NIT), Sulaymaniyah, KRG, Iraq, [2]Department of Applied Computer, College of Medical and Applied Sciences, Charmo University, Sulaymaniyah, KRG, Iraq

ABSTRACT

Thoracic surgery refers to the information gathered for the patients who have to suffer from lung cancer. Various machine learning techniques were employed in post-operative life expectancy to predict lung cancer patients. In this study, we have used the most famous and influential supervised machine learning algorithms, which are J48, Naïve Bayes, Multilayer Perceptron, and Random Forest (RF). Then, two ranker feature selections, information gain and gain ratio, were used on the thoracic surgery dataset to examine and explore the effect of used ranker feature selections on the machine learning classifiers. The dataset was collected from the Wroclaw University in UCI repository website. We have done two experiments to show the performances of the supervised classifiers on the dataset with and without employing the ranker feature selection. The obtained results with the ranker feature selections showed that J48, NB, and MLP's accuracy improved, whereas RF accuracy decreased and support vector machine remained stable.

Index Terms: Ranker feature selection, Information gain, Gain ratio, Supervised machine learning algorithms, Thoracic surgery, Cross-validation

1. INTRODUCTION

Tracking health results is fundamental to reinforce quality initiative, managing health care, and educating consumer. At present, employing computer applications in medical fields have had a direct impact on doctor's productivity and accuracy. Health results measurement is one of these applications. Health outcomes are playing an increasing role in health-care purchasing and administration. Nowadays and in most countries, cancer is becoming one of the leading causes of death. At present, lung cancer is the most common presage for thoracic surgery [1].

Access this article online

DOI: 10.21928/uhdjst.v5n2y2021.pp66-74

E-ISSN: 2521-4217

P-ISSN: 2521-4209

In the last several decades, there has been a lot of study in the field of medical science that has used various computing approaches. In the case of medical care, new approaches to data abstraction make data extraction quick and accurate, providing a larger opportunity to work with data for measuring health results. Cancer is a serious health threat that the world is confronting, thus knowing how to anticipate results is essential [2].

Selecting attribute and features in a massive amount of data and using machine learning approaches in recent medical technique might cause the computing process faster and decrease the amount of redundant data. Removing unnecessary data are advantageous since it decreases the difficulty of data processing. Attribute classifier of the data is significant, in the case of thoracic cancer, it leads to the extraction of varied information regarding a specific case of a patient. To reduce and control the victims of lung cancer and

Corresponding author's e-mail: Hezha M.Tareq Abdulhadi, Department of Information Technology, National Institute of Technology (NIT), Sulaymaniyah, KRG, Iraq. E-mail: Hezha.Abdulhadi@nit.edu.krd

Received: 25-07-2021 Accepted: 12-12-2021 Published: 15-12-2021

thoracic surgery patients, ranker feature selection techniques became an important and necessary method, because it can challenge and solve this kind of problems. In general, machine learning and ranker algorithms are a technique for classifying patient and disease datasets and separate the data to relevant and irrelevant. There are several studies worked on thoracic surgery. Therefore, this work shed a light on the success rate of machine learning algorithms with ranker feature selections in classifying thoracic surgery patients. The major goal is to obtain an accurate prediction of the result after employing different approaches [3].

This research is done by a famous tool which is WEKA, used for analyzing and classifying data with famous machine learning algorithms. Five different machine learning algorithms employed in this study which are J48, Random Forest (RF), Naïve Bayes, Multilayer Perceptron, and Support Vector Machine (SVM) with two famous ranker feature selections algorithms, information gain and gain ratio (GR). We have performed a classification on the thoracic surgery dataset through machine learning techniques and ranker algorithms.

The rest of this paper is organized as follows: Section 2 describes some background concepts relevant to our review. Section 3 describes the problem and proposed method. Section 4 will present the experiments and results, and finally, the conclusion is stated in Section 5.

2. LITERATURE REVIEW

Various studies have been published that emphasize the significance of methodology in the realm of medical diagnosis. This research used various methods to the problem and obtained reasonable classification accuracies. Following are some examples:

Several studies have been implemented in the medical field for analyzing data to discover patterns and predict outcomes. Techniques such as Synthetic Minority Over-sampling Technique (SMOTE) are used to rectify the unbalanced data. Various measures are used for predicting results. For balancing the data by oversampling the minority class, the comparison between prediction methods such as Artificial Neural Network (ANN), Naive Bayes techniques, and Decision Tree Algorithm is explained in [3] by employing 10-fold cross-validation and SMOTE. The receiver operating characteristics summed the classifier performance based on the true positives and true negatives error rates; the ANN achieves the highest accuracy in this scenario. Another 10 folds cross-validation study in life expectancy prediction was conducted by [1] using Naïve Bayes, Logistic Regression, and SVM with the RF concept, which uses the tree classification technique to average deep multiple trees that are trained using different fragments of the current training set.

Jahanvi Joshi *et al.* offered the detailed proof that K-nearest neighbor (KNN) provides preferable accuracy than expectation-maximization classification technique. Employing the Farthest first algorithm, they showed that 80% of patients were healthy and 20% of patients were sick, which are very close to KNN technique outcome [4].

Vanaja *et al.* explained that each feature selection approach has its effects and weak points inclusion of greater characteristics reduces accuracy. This survey was demonstrated that the feature selection algorithms improve the classifier accuracy consistently [5].

Zieba *et al.* employed boosted SVM to estimate post-operative life expectancy in their study. During the research, an Oracle-based technique to extract decision rules from the boosted SVM for solving problems with unbalanced data had been used [6].

Sindhu *et al.* analyzed thoracic surgical data using six classification techniques (Naive Bayes, J48, PART, OneR, Decision Stump, and RF). An experiment was done and discovered that RF provides the greatest classification accuracy with all split percentages [1].

Another research evaluated the performance of four machine learning algorithms (Nave Bayes, Simple logistic regression, Multilayer perceptron, and J48) with their boosted variants using various measures. The outcomes showed that the boosted simple logistic regression approach outperforms or is at least competitive with the other four machine learning techniques, with an average score of 84.5% [7].

In this work, four various machine learning algorithms will be used for post-life expectancy estimation after thoracic surgery, by employing two novel metrics which are information gain (IG) and GR that can be used to improve the accuracy of the algorithms and provide a reasonable result.

3. METHODOLOGY

In this work demonstrated in Fig. 1, the thoracic surgery dataset is used and pre-processed to remove unbalanced and useless data, then filling missing values. The pre-processed

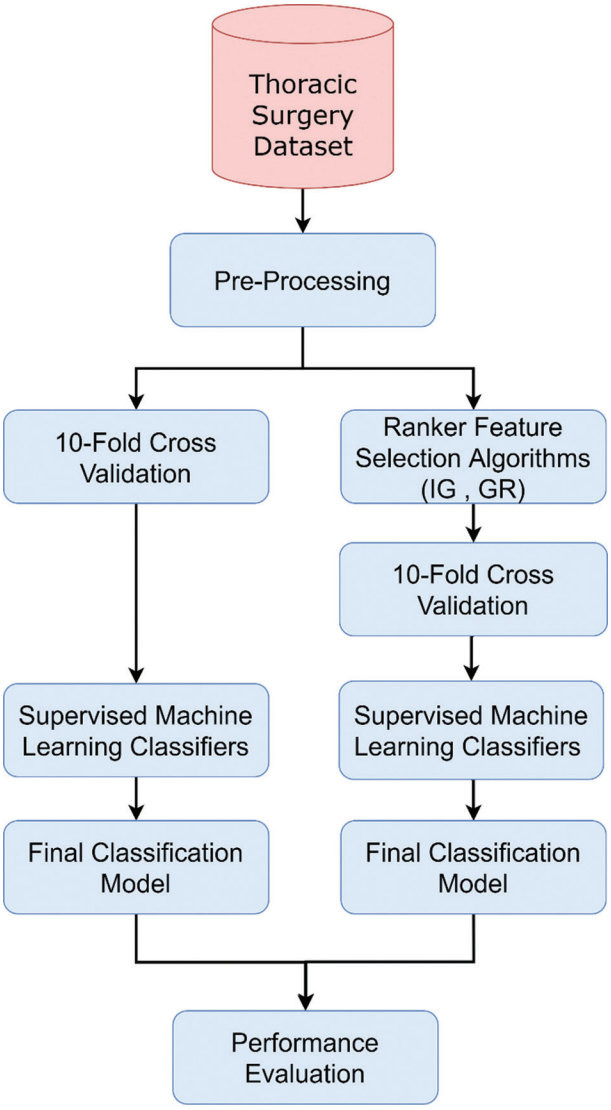

Fig. 1. Flowchart of the proposed method.

dataset will be used in two different tests. The two main purposes of this paper are as follows: First, to analyze the effect of number of attributes on accuracy of machine learning to solve the problem for prediction of the post-operative life in lung cancer patients reducing the number of attributes and increasing the accuracy is required to minimize the computational time of prediction techniques. Second, to make a comparison between the supervised classifiers performances before and after using ranker feature algorithms with employing 10-fold cross-validation technique for splitting the dataset. Notably, cross-validation is a method to evaluate a predictive model by partitioning the original sample into a training set to train the model and a validation/test set to evaluate it. The first test will be done on the dataset employing supervised machine learning classifiers then the results will be compared with the other test according to some measurement criteria. The second test will be done on the dataset using the attribute ranking methods (IG and GR) to eliminate the redundant and irrelevant attributes from the original set of attributes and to evaluate the importance of an attribute by measuring the IG and GR with regard to the class. After attribute evaluation, the dataset will be separated randomly by applying 10-fold cross-validation and then the classification process will begin with the supervised classifiers to find the best performance among them. The final classification model of both tests will be evaluated and compared based on some performance criteria explained in the next chapter.

3.1. Thoracic Surgery Corpus

The dataset used in this paper was collected from the information of patients who were suffering from lung cancer and underwent lung resections in 2007 and 2011 at the Center for Thoracic Surgery in Wroclaw, which, in turn, is affiliated with the Lower Silesian Center for Pulmonary Diseases and the Department of Thoracic Surgery at the University of Wroclaw medical. It is worth noting that this dataset has been extracted from Wroclaw Thoracic Surgery Centre that has been gathered by the National Lung Cancer Registry of the Polish Institute of Lung Diseases and Tuberculosis in Warsaw [8]. In general, the dataset consists of 17 attributes (14 nominal and three numeric) with 470 records, which are detailed in Table 1.

3.2. Pre-Processing

The dataset is pre-processed removing unbalanced and useless data through SMOTE, a bootstrapping algorithm to solve this issue (SMOTE). Other methods, ROS, are also being tested (random over sampler) for that issue. In this work, several new features are designed to better describe the underlying connections among different dataset features, resulting in enhanced model performance [9]. The operations of correcting discrepancies in the data reducing noise in outliers and filling in missing values using one of the data preprocessing methods called (data cleansing).

3.3. Ranker Feature Selection

The two basic principles of ranker-based feature selection algorithms are as follows: First, the evaluation of features related to their impact on the process of data classification or analysis. Second, building a ranking list based on its score using the desired features (the most influential on the accuracy of the algorithm performance) that were identified to create a subset. Among the different types of rank-based feature selection algorithms, two main types. GR and IG

TABLE 1: Descriptions of thoracic surgery dataset attributes

Attribute ID	Attribute name	Attribute type	Attribute description
1	DGN	Nominal	Diagnosis-specific combination of ICD-10 codes for primary and secondary as well multiple tumors if any (DGN3, DGN2, DGN4, DGN6, DGN5, DGN8, and DGN1)
2	PRE4	Numeric	Forced vital capacity – FVC
3	PRE5	Numeric	Volume that has been exhaled at the end of the first second of forced expiration – FEV1
4	PRE6	Nominal	Performance status – Zubrod scale (PRZ2, PRZ1, and PRZ0)
5	PRE7	Nominal	Pain before surgery (T,F)
6	PRE8	Nominal	Hemoptysis before surgery (T,F)
7	PRE9	Nominal	Dyspnea before surgery (T,F)
8	PRE10	Nominal	Cough before surgery (T,F)
9	PRE11	Nominal	Weakness before surgery (T,F)
10	PRE14	Nominal	T in clinical TNM – size of the original tumor, from OC11 (smallest) to OC14 (largest) (OC11, OC14, OC12, and OC13)
11	PRE17	Nominal	Type 2 DM – diabetes mellitus (T,F)
12	PRE19	Nominal	MI up to 6 months (T,F)
13	PRE25	Nominal	PAD – peripheral arterial diseases (T,F)
14	PRE30	Nominal	Smoking (T,F)
15	PRE32	Nominal	Asthma (T,F)
16	AGE	Numeric	Age at surgery
17	Risk1Y	Nominal	1 year survival period – (T)rue value if died (T,F)

were adopted and applied to check whether they had a positive effect in increasing the performance accuracy of the supervised algorithms used in this paper. Indeed, and through the obtained results, it was proved that after their application, there was a relative increase in the performance of the algorithms [10].

3.3.1. GR

It is an enhancement version of IG. It calculates the GR in connection with the class. Whereas the IG selects the feature with a huge number of value, this method's objective is to maximize the feature IG while decreasing the value numbers [11].

$$Gain\ Ratio\ (Feature) = \frac{Gain\ (Feature)}{SplitInfo\ (Feature)} \quad (1)$$

In the following, the value for splitting information is shown. It is the result of splitting the training dataset D into v partitions, each corresponding to v outcomes on the attribute feature:

$$SplitInfo\ (A) = -\sum_{j=1}^{v} \frac{|D_j|}{|D|} . Log_2 \frac{D_j}{D} \quad (2)$$

3.3.2. IG

The attribute values are evaluated by the IG method with the calculation of IG concerning the class which calculated the difference in information between cases where the feature's value is known and cases unidentified. Each feature will get an assigned score, indicating how much more information about the class is fetched when that feature is used [11].

$$InfoGain\ (Feature) = H\ (Class) - H\ (Class\ |Feature) \quad (3)$$

Where, H refers to entropy is:

$$H\ (X) = -\sum_{i-1}^{n} P\ (x_i) . Log_2 P\ (X_i) \quad (4)$$

3.4. 10-Fold Cross-Validation

Cross-validation is one of the standard machine learning techniques used in Weka workbench. Ten-fold cross-validation is a mechanism for evaluating predictive models by dividing the original dataset into two subsets: The training set and the test set in which the used dataset is randomly divided into 10 equal-sized of subparts, one subpart is kept as validation data for testing, and the remaining nine parts are used as training data. Hence, iterating the cross-validation process 10 times, the results for 10-fold can then be averaged to produce one evaluation. The advantage of this technique is that all the datasets will be used in both training set and testing set [12]. The reason for the selection of the cross-validation technique is that it reduces the variance in the estimation a lot more than the other techniques. Accordingly, the dataset used in this paper has been separated according to this technique. This ensures that we will obtain the necessary estimations as well as monitor the performance of the classifiers.

3.5. Supervised Machine Learning Classifiers

Supervised learning mechanism is a type of machine learning in which machines are trained employing labelled training data. In other words, when the used dataset is divided into Training and testing. The supervised learning mechanism is used on a training dataset consisting of known input data (X)

and output variable (Y) to build a module and implement it to predict the output variables (Y) of the testing data [13]. The following are the supervised learning algorithms that have been used in this paper.

3.5.1. RF

A RF algorithm, as its name suggests, is made up of a large number of individual decision trees that act as a set. Each tree in the RF emerges from the prediction of the class and becomes the class with the most votes the basic principle behind the RF algorithm is a simple but powerful concept – the wisdom of the majority crowd. In data science, the reason the RF model is so successful is that a large number of relatively uncorrelated (trees) models acting as a committee will outperform any of the single-component models. The low correlation coefficient between the models is key. Just like how investments with a low coefficient of correlation are aggregated, uncorrelated models can produce aggregate forecasts that are more accurate than any individual forecasts. The reason for this wonderful effect is that trees protect each other from their mistakes (as long as they don't all err in the same direction constantly). While some trees may be wrong, many others will be right so that the trees as a group can move in the right direction [14]. The mathematical formula of the algorithm is as follows [15].

$$RFfi_i = \frac{\sum_{j \in all\ trees} norm\ fi_{ij}}{T} \quad (5)$$

Where, RFfi sub(i)= the significance of feature i calculated from all trees in the RF model

normfi sub(ij) = the normalized feature importance for I in tree j.

3.5.2. J48

The process of classification using a decision tree uses gain information to divide the tree. The first step is to gain information for each attribute. The attribute with the largest amount of IG will be the node root of the decision tree. The decision tree technique aims to divide the database with a specific goal that has already been determined, and the presence of a certain element in one of the groups, which is represented here by the branches, becomes a result because it achieved the series of conditions set down to this branch and not only because it is similar to the rest of the elements [16]. Although, it has not been defined similarity in this case. The J48 and the algorithms that are used to produce it can be complex, but the results that lead to it can be shown in a simple, easy-to-understand form, and with a high level of utility. The algorithm steps are as follows:

First: If the instances belong to the same class, the leaf is tagged with a comparable class.

Second: The prospective data for each attribute will be calculated, and the data gain from the attribute test will be calculated.

Third: Eventually, based on the current selection parameter, the best attribute will be selected.

3.5.3. Naive Bayes

It is a classification model in machine learning fields which based on probability. A Naive Bayesian model is simple to construct and does not require iterative parameter estimation, making it ideal for huge datasets [17]. From P(c), P(x), and P(x|c), the Bayes theorem may be used to get the posterior probability, P(c|x). The effect of the value of a predictor (x) on a given class (c) is independent of the values of other predictors, according to the Naive Bayes classifier. The following is the formula of the model.

$$P(c|x) = \frac{P(x|c)P(c)}{P(x)} \quad (6)$$

$P(c|X) = P(x_1|c)\ P(x_2|c).P(x_n|c)\ P(c)$

P(c|x): Rear probability of class(target)given predictor(attribute).

P(c): The prior probability of class.

P(x|c): Likelihood which is the probability of predictor given class.

P(x): The prior probability of predictor.

3.5.4. Multilayer perceptron

It is a category of feedforward ANN which creates a set of outputs from a set of inputs. The perceptron, which comprises numerous inputs Xi multiplied by a scalar value known as weight Wij and a bias bj, was one of the earliest PEs constructed [18]. A specified activation function f is used to process the acquired result, which may be explained as follows:

$$Y_j = f\ [\sum_i (W_{ij} * X_{i+} b_i) \quad (7)$$

$$Y_j = f\ [\sum_i (W_{ij} * X_{i+} b_i)] \quad (8)$$

The hyperbolic tangent function tanh, which is represented as follows, is the most frequent activation function f utilized in perceptron.

$$tanh(x) = 2.\sigma(2x) - 1,$$

$$where\ \sigma(x)\ is\ \sigma(x) = \frac{e^{x}}{1 + e^{x}} \quad (9)$$

The MLP network is used to solve nonlinear separation issues by connecting numerous perceptions in one or more hidden layer topologies. The aim is to discover the error function with the lowest possible error in proportion to the connection weights. The error function is explained as follows:

$$E = \frac{1}{2} \sum_{j \in M} (\hat{y}_m - y_m)^2 \quad (10)$$

with $\hat{y}_m$ being the desired output of m'th y_m.

3.5.5. SVM

The SVM algorithm classifies data for two divisions by taking input data and generating output predictably. The best way for implementing this technique is to build a model to text corpuses while any training sample belonged to one of the classes. After that, the data will be divided into two categories with the way of constructing an N-dimensional hyperplane. To separate data, SVM will build two hyperplanes but they should be paralleled in both sides of the hyperplane while the separated hyperplane will increase the space between other hyperplanes [19]. SVM is capable of conducting regression analyze and extending it while performing a numerical calculation. The formula of the algorithm is shown below:

$$K(x, y) = (x.y + c) \quad (11)$$

4. EXPERIMENTS AND RESULTS

In machine learning, and specifically in the field of data classification, there are many commonly accepted criteria for measuring the classification performance for the machine learning algorithms. In this research, the scales shown in the following tables were used to explain the difference in the performance of the algorithms used to classify the data. Then, the performance of each algorithm is compared before and after applying each of the classifier feature selection algorithms GR and IG.

In general, through the results obtained in Tables 2 and 3 with Figs 2-5, it is clear that there is a difference in the stability and instability in the classification performance of algorithms with or without ranker feature selections in the process of classifying thoracic surgery datasets. To begin with regard

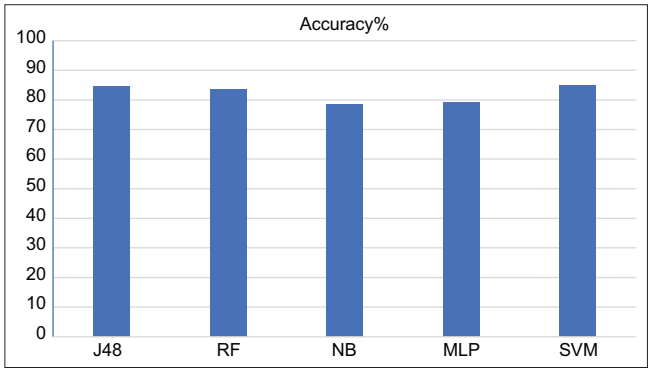

Fig. 2. Accuracy of the classifiers before feature selections.

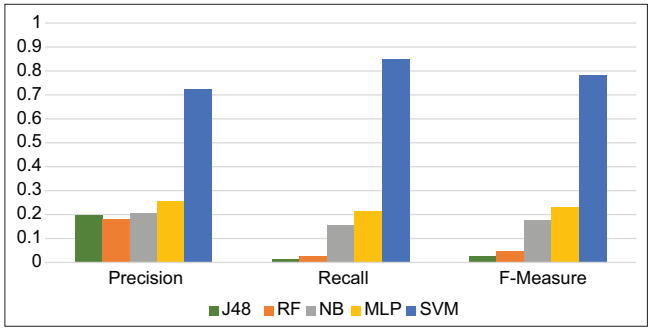

Fig. 3. Precision/recall and F-measure of the classifiers before feature selections.

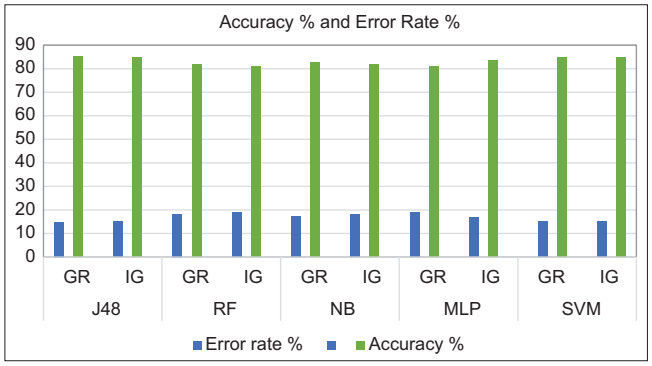

Fig. 4. Accuracy and error rate of the classifiers after using feature selections.

TABLE 2: Performance measurements before implementing ranker attribute evaluators

Supervised algorithms	Precision	Recall	F-measure	Accuracy
J48	0.200	0.014	0.027	84.46%
RF	0.182	0.029	0.049	83.62%
NB	0.208	0.157	0.179	78.51%
MLP	0.259	0.214	0.234	79.14%
SVM	0.724	0.849	0.782	84.89%

TABLE 3: Performance measurements after implementing ranker attribute evaluators (Gr)/(IG)

Performance measurements	J48		RF		NB		MLP		SVM	
	GR	IG	GR	IG	GR	IG	GR	IG	GR	IG
Precision	0.724	0.724	0.750	0.751	0.744	0.745	0.767	0.799	0.724	0.724
Recall	0.851	0.849	0.817	0.811	0.828	0.819	0.811	0.834	0.849	0.849
F-measure	0.783	0.782	0.777	0.776	0.777	0.776	0.785	0.811	0.782	0.782
Error rate %	14.893	15.106	18.297	18.936	17.234	18.085	18.936	16.595	15.106	15.106
Accuracy %	85.106	84.893	81.702	81.063	82.766	81.914	81.063	83.404	84.893	84.893

TABLE 4: Classification/time measurements before implementing ranker attribute evaluators

Classification measurements	J48	RF	NB	MLP	SVM
Correctly classified instances	397	393	369	372	399
Incorrectly classified instances	73	77	101	98	71
Time (milliseconds)	30	210	9	1820	90

TABLE 5: Classification/Time measurements after implementing ranker attribute evaluators (GR)/(IG)

Classification measurements	J48		RF		NB		MLP		SVM	
	GR	IG	GR	IG	GR	IG	GR	IG	GR	IG
Correctly classified instances	400	399	384	381	389	385	381	392	399	399
Incorrectly classified instances	70	71	86	89	81	85	89	78	71	71
Time (milliseconds)	40	10	140	90	10	**9**	1150	1290	30	40

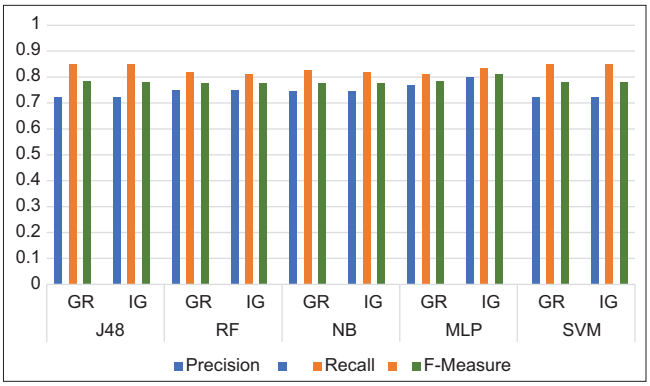

Fig. 5. Precision/recall and F-measure of the classifiers after ranker evaluators.

in J48, NB, and MLP algorithms, we noticed an increment in accuracy of the classification performance, in which the accuracy of J48 is 84.46% without using ranker feature selections, as shown in Table 2 and Fig. 2, this performance has been improved using ranker feature selections GR and IG to 85.106% and 84.893%, respectively, as shown in Table 3 and Fig. 4. Furthermore, the classification performance accuracy of NB is 78.51% without ranker, as shown in Table 2 and Fig. 2, the performance is raised with ranker GR and IG to 82.766% and 81.914%, respectively, as shown in Table 3 and Fig. 4. Moreover, the classification performance accuracy of MLP is 79.14% without ranker feature selections, as shown in Table 2 and Fig. 2, this accuracy is enhanced with ranker GR and IG to 81.063% and 83.404%, respectively, as shown in Table 3 and Fig. 4.

Another point to consider is with regard to the RF algorithm, we notice a decrement in performance accuracy of RF which was 83.62% without ranker feature selections, as shown in Table 2 and Fig. 2, the accuracy is raised after employing ranker GR and IG to 81.702% and 81.063%, respectively, as shown in Table 3 and Fig. 4. Whereas, in testing SVM algorithm, there are no changes observed in the accuracy during classification as it remains equal in both cases and its performance did not change with both feature selections, the accuracy without ranker selections was 84.89%, as shown in Table 2 and Fig. 2, and it remains stable with no any effectiveness with ranker selections with accuracy 84.89% for both feature selection algorithms GR and IG, as shown in Table 3 and Fig. 4.

In Table 4, it is clear that SVM is the most accurate algorithm in classifying instances correctly with 399 instances out of a total of 470 instances units without ranker feature selections. However, it is not the fastest in constructing the model, as it took 0.09 seconds for classifying the whole dataset records.

Besides, MLP is the slowest algorithm among the other algorithms in the classification process as it took 1.82 seconds without using ranker feature selections. In contrast, NB is the lowest in classifying instances correctly with 369 instances out of a total of 470 instances without using franker feature selections. However, it is the fastest in building the model, as it took 0.00 seconds to classify the whole dataset records.

In Table 5, a drastic change can be observed, it is clear that J48 is the most accurate algorithm in classifying instances correctly with 400 instances out of a total of 470 instances units with ranker feature. However, it is one of the fastest algorithms in constructing the model using IG which took 10 milliseconds for classifying the whole dataset records. In contrast, both RF using IG and MLP using GR are the lowest in classifying instances correctly with 381 instances out of a total of 470 instances without ranker feature. Furthermore, MLP remained the slowest in building the model, as it took 1290 milliseconds to classify the whole dataset records using IG. The NB remained the fastest algorithm among the others in the classification models as it took 0.00 seconds with IG. In contrast, both RF using IG and MLP using GR are the lowest in classifying instances correctly with 381 instances out of a total of 470 instances without franker feature selections. However, MLP remained the slowest in building the model, as it took 1290 milliseconds in classifying the whole datasets using IG. Finally, NB remained the fastest algorithm among the other algorithms in classifying the dataset as it took 9 milliseconds with using IG.

5. CONCLUSION

The comparison made in this paper showed a significant effect of the ranker features on supervised classification algorithms. Through the obtained results, we concluded that the use of ranker feature selections leads to improving the classification performance of particular algorithms, as done with J48, MLP, and NB algorithms. In contrast, ranker feature selection reduced the performance of RF. Moreover, specific algorithms such as SVM remained stable before and after ranker feature selection concerning classification performance. Similarly, as for the speed of building the model, the NB algorithm did not change its speed in both cases by recording the least time for data classification and the fastest among the other algorithms, 9 milliseconds. Eventually, the highest performance in the accuracy of classification was the J48 algorithm using GR, which amounted to 85.1%. Other feature selection algorithms can be employed to improve the used algorithms' performance in future work.

REFERENCES

[1] S. Prabha, S. Veni and S. Prabha. "Thoracic Surgery analysis using data mining techniques". *International Journal of Computer Technology and Applications*, vol. 5, no. 1, pp. 578-586, 2014.

[2] K. Kourou, T. P. Exarchos, K. P. Exarchos, M. V. Karamouzis and D. I. Fotiadisa. "Machine learning applications in cancer prognosis and prediction". *Computational and Structural Biotechnology Journal*, vol. 13, pp. 8-17, 2015.

[3] A. S.Dusky and L. M. El Bakrawy. "Improved prediction of post-operative life expectancy after Thoracic Surgery". *Advances in Systems Science and Applications*, vol. 16, no. 2, pp. 70-80, 2016.

[4] J. Joshi, R. Doshi and J. Patel. "Diagnosis of breast cancer using clustering data mining approach". *International Journal of Computer Applications*, vol. 101, no. 10, pp. 13-17, 2014.

[5] S. Vanaja and K. R. Kumar. "Analysis of feature selection algorithms on classification: A survey". *International Journal of Computer Applications*, vol. 96, no. 17, pp. 29-35, 2014.

[6] M. Zięba, J. Tomczak, M. Lubicz and J. Świątek. "Boosted SVM for extracting rules from imbalanced data in application to prediction of the post-operative life expectancy in the lung cancer patients". *Applied Soft Computing*, vol. 14, pp. 99-108, 2014.

[7] M. U. Harun and N. Alam. "Predicting outcome of thoracic surgery by data mining techniques". *International Journal of Advanced Research in Computer Science and Software Engineering*, vol. 5, no. 1, pp. 7-10, 2015.

[8] M. Lubicz, K. Pawelczyk, A. Rzechonek and J. Kolodziej. "*UCI Machine Learning Repository: Thoracic Surgery Data Data Set*", 2021. Available from: https://archive.ics.uci.edu/ml/datasets/thoracic+surgery+data [Last accessed on 2021 Oct 08].

[9] S. Xu. "*Machine Learning-Assisted Prediction of Surgical Mortality of Lung Cancer Patients*". The IEEE International Conference on Data Mining, 2019.

[10] S. Subbiah and J. Chinnappan. "An improved short term load forecasting with ranker based feature selection technique". *Journal of Intelligent and Fuzzy Systems*, vol. 39, no. 5, pp. 6783-6800, 2020.

[11] D. El Zein and A. Kalakech. "*Feature Selection for Android Keystroke Dynamics*". 2018 International Arab Conference on Information Technology, 2018.

[12] H. Talabani and A. V. C. Engin. "*Performance Comparison of SVM Kernel Types on Child Autism Disease Database*". International Conference on Artificial Intelligence and Data Processing, 2018.

[13] F. Y. Osisanwo, J. E. T. Akinsola, O. Awodele, J. O. Hinmikaiye, O. Olakanmi and J. Akinjobi. "Supervised machine learning algorithms: Classification and comparison". *International Journal of Computer Trends and Technology*, vol. 48, no. 3, pp. 128-138, 2017.

[14] M. Rathi and V. Pareek. "Spam mail detection through data mining a comparative performance analysis". *International Journal of Modern Education and Computer Science*, vol. 5, no. 12, pp. 31-39, 2013.

[15] J. Wong. "*Decision Trees Medium*", 2021. Available from: https://towardsdatascience.com/decision-trees-14a48b55f297 [Last accessed on 2021 Oct 08].

[16] A. Yadav and S. Chandel. "Solar energy potential assessment of Western Himalayan Indian state of Himachal Pradesh using J48 algorithm of WEKA in ANN based prediction model". *Renewable Energy*, vol. 75, pp. 675-693, 2015.

[17] K. Vembandasamy, R. Sasipriya and E. Deepa. "Heart diseases

detection using Naive Bayes algorithm". *International Journal of Innovative Science, Engineering and Technology*, vol. 9, no. 29, pp. 441-444, 2015.

[18] M. Khishe and A. Safari. "Classification of sonar targets using an MLP neural network trained by dragonfly algorithm". *Wireless Personal Communications*, vol. 108, no. 4, pp. 2241-2260, 2019.

[19] H. Talabani and A. V. C. Engin. "*Impact of Various Kernels on Support Vector Machine Classification Performance for Treating Wart Disease*". International Conference on Artificial Intelligence and Data Processing, 2018.

p-ISSN 2521-4209
e-ISSN 2521-4217

جامعة التنمية البشرية
UNIVERSITY OF HUMAN DEVELOPMENT

UHD Journal of Science and Technology

A Scientific periodical issued by University of Human Developement

Vol.5 No.(2) December 2021

2021 2721

e.mail:jst@uhd.edu.iq