ICIC Express Letters
Part B: Applications
Volume **13**, Number **12**, December **2022**

 ISSN 2185-2766
pp. **1321–1330**

A NOVEL APPROACH TO WIKIPEDIA REFERENCES CLASSIFICATION

Batool AlNajrani, Amnah Alghamdi, Meera AlOtaibi, Safa Aldawod
Atta-ur-Rahman* and Majed Nabil

Department of Computer Science
College of Computer Science and Information Technology
Imam Abdulrahman Bin Faisal University
P.O. Box 1982, Dammam 31441, Saudi Arabia
{ 2200500057; 2200500067; 2200500087; 2190500146; 2160007950 }@iau.edu.sa
*Corresponding author: aaurrahman@iau.edu.sa

Received April 2022; accepted July 2022

Abstract. *Wikipedia is becoming a significant source of obtaining an immediate, to the point and state of the art information about an entity like object, person, or topic. Being an open source, anyone can edit the information anytime, that makes the Wikipedia unreliable. Though Wikipedia has its own mechanism for vetting the information by means of splitting the resources into primary and secondary for cross referencing. However, still there are several loopholes. Researchers spend time in collecting references that are reliable and trusted to rely on and to build up their research. In this paper, Wikipedia references classifier has been proposed to help the researcher searching process easier by extracting and classifying the references from the XML files. The Wikipedia dump dataset has been selected and it is stored as XML file after retrieval. In this paper, the proposed approach is rule based approach for extracting the information by using the regular expressions. This application can extract the references and classify them based on their type like whether it is reference to a book, website, journal, conference, or thesis. Moreover, it helps in classification of citations in terms of their relevance to the topic, year, quartile, impact factor and other metrics. The evaluation of the proposed technique possesses 97.5% average accuracy.*
Keywords: Wikipedia reference classifier, XML, Information extraction, Rule based

1. **Introduction.** Research is creative and intelligence work that increases knowledge and adds value to a different area. Nowadays, the web has become the most common source of information mining for ease of use and rapid response [1]. With this spread, knowledge has grown greatly, and most of the information has become available in the form of unorganized natural language documents in digital libraries, digital documents, digital encyclopedias, and others. So, the need for methods to extract information from natural languages or unstructured texts increased. Information mining techniques have become an important research field. Text extraction is like data extraction, but it is aimed at unstructured texts. Text extraction techniques can convert unstructured text into an organized textual database. Our goal is to extract information in a meaningful way from Wikipedia by means of classification. Following highlights study background.

Wikipedia is defined as a free electronic encyclopedia, and it is included under the concept of open source. Wikipedia was launched in 2001 under the supervision of the non-profit company called Wikimedia. The aim of its launch was to organize knowledge and share information as it aims to create electronic encyclopedias in multiple languages [2]. Wikipedia pages contain 250 languages, and articles in the English language are only a quarter of them. Wikipedia has emerged as an open source consisting of articles in

DOI: 10.24507/icicelb.13.12.1321

various fields. The strength of Wikipedia is that it contains a huge collection of trusted threads [3]. Therefore, Wikipedia has become one of the most visited websites on the Internet, which can be useful for researchers and students of knowledge. Wikipedia pages are written in the natural language to which methods of extracting information can be applied. Obtaining a method for extracting information from Wikipedia in an organized and fast way that will facilitate researchers in plenty of way for finding the consolidated information and save their time and efforts. Information extraction is defined as the use of computational methods to recognize relevant information in any text that is generated for human use, and then convert the identified information into a representation that is suitable for computer processing like storage and retrieval [4]. The input for information extraction system is a collection of text documents (research papers, articles, web pages, reports, resumes, and so on) and then the output is a representation for the related information that was extracted from the text document source. The goal for information extraction system is to analyze free text and find relevant information in the form of structured information [4]. Information extraction system can extract pre-specified types of entities, relationships, or events to build more meaningful representation of its semantic content. There are two generic types of information extraction system approaches, knowledge engineering approach and machine learning approach. The knowledge engineering strategy is applied to the construction of the grammar rules and to discovering the domain patterns. However, for named entity recognition, prior training of the machine learning algorithm is required [5].

Regular expressions also called as Regex or Regexp are a pattern rule for matching text that are written in form of metacharacters, plain text or quantifiers. The regular expressions are mainly used for finding patterns and replacing the matched pattern. Also, it uses in extraction information from any text by searching of a specific search pattern. For example, searching for a sequence of ASCII characters. Several programming languages like Java, provide in built support to write the regular expressions for finding specific patterns out of the provided text. There are tremendous applications of regular expressions in various types of information extraction systems. Regular expressions have been utilized in many other applications of information extraction such as from scientific documents, email text, scientific reports, and variety of other documents to get the desired patterns. As mentioned above, Wikipedia has a massive number of pages and articles. Thus, this website could help researchers to have references for their research topic. Moreover, references that researcher cite should be highly reliable so that their papers become more powerful and logical. Hence, not all Wikipedia's articles meet this requirement. Wikipedia gives the authority to everyone to add and edit which makes Wikipedia's articles less reliable. Furthermore, looking into articles that could contain unreliable references could waste researchers time and efforts. From that point, helping researchers in investigating all Wikipedia's articles reference and classify them based on categories (Web, Journal, Thesis and Book). The use of information extraction technology facilitates this work and cuts efforts.

In this paper the proposed method aims to use the technology information extraction using the regular expression to extract all references for each article that was searched and classified according to the category (Web, Book, Journal and Thesis). This information is extracted from a database containing XML files. The proposed method would facilitate the task for researchers and save their time. To help researcher to overcome this problem, we are motivated by a vision of creating information extraction system which can extract article references information like titles, authors, date accessed, and ISBN. Then, the extracted information especially from the journal will be investigated again to decide whether this reference is reliable or not. This will help Internet users, students, and researchers by making their searching process easier and more accuracy and less time

consuming. Also, the use of regular expression has two main advantages. First, the expression provides natural mechanism for domain user to provide domain knowledge of the structure of the extracted entity. Second, by using the regular expression, it also gives the opportunity to restrict the search space.

The structure of the rest of paper is as follows. Section 2 covers the literature review of the most recent attempts and methods used in the field of information extraction. Section 3 provides the methodology and approaches. Section 4 shows and discusses the result. Finally, Section 5 gives the conclusion and future work.

2. **Literature Review.** In [6], researchers are extracting text information from articles. The articles that have been used were collected from different resources: IEEE, springer, Wiley, Science Direct, Cambridge and SAGE. They were focusing only on a single term which is "Mobile Learning in Higher Education". After retrieving the data, it is saved into six different folders based on the article resource. Then, researchers applied two techniques: word cloud and word frequency. As a result, the word "Learning" was the most frequent keyword from all the resources. Researchers in [7] have been focusing on extracting the biological information from scientific articles. The framework was not only extracting the authors' information but also getting into the body structure and the bibliography. The framework was extracting the scientific information by dividing the articles into three sections. The first section is the metadata where they extract the title, author's name, affiliation, and email. After that, structure section is where they extract text from heading and body, URL, headings of the tables and figures, and footnotes. Finally, the bibliography is where they extract citation and references. The framework was implemented by python. This work was using two dataset: Google Scholar and PubMed. The accuracies were 13% and 44%, respectively. In [8], authors presented Wikipedia-based Open Extractor (WOE) technique. This is the first system that can transfer information from collaboratively editing Wikipedia to learn an open information extractor. This approach produces relation-specific training instances through comparing Infobox attribute values and match them to the corresponding sentences. Then epitomize these instances to relation separate training data to train an unlexicalized extractor, like TextRunner. WOE is working in two styles, Conditional Random Field (CRF) extractor WOE (WOE^pos), which is utilized in perfunctory features such as part-of-speech (POS) tags, and pattern classifier (WOE^parse), which is trained on dependency path patterns. When it is adjusted to employ dependency-parse features, it gives higher accuracy. They generated positive and negative training data using four various methods and one learning algorithm, and then they compare the results with TextRunner's result. They conclude the following, (WOE^pos) and TextRunner are equal in speed; however, (WOE^pos) gains higher F-measure with 18% and 34% percentage on three bulks. And (WOE^parse) reaches an F-measure higher than TextRunner with 72% and 91%, on the other hand it is slower than TextRunner 30X times approximately, this is because it needs more time for parsing. In [9], Mihálta developed a large-scale and reliable information extraction system and implemented it on Wikipedia online encyclopedia, which can get the information that is unreachable from the organized sections such as info boxes, tables, and category labels of the article pages. Using templates, or frames which contain slots that match the entities that are in the given relation to process texts. The system relies on natural language processing tools, for example, syntactic analyzing, and named entity recognition. To evaluate the performance of the linguistic analysis and to explore the potentials of the verb frame-based information extraction approach, they tested system that uses hand-crafted extraction patterns, and they examined a technique to learn extraction patterns automatically. The results indicated that while accuracy is comparable for the approaches, they prefer the recall for different approaches. The experiments

discovered that according to recall, manual patterns beat machine learning and automatically extracted patterns; however, machine learning has superior recall compared with automatically extracted patterns. The authors in [10] have proposed a method for extracting lack of information of one language by comparing native article with non-native article in Wikipedia. The proposed system is dividing the articles into segments based on the table of contents then comparing the similarity of multilingual Wikipedia articles based on link graph and article structure. The results showed that the system can also extract relevant passages of different articles from Wikipedia. In the research of [11] the authors used Japanese Wikipedia XML dump files to propose method for extracting IsA assertions which is hyponymy relations, Atlocation assertions like the information of an object or place, LocatedNear assertions to inform neighboring location and Createdby assertions to inform the creator of new object. The authors used Hyponymy extraction tool that can make an analysis of definition, category, and hierarchy structures of Wikipedia articles to extract [IsA, Atlocation, LocatedNear, Createdby] assertions and produce an information-rich taxonomy. The results showed an effective precision achieved from extracted assertions. Researchers in [12] provided a method using Map-reduce to extract the titles from large bibliographic datasets in XML format. They used two mapper functions: one was sorted dataset based on year, and the second used to extract titles from a dataset which sorted. They found the performance of the proposed approach presented quick, efficient, and highly scalable titles extraction by using Hadoop Map-reduce compared with regular XML parsing without Map-reduce. Authors in [13] proposed a technique to extract the citations from google scholar and PubMed datasets. The scheme was better in form of information extraction in general but suffered from low accuracy in certain areas. In [14], authors proposed an ontological information extraction approach to extract the data from published scholarly articles of various societies [15]. The scheme was promising in terms of accuracy and eventually the information was stored in digital library [16,17]. Authors in [18], extracted the table of contents from the PDF books using the regular expression, originally motivated by [19]. Similarly, a technique for multilabel classification of computer science papers was presented in [20] by exploiting the references of the papers. The scheme was promising in terms of article classification; however, information extraction was not in the scope of the scheme. Also, several other schemes have been presented in the literature in classification in various fields [21-30].

3. **Proposed Wikipedia Reference Classification.** This paper aims to extract the references type and their details that have been used in Wikipedia articles. This paper proposed an application to extract Wikipedia reference information by using regular expression. Following steps are carried out in this study.

3.1. **Data collection.** The dataset used in this paper is not structured data, instead it is semi-structured. Semi-structured data means that the data used do not conform to the formal data structure model related with relational database. However, it still contains tags and indicators to separate and identify semantic elements. This section provides a detailed description of the dataset used. Wikipedia dumps dataset contains all Wikipedia articles and their revisions represented in an XML file. Data are identified using tags and formatted as shown in Figure 1. Similarly, Figure 2 shows the sample reference format in the dump file.

Table 1 connects each tag in the Wikipedia reference dump XML file along with its description.

3.2. **Rules-based matching.** As shown in the previous subheading, the dataset that has been used in this paper is an XML file. The proposed application works as follows. First, the user is asked whether to enter specific subject to search for its related article or to extract all the information from the XML file. In case the user entered a specific

```
<Book>
        <BookTitle>The Craft of Information Visualization Readings and
        Reflections </BookTitle>
        <ISBN>9781558609150</ISBN>
        <AuthorList>
                <Author>Benjamin B. Bederson, Ben Shneiderman</Author>
        </AuthorList>
        <Publisher>Morgan Kaufmann</Publisher>
        <PubYear>2003</PubYear>
        <Description>Interactive Technologies</Description>
</Book>
```

Figure 1. XML format file

```
<page>
<title>YYYY</title>
{{cite book |last=XXXX|first=XXXXXX -link=XXXXXXX|title=XXXXXXXXX|title-link=XXXXXXXXX|year=XXXXX|orig-
year=XXXXXXX|publisher=XXXXXXX|location=XXXXXXXX }}
{{cite journal |last=XXXX| first=XXXXXX -link=XXXXXXX|title=XXXXXXXXX|title-link=XXXXXXXXX|year=XXXXX|orig-
year=XXXXXXX|publisher=XXXXXXX|location=XXXXXXXX }}
</page>
```

Figure 2. XML references format

Table 1. XML file tags and their meaning

Tag	Description
<page></page>	Contain the whole article.
<title></title>	Contain the title of this article.
<revision></revision>	Contain the revision identifier of this article.
<timestamp></timestamp>	Contain the timestamp this revision added.
<contributor></contributor>	Contain the user who added this revision.
<comments> </comments>	Contain any comment the user added to the revision.
<text></text>	Contain the contents of this article.

subject, the proposed work will extract the information like the title, and author name. Then, the result will be classified based on their type. If there were no references for the subject that was entered by the user, it will display "No reference found" message. All the references are presented between the text tag as full sentence starting with "{{" and ending with "}}" and between each element "|" separated them and Figure 2 illustrates the format. To extract the information needed from the tag, the matching methods are listed in Table 2.

4. **Results and Discussion.** In this paper, the result section is divided into sections. The first section is about type of information being extracted while the analyses are performed in the second section. The proposed method will extract Title, Author, URL, Publisher, ISBN, Date, Journal, Page, etc. The second section is where the application will classify the references based on their type that weather it is thesis, book, website, or journal. Figure 3 to Figure 5 show the extracted references information result based on each category reference. Figure 3 represents web reference result where a website

TABLE 2. Matching method for extracting information

Section	Regular expression
Journal	(?:\\{\\{\\s*(?i)cite journal\\b[^|]*|\\G(?!^))(?=[^}]*}})\\\|([^=]+)=([^|}]+)*|\\s"
Book	(?:\\{\\{\\s*(?i)cite book\\b[^|]*|\\G(?!^))(?=[^}]*}})\\\|([^=]+)=([^|}]+)*|\\s"
Web	(?:\\{\\{\\s*(?i)cite web\\b[^|]*|\\G(?!^))(?=[^}]*}})\\\|([^=]+)=([^|}]+)*|\\s"
Thesis	(?:\\{\\{\\s*(?i)cite thesis\\b[^|]*|\\G(?!^))(?=[^}]*}})\\\|([^=]+)=([^|}]+)*|\\s"

```
++  Web Cites
------------------------------------------------------------
++  Web ---->
title = altruism (n .)
url = http://www.etymonline.com/index.php?term=altruism
publisher = Douglas Harper
------------------------------------------------------------

++  Web ---->
title = Biological Altruism
url = http://plato.stanford.edu/entries/altruism-biological/#2
accessdate = 13 May 2011
------------------------------------------------------------
```

FIGURE 3. Web reference result

```
++  Book Cites
------------------------------------------------------------------------
++  Book ---->
title = Political Activists in America: The Identity Construction Model
first1 = Nathan
last1 = Teske
url = https://books.google.com/books?id=7B38I2UEPa0C&dq
publisher = Pennsylvania State University Press
isbn = 9780271035468
page = 101
------------------------------------------------------------
```

FIGURE 4. Book reference result

```
++  Journal Cites
------------------------------------------------------------------------------------
++  Journal ---->
title =  Altruism in medicine: its definition, nature, and dilemmas
first =  David
last =  Steinberg
url =
journal =  Cambridge Quarterly of Healthcare Ethics
------------------------------------------------------------

++  Journal ---->
title = The evolution of reciprocal altruism
first1 = R.L.
last1 = Trivers
url = https://www.academia.edu/1591793
journal = Quarterly Review of Biology
year = 1971
------------------------------------------------------------

++  Journal ---->
title = The evolution of cooperation
journal = [[Science (journal)
------------------------------------------------------------

++  Journal ---->
title = Evolution of indirect reciprocity
url = https://www.academia.edu/17949833
journal = [[Nature (journal)
------------------------------------------------------------
```

FIGURE 5. Journal reference result

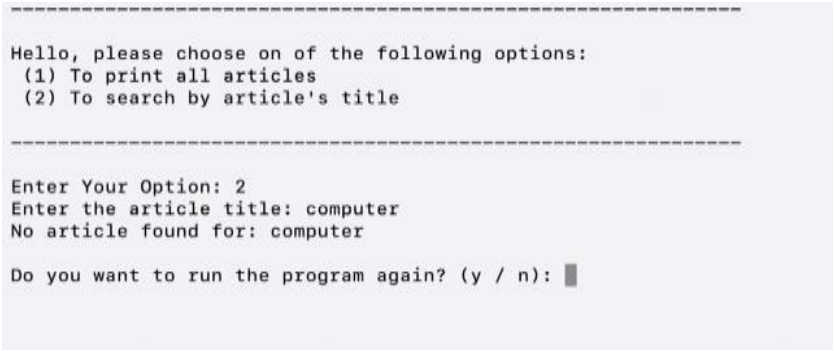

FIGURE 6. Missing reference

was cited in the Wikipedia text. Figure 4 and Figure 5 represent the citations related to book and journals, respectively. Based on this extracted information, the citation is then classified to resultant category of the citation. This not only helps in categorization but also integrity of the reference and date of publishing can be obtained. The title, source URL and publisher information can be used for authentication. Moreover, the proposed scheme will notify if there are no references for the topic that entered by the user for the given Wikipedia page, as shown in Figure 6. This information can be used as the page is relatively new and less information is found, and this page has less reliability in terms of the information contained.

Analysis of the result. To evaluate the proposed methods, several articles pages have been analyzed to extract the references in each page. The regular expression patterns are carefully carved to extract relevant most type of information out the given Wikipedia page. In this experiment, in total, 264 references were successfully extracted and then classified as: Book: 105, Web: 76, Journal: 82, Thesis: 1 reference, respectively.

Moreover, the proposed work was evaluated by considering 80% Wikipedia dataset pages for training and 20% pages for testing purpose. Table 3 shows the accuracy for each information type of extracted information. Relatively low accuracy was observed for the author name that is considered as a challenge in many approaches; however, still 84% is considerable accuracy. The average accuracy was 97.5%. Further, Table 4 shows a comparison between the proposed work in this paper and in [13] where authors tried to extract citation and reference information from two different dataset which are Google Scholar and PubMed. Table 4 shows the accuracy comparison. Moreover, authors in [13] extract only journal references and from the journal authors extracted URL of the reference and number of citations. Table 5 shows another comparison of what type or volume of information was extracted from each paper and the proposed scheme extracts more type and volume of information in contrast to the scheme in [13].

TABLE 3. Accuracy of the proposed approach

Extracted information	Accuracy for training	Accuracy for testing
Title	100%	100%
Author name	84%	80%
Publisher	99%	99%
URL	100%	100%
Journal	100%	100%
ISBN	99%	100%
Access date	100%	100%
Page	100%	100%
Total	98.5%	97.5%

Table 4. Accuracy comparison

Dataset	Accuracy
Google Scholar [13]	44%
PubMed [13]	13%
Proposed work	97.5%

Table 5. Extracted comparison

Section	[13]	Proposed work
Title	No	Yes
Author name	No	Yes
URL	Yes	Yes
Publisher	No	Yes
ISBN	No	Yes
Access date	No	Yes
Journal	No	Yes
Page	No	Yes
No. of citation	Yes	No

5. **Conclusion.** Information extraction provides a better representation of the massive information that could give more meanings of the data. This paper proposed a technique that uses the rule-based approach to extract the required references information from a Wikipedia Reference dump dataset available in the XML files format. The dataset consists of several Wikipedia web pages and 264 references in total, where 105 references were from the book category, 76 references were from web category and 82 references were from the journal category and one from the thesis category. The proposed scheme achieved 97.5% average accuracy for all the references type. For future work, the proposed scheme can be enhanced by including the intelligent approaches such as [31-35]. Moreover, the extracted information (reference) can be further classified based on the reputation and/or indexing of the reference type in say Web of Science and Scopus etc. That will help categorize the Wikipedia page being more or less trusted. Moreover, based on the references the topic can be further classified and/or labeled.

REFERENCES

[1] G. C. Rodi, V. Loreto and F. Tria, Search strategies of Wikipedia readers, *PLoS One*, vol.12, no.2, pp.1-16, 2017.

[2] B. Academic, *Wikipedia*, https://academic-eb-com.library.iau.edu.sa/levels/collegiate/article/Wikipedia/433599, Accessed on February 23, 2020.

[3] C. Keßler, Extracting central places from the link structure in Wikipedia, *Transactions in GIS*, vol.21, no.3, pp.488-502, 2017.

[4] S. Singh, Natural language processing for information extraction, *arXiv.org*, arXiv: 1807.02383, 2018.

[5] A. Gelbukh (Ed.), Computational linguistics and intelligent text processing, *Proceedings of the 5th International Conference, CICLing 2004*, Seoul, Korea, LNCS, vol.2945, 2004.

[6] K. Shaalan, A. E. Hassanien and F. Tolba, Intelligent natural language processing: Trends and applications, *Studies in Computational Intelligence*, vol.740, no.2, pp.380-392, 2018.

[7] M. Singh et al., OCR++: A robust framework for information extraction from scholarly articles, *Proc. of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, Osaka, Japan, pp.3390-3400, 2016.

[8] F. Wu and D. S. Weld, Open information extraction using Wikipedia, *Proc. of the 48th Annual Meeting of the Association for Computational Linguistics*, Uppsala, Sweden, pp.118-127, 2010.

[9] M. Miháltz, Information extraction from Wikipedia using pattern learning, *Acta Cybernetica*, vol.19, no.4, pp.677-694, 2010.

[10] M. Krawczyk, R. Rzepka and K. Araki, Extracting location and creator-related information from Wikipedia-based information-rich taxonomy for ConceptNet expansion, *Knowledge-Based Systems*, vol.108, pp.125-131, 2016.

[11] Y. Fujiwara, Y. Konishi, Y. Suzuki and A. Nadamoto, Extracting lack of information on Wikipedia by comparing multilingual articles, *Proc. of the 14th International Conference on Information Integration and Web-Based Applications and Services*, pp.395-398, 2012.

[12] K. P. Swaraj and D. Manjula, Fast extraction of article titles from XML based large bibliographic datasets, *Procedia Technology*, vol.24, pp.1263-1267, 2016.

[13] S. Liu, C. Chen, K. Ding, B. Wang, K. Xu and Y. Lin, Literature retrieval based on citation context, *Scientometrics*, vol.101, no.2, pp.1293-1307, 2014.

[14] G. Zaman, H. Mahdin, K. Hussain, A. Rahman, J. Abawajy and S. A. Mostafa, An ontological framework for information extraction from diverse scientific sources, *IEEE Access*, vol.9, pp.42111-42124, DOI: 10.1109/ACCESS.2021.3063181, 2021.

[15] G. Zaman, H. Mahdin, K. Hussain and A. Rahman, Information extraction from semi and unstructured data sources: A systematic literature review, *ICIC Express Letters*, vol.14, no.6, pp.593-603, 2020.

[16] G. Zaman, H. Mahdin, K. Hussain, A. Rahman, N. Ibrahim and N. Z. M. Safar, Digital library of online PDF sources: An ETL approach, *International Journal of Computer Science and Network Security*, vol.20, no.11, pp.172-181, 2020.

[17] A. Rahman and F. A. Alhaidari, The digital library and the archiving system for educational institutes, *Pakistan Journal of Information Management and Libraries*, vol.20, no.1, pp.94-117, 2019.

[18] H. Alghamdi, W. Dawwas, T. H. Almutairi and A. Rahman, Extracting ToC and metadata from PDF books: A rule-based approach, *ICIC Express Letters, Part B: Applications*, vol.13, no.2, pp.133-143, 2022.

[19] A. Alamoudi, A. Alomari, S. Alwarthan and Atta-ur-Rahman, A rule-based information extraction approach for extracting metadata from PDF books, *ICIC Express Letters, Part B: Applications*, vol.12, no.2, pp.121-132, 2021.

[20] N. A. Sajid, M. Ahmad, M. T. Afzal and A. Rahman, Exploiting papers' reference's section for multi-label computer science research papers' classification, *Journal of Information and Knowledge Management*, vol.20, no.2, pp.1-21, 2021.

[21] M. Gollapalli, A. Rahman, D. Musleh, N. Ibrahim et al., A neuro-fuzzy approach to road traffic congestion prediction, *Computers, Materials and Continua*, vol.72, no.3, pp.295-310, 2022.

[22] A. Rahman, M. Mahmud, T. Iqbal, L. Saraireh, H. Kholidy et al., Network anomaly detection in 5G networks, *Mathematical Modelling of Engineering Problems*, vol.9, no.2, pp.397-404, 2022.

[23] N. M. Ibrahim, D. G. I. Gabr, A. Rahman, S. Dash and A. Nayyar, A deep learning approach to intelligent fruit identification and family classification, *Multimedia Tools and Applications*, vol.81, no.19, pp.27783-27798, 2022.

[24] S. M. Alotaibi, A. Rahman, M. I. Basheer and M. A. Khan, Ensemble machine learning based identification of pediatric epilepsy, *Computers, Materials and Continua*, vol.68, no.1, pp.149-165, 2021.

[25] A. Rahman, GRBF-NN based ambient aware realtime adaptive communication in DVB-S2, *Journal of Ambient Intelligence and Humanized Computing*, DOI: 10.1007/s12652-020-02174-w, 2020.

[26] A. Rahman, S. Abbas, M. Gollapalli, R. Ahmed, S. Aftab et al., Rainfall prediction system using machine learning fusion for smart cities, *Sensors*, vol.22, no.9, pp.1-15, 2022.

[27] F. Alhaidari, A. Rahman and R. Zagrouba, Cloud of things: Architecture, applications and challenges, *Journal of Ambient Intelligence and Humanized Computing*, DOI: 10.1007/s12652-020-02448-3, 2020.

[28] D. Musleh, R. Ahmed, A. Rahman and F. Al-Haidari, A novel approach to Arabic keyphrase extraction, *ICIC Express Letters, Part B: Applications*, vol.10, no.10, pp.875-884, 2019.

[29] A. Rahman, S. Dash, A. K. Luhach, N. Chilamkurti, S. Baek and Y. Nam, A neuro-fuzzy approach for user behavior classification and prediction, *Journal of Cloud Computing*, vol.8, 17, 2019.

[30] A. Rahman and F. A. Alhaidari, An electronic data interchange framework for educational institutes, *ICIC Express Letters*, vol.13, no.9, pp.831-840, 2019.

[31] A. Rahman, D. Musleh, M. Nabil, H. Alubaidan, M. Gollapalli, G. Krishnasamy et al., Assessment of information extraction techniques, models and systems, *Mathematical Modelling of Engineering Problems*, vol.9, no.3, pp.683-696, 2022.

[32] M. A. A. Khan, M. AlAyat, J. AlGhamdi, S. M. AlOtaibi, M. AlZahrani et al., WeScribe: An intelligent meeting transcriber and analyzer application, in *Proceedings of the 3rd International Conference on Computing, Communications, and Cyber-Security, Lecture Notes in Networks and*

Systems, P. K. Singh, S. T. Wierzchoń, S. Tanwar, J. J. P. C. Rodrigues and M. Ganzha (eds.), vol.421, pp.755-766, Springer, Singapore, https://doi.org/10.1007/978-981-19-1142-2_59, 2022.

[33] A. Rahman, R. N. Asif, K. Sultan, S. A. Alsaif, S. Abbas et al., ECG classification for detecting ECG arrhythmia empowered with deep learning approaches, *Computational Intelligence and Neuroscience*, vol.2022, Article ID 6852845, https://doi.org/10.1155/2022/6852845, 2022.

[34] A. Rahman, M. U. Nasir, M. Gollapalli, M. Zubair, M. A. Saleem, et al., Advance genome disorder prediction model empowered with deep learning, *IEEE Access*, vol.10, pp.70317-70328, DOI: 10.1109/ACCESS.2022.3186998, 2022.

[35] A. Rahman, M. Ahmed, G. Zaman, T. Iqbal, M. A. A. Khan, M. Farooqui et al., Geo-spatial disease clustering for public health decision making, *Informatica*, vol.46, no.6, pp.21-32, 2022.