ICIC Express Letters
Part B: Applications
Volume **14**, Number **10**, October **2023**

ICIC International ©2023 ISSN 2185-2766
pp. 1029–1036

SYSTEM ARCHITECTURE FOR CHEMICAL COMPOUND IDENTIFICATION DATA LC-MS OF MEDICINAL PLANT

Iwan Binanto* and Agung Hernawan

Informatics Department
Sanata Dharma University
Kampus 3, Jl. Paingan, Krodan, Maguwoharjo, Kec. Depok, Kabupaten Sleman, Daerah Istimewa
Yogyakarta 55281, Indonesia
agung.h@usd.ac.id
*Corresponding author: iwan@usd.ac.id

Received January 2023; accepted March 2023

Abstract. *The main problem with supervised learning is data labeling, an activity that seems trivial when the data is small, but not if the data is very large, such as LC-MS (Liquid Chromatography-Mass Spectrometry) data. This task requires high concentration and accuracy if done by humans and impacts processing time. This paper discusses a method to automate labeling of LC-MS data to speed up processing time. In this case, webscraping technique is utilized to retrieve the labels because they are stored in an online database. It has been done in previous studies, but the results are not satisfactory because it still takes a long time to get the required label which is the name of the chemical compound. This is due to frequent disconnections. To solve this problem, a local mirror database is built so that it can be accessed locally. We built two system architectures. The first utilizes two separate computers as a server and client. They are connected to the access point. The second is to utilize a single computer, acting as both server and client at the same time. Theoretically, this will reduce the distance and save labeling time. The system architecture has succeeded in labeling the required data and has a time efficiency of 96.4% and 96.67%, respectively, compared to previous studies. This is a massive time saver.*

Keywords: System architecture, Medicinal plant, Compound identification, Webscraping, Efficiency, Latency

1. **Introduction.** Currently, treatment using medicinal plants has been widely adapted and has shown positive results. Medicinal plants are preferred in many medical systems because they are renewable sources, generally considered safer, and available worldwide. They are the source of thousands of chemicals that have their own functional benefits that make plants one of the medicinal sources of choice in alternative and traditional medicine systems [1]. However, it is necessary to clarify the medicinal chemical content in medicinal plants. There are several ways to do this task, one of them is using Liquid Chromatography-Mass Spectrometry (LC-MS) technology.

Liquid Chromatography-Mass Spectrometry (LC-MS) is widely used, especially in the interpretation or identification of chemical compounds in biological samples [2-6]. Raw data of Liquid Chromatography-Mass Spectrometry (LC-MS) contains millions of data points and there are hundreds to thousands of chromatographic peaks, after peak integration and extraction. These raw data provide highly complex biological samples although only have features: mass per charge (m/z), retention time and intensity [2,7,8].

These features are useless data when they are not interpreted for the identification of the chemical compounds present. Identification requires a lot of precision and time. Therefore, the identification of compounds remains a major obstacle in metabolomics [9].

DOI: 10.24507/icicelb.14.10.1029

This identification can also be called data labeling which is useful for further processing using machine learning because the results are labeled data.

Supervised learning is one of the machine learning methods that requires labeled data. Data labeling is not a difficult thing, but it requires thoroughness, patience, and time-consuming task, especially very large data and is done manually. This is what makes it not a trivial process because of the tension between complexity and simplicity [10,11].

In previous studies, automation has been carried out for identification or data labeling of chemical compounds in the Liquid Chromatography-Mass Spectrometry (LC-MS) data, but there are shortcomings [12,13]. It uses webscraping technique, because there is no availability of API (Application Programming Interface) from the MassBank server. The most prominent weakness is the frequent disconnection from this server [13]. MassBank is an online combined database website which is the official distributed database of the Mass Spectrometry Society of Japan. Data is obtained from each research group which is then distributed on the Internet [14-16].

Improvements are needed from the previous one that utilizes MassBank for data labeling by taking chemical compound names based on mass per charge (m/z) features as input that can be used to identify chemical compound names [13]. The improvement addresses frequent disconnections and slow data fetching when using web scraping techniques. This is the major contribution and innovation of this research.

This paper is categorized as follows: Section 2 describes the related works, Section 3 describes the methods, Section 4 describes the result, and Section 5 focuses on conclusions and future work.

2. Related Works.

Liquid Chromatography-Mass Spectrometry (LC-MS) provides quantitative data that makes a major contribution to biologically and clinically oriented research. Although it still requires highly specialized skills for instrument operation, data acquisition and analysis [17]. Liquid Chromatography-Mass Spectrometry (LC-MS) was used to analyze PPCP (Pharmaceutical and Personal Care Product) samples in the Aquatic ecosystem and has detected very low levels of the chemical [18]. Kharyuk et al. utilized Liquid Chromatography-Mass Spectrometry (LC-MS) to obtain a data set of medicinal plants that was used to train and validate plant species identification algorithms [7,19]. Identification of bacterial species in urine speciment was carried out by Roux-Dalvai et al. by using Liquid Chromatography-Mass Spectrometry (LC-MS) whose data is processed using machine learning [20].

Webscraping is the automation of manual copy-paste jobs from a website. This work is carried out at a computer speed that is super fast compared to human speed [21]. This technique is utilized to get content from websites to analyze certain structured or unstructured data. It was developed in the private sector for business purposes, especially in market analysis, although it is also useful for those seeking specific information [22-24]. Accessing a website means sending requests using HTTP at human speed. Sending requests using HTTP at computer speed, can be problematic as the server will receive many requests in a very short time. It will be considered by the server that someone is attempting a Denial-of-Service attack [25].

In cloud-based applications, latency may lead slow response, performance degradation, and power consumption [26-29]. Managing edge-cloud latency is to minimize the delay by shifting the processing task to numerous smaller clusters located nearer to the end-user devices [28]. Despite significant attempts to enhance network communication and mitigate the effects of network conditions on Machine Learning (ML) applications, there is a need to assess the influence of network latency on their performance, particularly in the context of the irregularities of network conditions in cloud environments [26].

Computer communication on a computer network will experience latency. Latency is determined by the wave propagation through a medium and the nodal process that occurs at the nodes along the router's path. The latency on physical media, whether wired or wireless, is relatively constant, but the in nodal processing latency varies depending on the computational load. Latency causes slow responses, but variations in latency (jitter) can result in unpredictable responses. The larger the network, the greater the problem of latency and jitter. This would not be an issue if the computation is performed on a single computer [30].

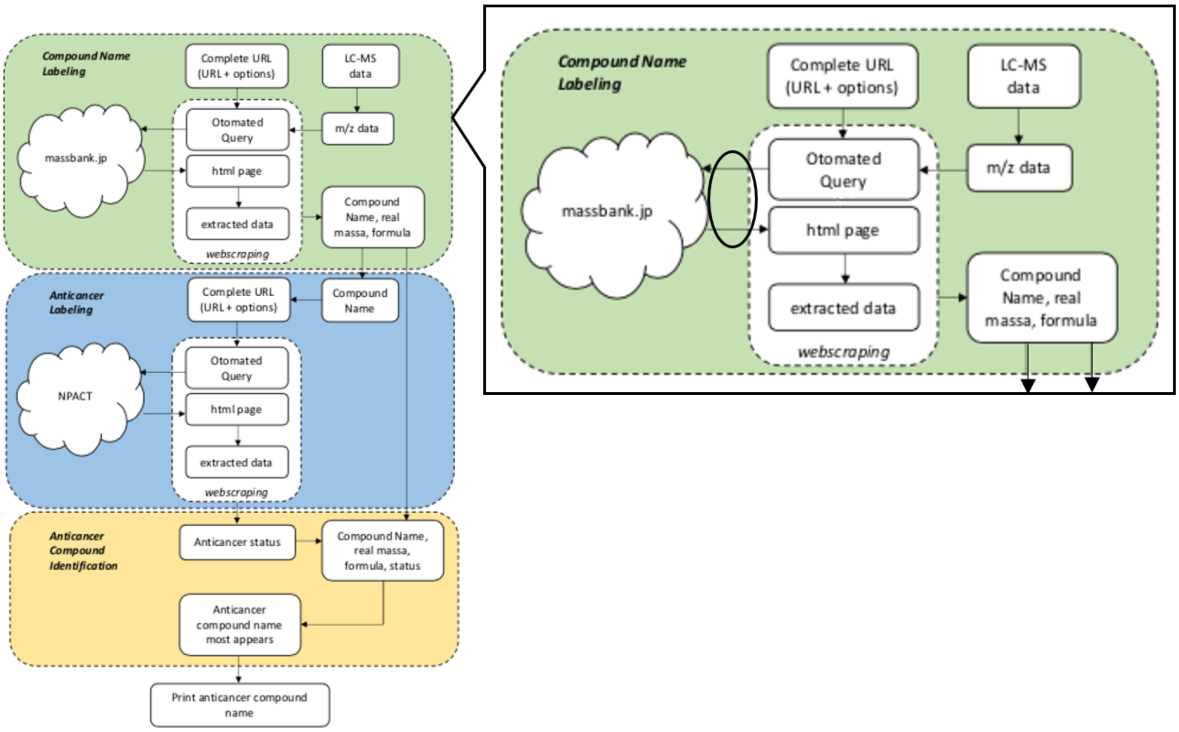
Data labeling is important and time-consuming, especially with large and complex data. Several studies have produced methods and frameworks for labeling data. Sarr et al. utilized deep learning methods for data labeling. It helped human experts refine the essence of echogram data [31].

Many of the most recent published papers in the field of machine learning and Activity Recognition (AR) rely heavily on labeled data sets. For this reason, the synchronization approach using visual key and synchronization using real-time clocks were made to label the obtained data [32].

Our previous study utilized webscraping technique to label Liquid Chromatography-Mass Spectrometry (LC-MS) data and found problems that needed to be solved [12,13]. This paper solved the existing problems.

3. **Methods.** Frequent disconnections and long processing times were the main problems in previous studies. The cause of frequent disconnections is suspected to be due to a bad network or is considered a denial-of-service attack.

In previous study, we developed a tool based on the model as shown in Figure 1(a) [12,13]. This model is connected to an online database, namely MassBank.jp via the Internet. The model consists of 3 stages, namely Compound Name Labeling, Anticancer Labeling, and Anticancer Compound Identification. The first stage, called Compound Name Labeling, takes the longest time because it uses web scraping techniques with a large amount of data, often resulting in frequent disconnections, as it is considered as an attack by the related server. Because the client makes numerous requests at a computer



(a)

(b)

Figure 1. (a) Previous model [12,13]; (b) the first stage of previous model pace that is extremely faster than human speed, it is seen as an attack. This is the main problem of the model. So, in this paper we focus on the first stage, namely Labeling of Compound Names as shown in Figure 1(b).

The direct connection to the online database as circled in black as seen in Figure 1(b), connected to the massbank.jp [14], is a major problem. It causes frequent disconnection as considered a Denial-of-Service attack or bad connection. We cannot control connections via the Internet to the server because they involve many parties and differ geographically. The solution that will be implemented is to build a mirror database so that the first stage of the model is not directly connected to the Internet.

Documentation for building a mirror database has been provided by MassBank [33]. This documentation provides an overview of how the mirror database will be set up. The mirror database is placed as an intermediary to the online database server, and acts like a proxy server as circled in black in Figure 2. This model will synchronize to keep the mirror database up to date. The mirror database will sync when the server is turned on, so for synchronization simply turn the server off and then on again. It takes about 20 minutes.

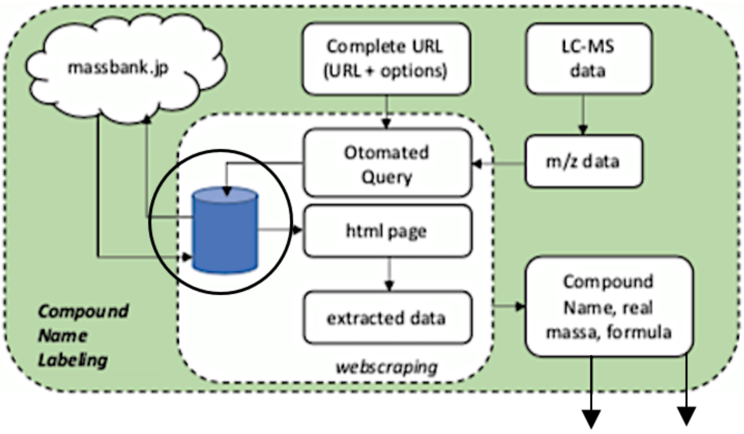


Figure 2. New first stage of model

To implement the new first-stage model, the system architecture is designed by designing the physical connectivity of the computer as a local computer network consisting of one server and one client as shown in Figure 3. It is referred to System Architecture 1.

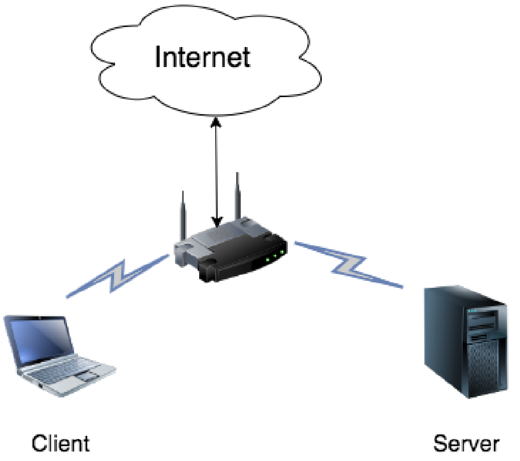


Figure 3. System Architecture 1

In this paper, a local computer network is a network where all computers are located in the same geographic location [34].

In addition, a simple system architecture will be designed with only one computer, as shown in Figure 4, with the single computer acting as both a server and a client at the same time. It is referred to as “System Architecture 2”. This system architecture does not require media to physically connect to other computers. Theoretically, network performance with this system architecture is faster because there is no transmission medium.

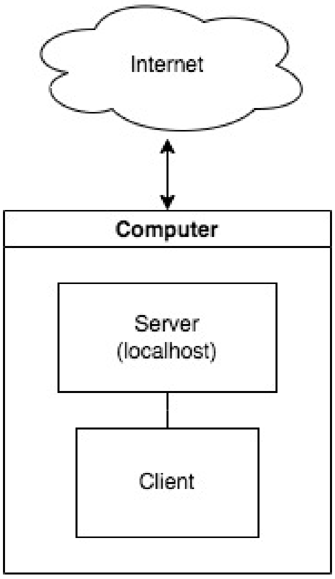


Figure 4. System Architecture 2

Both of these system architectures will connect to the Internet when needed, which is to synchronize with the actual server. These system architectures and mirroring databases are the major contribution of this research.

4. **Results and Discussions.** Both system architectures were successfully built. There are some obstacles, but they can be solved. Likewise, with the old model in the first stage, there were some obstacles too, but they were solved. The first stage model algorithm is shown in Figure 5.

```
Algorithm 1. Labeling the compound name
1: Read the data set only on “m/z” column
2: Repeat until all data is retrieved in the “m/z” column
     Use that data to retrieve data from websites using BeautifulSoup
     Retrieve the desired data in the table with the attributes “treeLayout2”, “width= 142”
     If this attribute is found, then look for the closest value and retrieve data on the name
     and formula of the compound, as well as the actual m/z.
```

Figure 5. First stage algorithm of model

Mirror database is built using a web server, like the original MassBank.jp. All requirements are downloaded from GitHub as provided in the documentation. When everything is configured properly, data retrieval can be done properly.

In this study, ten thousand rows of data were used as in previous studies [12,13]. To provide chemical compound name labels, it took around 3.5 hours with MassBank.jp via Internet.

Utilizing Architecture 1, it takes about 7.5 minutes to get the chemical compound name label, while utilizing Architecture 2, it takes about 7 minutes. The obtained time is generated from the time counter placed within the software that is created. The time counter starts at the beginning of data retrieval and ends when the process is completed.

So in this experiment, the efficiency was 96.4% using Architecture 1 and 96.67% using Architecture 2. This is consistent with the computer network theory that media is a data speed constraint [30]. Especially when utilizing the Internet, where it is not known exactly what media and devices are utilized [26-29].

Utilizing this system architecture – both Architecture 1 and Architecture 2 – makes the first stage of the model very efficient. This is because they utilized one medium only and not many connecting devices like those on the Internet. This affects the overall efficiency of the model because the first stage takes the longest to complete the overall identification process.

5. **Conclusions.** Changes in system architecture as an alternative to improving the performance of the previously developed model were successfully built and used. Although the webserver configuration is a bit constrained due to incomplete documentation, the web scraping technique is still utilized in this research and is useful in retrieving data for labeling chemical compound names. The time efficiency obtained is very large, with 96.4% utilizing Architecture 1 and 96.67% utilizing Architecture 2. It will affect the efficiency of the overall model.

Further researches are to make the second stage efficient by mirroring the NPACT database and utilizing machine learning for this identification.

Acknowledgments. Researchers express many thanks to Dr. Nesti F. Sianipar, SP, M.Sc. and Ms. Khoirunnisa Assidqi, S.Pi., M.Sc. who have supported in previous research and LC-MS data.

REFERENCES

[1] M. M. Babar, N.-S. S. Zaidi, V. R. Pothineni, S. F. Z. Ali, K. R. Hakeem and A. Gul, Application of bioinformatics and system biology in medicinal plant studies, in *Plant Bioinformatics*, K. Hakeem, A. Malik, F. Vardar-Sukan and M. Ozturk (eds.), Cham, Springer, DOI: 10.1007/978-3-319-67156-7, 2017.

[2] M. Brown, D. C. Wedge, R. Goodacre, D. B. Kell, P. N. Baker, L. C. Kenny, M. A. Mamas, L. Neyses and W. B. Dunn, Automated workflows for accurate mass-based putative metabolite identification in LC/MS-derived metabolomic datasets, *Bioinformatics*, vol.27, no.8, pp.1108-1112, DOI: 10.1093/bioinformatics/btr079, 2011.

[3] M. Gerlich and S. Neumann, MetFusion: Integration of compound identification strategies, *Journal of Mass Spectrometry*, vol.48, no.3, pp.291-298, DOI: 10.1002/jms.3123, 2013.

[4] C. Guijas, J. R. Montenegro-Burke, X. Domingo-Almenara, A. Palermo, B. Warth, G. Hermann, G. Koellensperger, T. Huan, W. Uritboonthai, A. E. Aisporna, D. W. Wolan, M. E. Spilker, H. P. Benton and G. Siuzdak, METLIN: A technology platform for identifying knowns and unknowns, *Analytical Chemistry*, vol.90, no.5, pp.3156-3164, DOI: 10.1021/acs.analchem.7b04424, 2018.

[5] B. Zhou, J. F. Xiao, L. Tuli and H. W. Ressom, LC-MS-based metabolomics, *Molecular BioSystems*, vol.8, no.2, pp.470-481, DOI: 10.1039/C1MB05350G, 2012.

[6] J. Listgarten and A. Emili, Statistical and computational methods for comparative proteomic profiling using liquid chromatography-tandem mass spectrometry, *Molecular & Cellular Proteomics*, vol.4, no.4, pp.419-434, DOI: 10.1074/mcp.R500005-MCP200, 2005.

[7] P. Kharyuk, D. Nazarenko, I. Oseledets, I. Rodin, O. Shpigun, A. Tsitsilin and M. Lavrentyev, Employing fingerprinting of medicinal plants by means of LC-MS and machine learning for species identification task, *Scientific Reports*, vol.8, no.1, pp.1-12, DOI: 10.1038/s41598-018-35399-z, 2018.

[8] F. Fernández-Albert, *Machine Learning Methods for the Analysis of Liquid Chromatography-Mass Spectrometry Datasets in Metabolomics*, Ph.D. Thesis, Universitat Polit' Ecnica De Catalunya, 2014.

[9] I. Blaženovi, T. Kind, J. Ji and O. Fiehn, Software tools and approaches for compound identification of LC-MS/MS data in metabolomics, *Metabolites*, vol.8, no.2, DOI: 10.3390/metabo8020031, 2018.

[10] C. M. Tseng, T. W. Huang and T. J. Liu, Data labeling with novel decision module of tri-training, *2020 2nd International Conference on Computer Communication and the Internet (ICCCI2020)*, pp.82-87, DOI: 10.1109/ICCCI49374.2020.9145968, 2020.

[11] R. Cowie, C. Cox, J. C. Martin, A. Batliner, D. Heylen and K. Karpouzis, Issues in data labelling, in *Emotion-Oriented Systems. Cognitive Technologies*, R. Cowie, C. Pelachaud and P. Petta (eds.), Springer, Berlin, Heidelberg, DOI: 10.1007/978-3-642-15184-2_13, 2011.

[12] I. Binanto, H. L. H. S. Warnars, N. F. Sianipar and W. Budiharto, Anticancer compound identification model of Rodent Tuber'S Liquid Chromatography-Mass Spectrometry data, *ICIC Express Letters*, vol.16, no.1, pp.9-16, DOI: 10.24507/icicel.16.01.9, 2022.

[13] I. Binanto, H. L. H. S. Warnars, N. F. Sianipar and W. Budiharto, Webscraping data labeling system on Liquid Chromatography-Mass Spectrometry of Rodent Tuber for efficiency of supervised learning preprocessing, *ICIC Express Letters, Part B: Applications*, vol.13, no.1, pp.107-114, DOI: 10.24507/icicelb.13.01.107, 2022.

[14] H. Horai, M. Arita, S. Kanaya, Y. Nihei, T. Ikeda, K. Suwa, Y. Ojima, K. Tanaka, S. Tanaka, K. Aoshima, Y. Oda, Y. Kakazu, M. Kusano, T. Tohge, F. Matsuda, Y. Sawada, M. Y. Hirai, H. Nakanishi, K. Ikeda, N. Akimoto, T. Maoka, H. Takahashi, T. Ara, N. Sakurai, H. Suzuki, D. Shibata, S. Neumann, T. Iida, K. Tanaka et al., MassBank: A public repository for sharing mass spectral data for life sciences, *Journal of Mass Spectrometry*, vol.45, no.7, pp.703-714, DOI: 10.1002/jms.1777, 2010.

[15] R. Arakawa, H. Adachi, Y. Shida, K. Shiratsuchi, T. Takeuchi, T. Nishioka and Y. Wada, Proposal: Recommendation on measuring and providing mass spectra as chemical information of organic molecules (secondary publication), *Mass Spectrometry*, vol.8, no.1, pp.1-6, DOI: 10.5702/massspectrometry.A0076, 2019.

[16] *MassBank Project*, MassBank | European MassBank (NORMAN MassBank) Mass Spectral DataBase, 2018.

[17] O. T. Schubert, H. L. Rst, B. C. Collins, G. Rosenberger and R. Aebersold, Quantitative proteomics: Challenges and opportunities in basic and applied research, *Nature Protocols*, vol.12, no.7, pp.1289-1294, DOI: 10.1038/nprot.2017.040, 2017.

[18] M. A. Mottaleb, Use of LC-MS and GC-MS methods to measure emerging contaminants pharmaceutical and personal care products (PPCPs) in fish, *Journal of Chromatography & Separation Techniques*, vol.6, no.3, DOI: 10.4172/2157-7064.1000267, 2015.

[19] D. V. Nazarenko, P. V. Kharyuk, I. V. Oseledets, I. A. Rodin and O. A. Shpigun, Machine learning for LC-MS medicinal plants identification, *Chemometrics and Intelligent Laboratory Systems*, vol.156, pp.174-180, DOI: 10.1016/j.chemolab.2016.06.003, 2016.

[20] F. Roux-Dalvai, C. Gotti, M. Leclercq, M.-C. Hélie, M. Boissinot, T. N. Arrey, C. Dauly, F. Fournier, I. Kelly, J. Marcoux, J. Bestman-Smith, M. G. Bergeron and A. Droit, Fast and accurate bacterial species identification in urine specimens using LC-MS/MS mass spectrometry and machine learning, *Molecular & Cellular Proteomics*, vol.5, 2019.

[21] A. V. Saurkar, K. G. Pathare and S. A. Gode, An overview on web scraping techniques and tools, *International Journal on Future Revolution in Computer Science & Communication Engineering*, vol.4, no.4, pp.363-367, 2018.

[22] R. McAlister, Webscraping as an investigation tool to identify potential human trafficking operations in Romania, *Proceedings of the 2015 ACM Web Science Conference*, DOI: 10.1145/2786451.2786510, 2015.

[23] M. Herrmann and L. Hoyden, Applied webscraping in market research, *The 1st International Conference on Advanced Research Methods and Analytics*, Valencia, DOI: 10.4995/carma2016.2016.3131, 2016.

[24] M. Shreesha, S. B. Srikara and R. Manjesh, A novel approach for news extraction using webscraping technique, *The 3rd National Conference on Image Processing, Computing, Communication, Networking and Data Analytics (NCICCNDA2018)*, pp.359-362, DOI: 10.21467/proceedings.1.56, 2018.

[25] F. J. A. P. Mattosinho, *Mining Product Opinions and Reviews on the Web*, Master Thesis, Technische Universitat Dresden, 2010.

[26] D. A. Popescu, N. Zilberman and A. W. Moore, *Characterizing the Impact of Network Latency on Cloud-Based Applications' Performance*, Technical Reports, University of Cambridge Computer Laboratory, 2017.

[27] C. Caiazza, S. Giordano, V. Luconi and A. Vecchio, Edge computing vs centralized cloud: Impact of communication latency on the energy consumption of LTE terminal nodes, *Computer Communications*, vol.194, no.10, pp.213-225, DOI: 10.1016/j.comcom.2022.07.026, 2022.

[28] M. S. Bali and S. Khurana, Effect of latency on network and end user domains in cloud computing, *Proc. of the 2013 International Conference on Green Computing, Communication and Conservation of Energy (ICGCE2013)*, pp.777-782, DOI: 10.1109/ICGCE.2013.6823539, 2013.

[29] L. Bulej, T. Bureš, A. Filandr, P. Hnìtynka, I. Hnìtynková, J. Pacovský, G. Sandor and I. Gerostathopoulos, Managing latency in edge-cloud environment, *Journal of Systems and Software*, vol.172, 110872, DOI: 10.1016/j.jss.2020.110872, 2021.

[30] B. A. Forouzan, *Data Communications and Networking with TCPIP Protocol Suite*, 6th Edition, McGraw-Hill, 2022.

[31] J. M. A. Sarr, T. Brochier, P. Brehmer, Y. Perrot, A. Bah, A. Sarré, M. A. Jeyid, M. Sidibeh and S. El Ayoubi, Complex data labeling with deep learning methods: Lessons from fisheries acoustics, *ISA Transactions*, DOI: 10.1016/j.isatra.2020.09.018, 2020.

[32] J. W. Kamminga, M. Jones, K. Seppi, N. Meratnia and P. J. M. Havinga, Synchronization between sensors and cameras in movement data labeling frameworks, *Proceedings of the 2nd ACM Workshop on Data Acquisition to Analysis (DATA2019), Part of SenSys 2019*, no.11, pp.37-39, DOI: 10.1145/3359427.3361920, 2019.

[33] *Installation of MassBank | MassBank-Documentation*, https://massbank.github.io/MassBank-documentation/developer_documentation.html, 2022.

[34] J. M. McCoy, S. L. French, R. Abnous and M. J. Niccolai, A local computer network simulation, *Proc. of the 12th SIGCSE Technical Symposium on Computer Science Education (SIGCSE'81)*, pp.263-267, 1981.