Available online at http://scik.org

Commun. Math. Biol. Neurosci. 2024, 2024:38

https://doi.org/10.28919/cmbn/8462

ISSN: 2052-2541

INDONESIAN COLORECTAL CANCER CONSORTIUM DATA SCIENCE SYSTEM

BENS PARDAMEAN[1,*], ARIF BUDIARTO[1], ALAM AHMAD HIDAYAT[1], NUR ADHIANTI HERYANTO[1], CARISSA IKKA PARDAMEAN[1], JAMES WILLIAM BAURLEY[1,2]

[1]Bioinformatics and Data Science Research Center, Bina Nusantara University, Jakarta 11480, Indonesia

[2]BioRealm, Culver City, CA 90230, United States

Abstract: Cancer registry has become an important part of cancer research. It helps researchers and clinicians track data related to cancer incidences in hospitals and communities. Indonesian Colorectal Cancer Consortium (IC3) is an initiative founded by Hasanuddin University and Bina Nusantara University to establish a collaboration among universities and hospitals throughout Indonesia in conducting colorectal cancer research. A data science system has been created to support this research as implementation and expansion of a site-specific cancer registry. SCRUM, as the most common agile software development technique, was used as the approach to building this system. In the beginning, a data dictionary was created based on the medical records and questionnaires filled by clinicians following patient answers. This web system was built using the PHP CodeIgniter framework. The overall system consists of two main parts: (1) data storage and (2) data analysis that allows anyone involved in the consortium to manage and analyze the data. All data are stored in the PostgreSQL database and can be accessed by RStudio Server as the analysis tool. This system was successfully implemented in the pilot research project with 355 samples. The implemented web-

*Corresponding author

E-mail address: bpardamean@binus.edu

Received January 26, 2024

based system can help researchers and clinicians store and analyze cancer-related data collected in Indonesia.

Keywords: cancer; cancer registry; data science; information system; web application.

2020 AMS Subject Classification: 92B15.

1. INTRODUCTION

Colorectal cancer is among the top ten most frequently diagnosed cancers in Indonesia [1] and remains one of the most fatal cancers in the world [2]. The number of colorectal cancer incidences worldwide increased significantly above the age of 50 years, leaving 3% of the cases for patients under 40 years [3], which is quite different compared with the trend in Indonesia. Young patients (under 40 years) constitute more than 30% of the cases, exhibiting more progressive disease and subpar response towards chemotherapy treatment [3]. The prevalence number of colorectal cancer cases per 100,000 people is about 3.15 for females and 4.13 for males [4]. This number is relatively low compared to countries with higher human development index, such as Australia, New Zealand, and Western Europe. However, since Indonesia is the fourth most populous country in the world with more than 270 million people, then the prevalence number becomes more significant [5].

One of the contributing factors that increase the figure of colorectal cases within a population is unhealthy habits [2,6]. Further, various genetic factors play an important role in elevating cancer risk as a lot of well-established studies have suggested that an individual is likely to develop cancer if there is a cancer history in his/her family [7–10]. Extensive studies are required to understand the role of lifestyle and genetic factors in the progression of colorectal cancer cases within diverse populations [7,10–16]. The data collected from the studies can be employed to build statistical models such as polygenic models to predict the risk of cancer incidences using both genetic data and phenotype data [11,17–19]. This implies the importance of colorectal cancer data science systems such as cancer registry as a tool to collect and analyze cancer-related data [20]. A systematic cancer registry can help clinicians and researchers classify cases, evaluate risk factors, and learn the effect of treatments by studying patient's medical data and genetic profiles. Moreover, it can be used to develop a preventive program planned for an individual with a high risk of colorectal cancer.

Hasanuddin University along with Bina Nusantara University has established the Indonesian Colorectal Consortium (IC3) focusing on colorectal cancer-related studies in Indonesia. This consortium is intended to collect and analyze colorectal-related data throughout hospitals in Indonesia for gaining insights in prevention and treatment schema. As a pilot project, clinical data and blood samples from 355 respondents were collected in a case-control design study. These data include medical records from hospitals, personal data directly gathered from the respondents using questionnaires, and genetic data from DNA sequencing results. Therefore, in this work, a systematic web-based query and analytic system to store and analyze all data collected from IC3 studies was developed. The ultimate goal of this system is to make a personalized follow-up action plan based on the clinical and genetic data and to perform as a tool to do epidemiological research [8].

2. Information Technology for Cancer Registry

A cancer registry is used to maintain a reporting system for cancer incidences, as a reference to investigate the causes of cancer and to help clinicians in planning appropriate treatments for patients. Also, a cancer registry can be categorized as a part of a surveillance system [21]. The development of a web-based registry system allows centralized database management and statistical tools that can be accessed by other researchers and clinicians. Such web applications also have been implemented and assessed in various types of previous studies, for examples: a stroke prevention system [22], online learning applications for early disease detection [20,23,24], health wearable data analysis [25], electronic health record sharing [26], child growth and malnutrition monitoring system [27], information system for rice genetics [28], and cloud computing [29].

There are two general types of cancer registry [21,30]: hospital-based registries and population-based registries. A hospital-based cancer registry collects cancer data from selected hospitals, in which the data can be used to assess the effectiveness of diagnosis and treatments [30]. Information collected in this registry is mainly related to diagnosis, stage distribution, treatment methods, and survival [21]. On the other hand, a population-based cancer registry collects cancer information

based on geographical factors [30]. It is usually focused on the incidence cases and trends [21]. Each geographical area can have its cancer registry system as its consideration to build reactive and protective planning from a population point of view. These data are usually gathered by census and interviews with the patients [30].

In Indonesia, a population-based cancer registry has been developed by Diponegoro University since 1970 in the city of Semarang [31]. However, due to a bureaucratic issue at the government level, the development of a cancer registry was stopped. The Ministry of Health restarted a program in 2007 to develop a cancer registry with a more sustainable approach. The program was only implemented in Jakarta as a hospital-based cancer registry but then expanded to be a population-based cancer registry. Later, two cancer hospitals in Indonesia (i.e., National Cancer Center and the Hospital for Cancer Registration Center) successfully developed a cancer registration system called CanReg5 [32,33].

The current development of cancer registries is only focused on phenotype factors to assess and control cancer incidences. Recently, several studies have included genetic testing to assess cancer risks. A colorectal cancer assessment that integrates genetic testing demonstrates more accurate results in detecting patients with colorectal cancer cases [34]. The main advantage of genetic testing in a cancer predisposition study is the unique characteristic of genes [35], allowing researchers and clinicians to develop more personalized treatments for patients. Since genetic testing is considered to be a complex analysis, the combination with cancer registries requires an efficient and robust approach to accommodate all issues.

3. METHODOLOGY

A web-based data science system that combines the features from the cancer registry and genetic testing was developed. A web-based approach was selected to give easy access to everyone involved in the colorectal cancer study. The whole development process of this web-based application was based on the Agile Software Development approach [36]. It is the most common approach used in developing a computer system. The most important part of it, compared to the earlier approach such as the Waterfall technique, is the capability to adapt to constant changes

during the software development process [37–40]. More specifically, SCRUM, as the most popular technique in the agile software development approach, was used as the guideline in building this web system [41].

The major consideration in developing the system was accessibility and usability. In respect of the accessibility aspect, a web-based system was selected to ensure that everyone involved in this research can access the system via the Internet. This system includes features to store, analyze, and report colorectal cancer data. The entire workflow of this system can be seen in Fig. 1.

The system was mainly developed using the PHP CodeIgniter framework while PostgreSQL was used to build the database. In order to provide a statistical analysis feature, this system was connected to the Rstudio server. It allows the system to perform statistical analysis, then displays its result to the users on the system dashboard. These software architectures were deployed in an HTTP Apache web server.

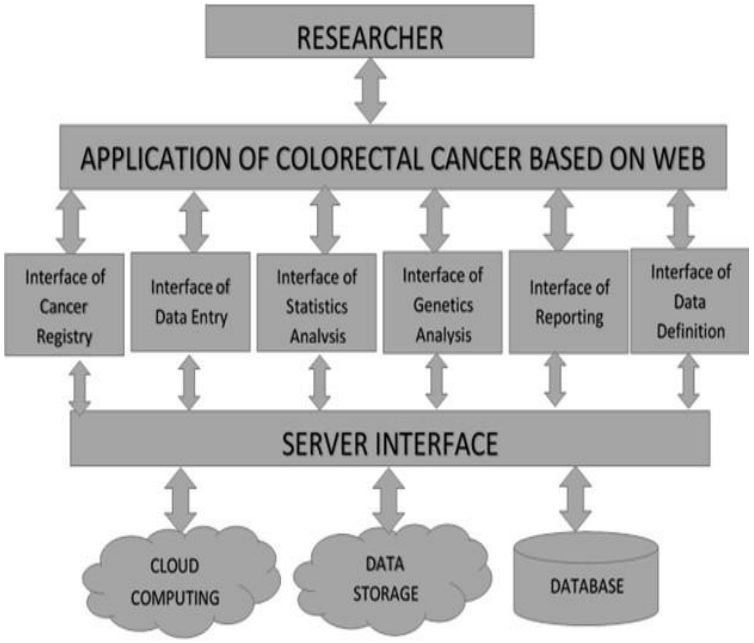

FIGURE 1. System Workflow

4. RESULTS AND DISCUSSION

The system has been implemented in pilot research at Hasanuddin University in Makassar, Indonesia. The study was a case-control study to assess colorectal cancer incidences in one specific

geographic area aiming to compare both clinical data and genetics data between a control case and a group case. The respondents in the case group were diagnosed with colorectal cancer while the control group consisted of respondents who did not have colorectal cancer. This can provide an example of combining population-based cancer registries and hospital-based cancer registries because the data collected in this study were focused on both clinical and geographic-related data. The system developed in the study enables clinicians and researchers to identify specific factors and biomarkers related to colorectal cancer. The final output is a model that can help clinicians plan personalized medical treatment based on genetic data. The data contain 355 samples which consist of 162 samples in the case group and 193 samples in the control group.

The data were extracted from questionnaires that contained medical records filled by the clinicians and personal data filled based on respondents' answers. The number of questions for both groups was different because some questions were not relevant to the control group which implied a different number of variables for each group dataset. The case dataset had 386 variables while the control dataset had 323 variables.

A data dictionary was created to help in understanding the dataset. All variables in the dataset were divided into 25 sections based on the questions' context. These sections include Patient Identity, Anamnesis, Menstruation and Reproductive History, Family History of Cancer, and more, which are detailed in Table 1.

The creation of the data dictionary is part of the requirements gathering at the beginning of the development process. Based on this dictionary, then a database was constructed for data storage and retrieval of the system. A relational database was selected because of its ease of deployment but at the same time, it can manage large-scale data. In total, there are 6 tables in the database consisting of 4 dataset-related tables and 2 user-related tables. The session table can store activity logs from all users so that it can be used to track any changes in the database. A level attribute in the user table was used to specify the user's read and write permission.

All respondents' data were stored in the CRC case and the CRC control table based on the respondent's group. These tables were then associated with the dataset table so it could be used

for analytic purposes through the analytic table. The illustration of this entity relationship diagram can be seen in Fig. 2. This web-based system enabled everyone involved in the research to store and access the data as well as to perform statistical analysis.

TABLE 1. Data Dictionary

Section	Description	Number of Variables
Patient Identity	Demographic information	31
Anamnesis	Clinical diagnosis	60
Menstruation and Reproduction History	Information related to menstruation and reproduction and only applicable for female respondents.	23
Family History of Cancer	Information related to cancer history of respondents family including father, mother, brother, sister, grandparents, and any other relatives.	52
Eating Habits	Information related to eating habits for several types of foods, such as vegetables, meat, spicy food, etc.	19
History of alcohol and coffee consumption	Information related to alcohol and coffee habits including the frequency, type of beverages, etc.	8
Smoking habits	Smoking habits information including frequency (cigarettes per day), smoking-related illness, etc.	11
Exercise	Information related to exercise habits	1
Physical Examination	Medical record focusing on physical examination	7
Laboratory Examination	Information related to laboratory examination including hemoglobin, leukocyte, etc.	12
Supporting Examination	Other supporting examination	9
Tumor Marker	Tumor identification	3
Diagnosis of Pre Surgery	Pre-surgery diagnosis by a clinician	1
Surgery Action	Whether a surgery has been taken	1
Type of Surgery Action	Type of surgery action has been taken	1
Surgery Time	Time of surgery action	1
Diagnosis Post Surgery	Post-surgery diagnosis by a clinician	1
Tumor Location	Location of tumor	1
Staging Post Surgery	Tumor staging after surgery	3
Anatomical Pathology	Anatomical pathology identification	3
Molecular Biology	Molecular biology identification	6
Durante Surgery Finding	Finding during surgery	1
Complications	Disease complication information	1
Chemotherapy	Chemotherapy history	1
Radiotherapy	Radiotherapy history	1

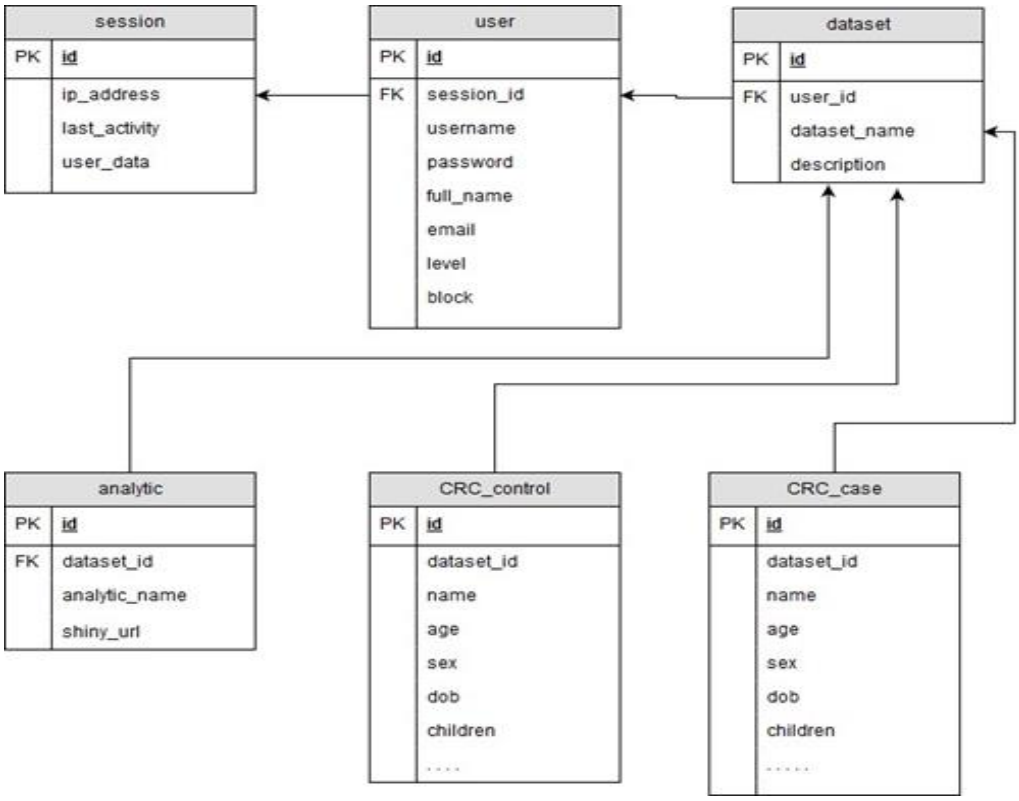

FIGURE 2. Entity Relationship Diagram

Users need to go to the login page to access the entire system. It can be used to differentiate the user's role to specify the permission level for each user. The role was defined by the level attribute in the user table. This level determines the extent to which users can add, edit, or delete data. There is no signup feature to allow someone to create an account in this system. The administrator is the only one who can create the user and set the role level to prevent intruders from accessing the system due to the confidential and sensitive nature of medical records.

There are two main parts of the data science system that can be accessed by the users. The first part is the dashboard that displays statistical analysis results. The majority of the analyses are descriptive statistical analyses that solely depend on the type of data used. The dashboard page can be viewed in Fig. 3. The second part of the system is the data table which is illustrated in Fig. 4. This part is divided into case and control groups based on the study design. In each group, the users can view all respondents' detailed data. Only users with a specific role can edit the records. The users also can download the data into CSV files if needed.

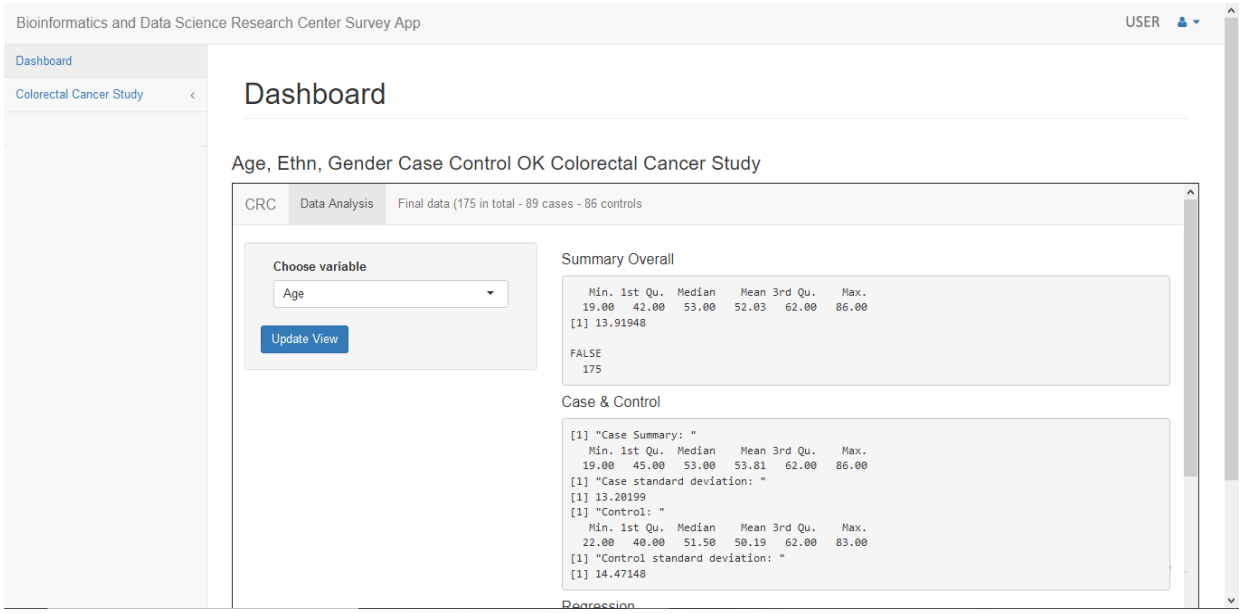

FIGURE 3. Dashboard Page

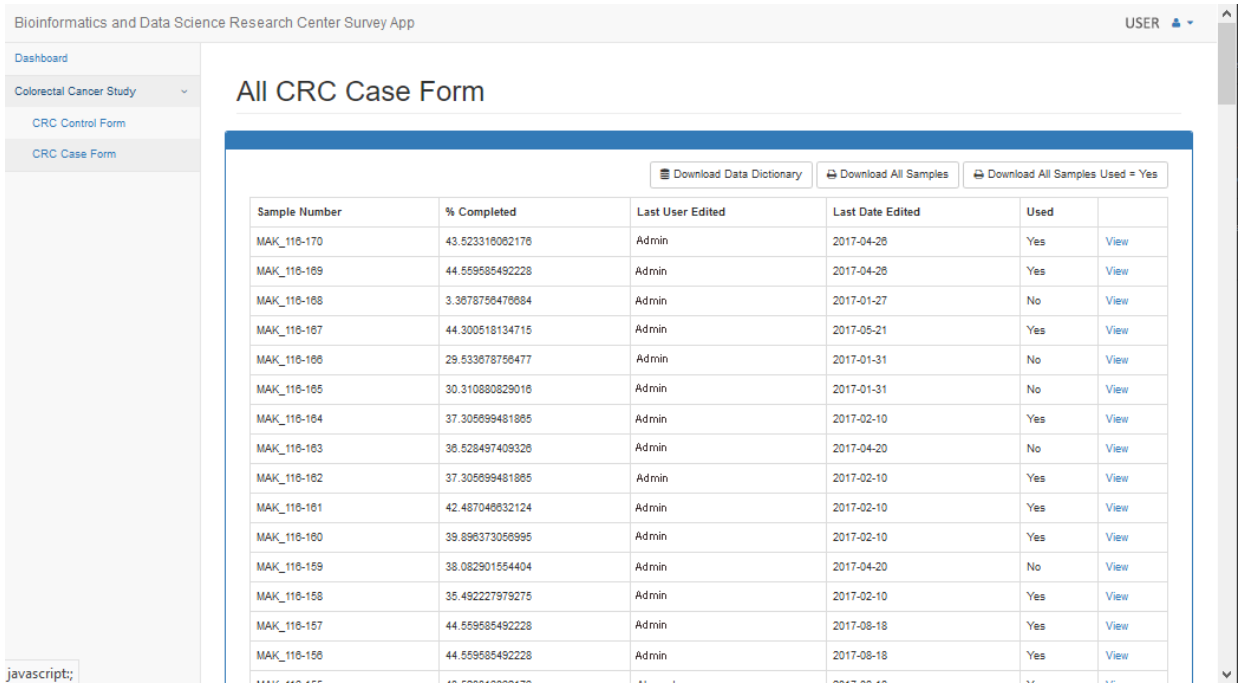

FIGURE 4. Data Table Page

5. CONCLUSION

The web-based data science system has been developed to help clinicians and researchers maintain all data related to colorectal cancer incidences that include clinical and genetic data. This

system was successfully implemented in the pilot research project with 355 samples. The data extracted from questionnaires can be recorded easily to the system for further use and analysis. Researchers can also get the statistical results from the system because it was connected to the Rstudio server. This system was built to allow researchers to be involved in every stage of research, starting from the data preparation to the reporting stage. Since the genetics data have not been recorded into the system, it was impossible to measure the system's performance to do a more sophisticated and complex statistical analysis.

Some useful feedbacks were also gathered from researchers and clinicians as the users of the system. Mainly, their feedbacks were focused on the data input mechanism since the input process was exhaustive and contained a large number of variables to be inputted into the system. For future improvement, these issues will be put as the main focus. These issues will potentially become the most significant barrier since this system will be implemented for multi-center research. A better user interface (UI) and user experience (UX) must be developed to provide a better system flow for the users. It also will avoid any data failure caused by data input errors.

CONFLICT OF INTERESTS

The authors declare that there is no conflict of interests.

REFERENCES

[1] D. Khairina, E. Suzanna, D. Triana, et al. Profile of Colorectal Cancer in 14 Provinces in Indonesia, J. Glob. Oncol. 4 (2018), 66s–66s. https://doi.org/10.1200/jgo.18.64300.

[2] H. Sung, J. Ferlay, R.L. Siegel, et al. Global cancer statistics 2020: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries, CA: Cancer J Clinicians 71 (2021), 209–249. https://doi.org/10.3322/caac.21660.

[3] A.W. Sudoyo, B. Hernowo, E. Krisnuhoni, et al. Colorectal cancer among young native Indonesians: A clinicopathological and molecular assessment on microsatellite instability, Med. J. Indones. 19 (2010), 245-251. https://doi.org/10.13181/mji.v19i4.411.

[4] M. Wahidin, R. Noviani, S. Hermawan, et al. Population-based cancer registration in Indonesia, Asian Pac. J.

Cancer Prevent. 13 (2012), 1709-1710. https://doi.org/10.7314/apjcp.2012.13.4.1709.

[5] M. Abdullah, A.W. Sudoyo, A.R. Utomo, et al. Molecular profile of colorectal cancer in Indonesia: is there another pathway? Gastroenterol. Hepatol. Bed Bench. 5 (2012), 71-78.

[6] B. Benarba, Red and processed meat and risk of colorectal cancer: An update, EXCLI J. 17 (2018), 792-797. https://doi.org/10.17179/excli2018-1554.

[7] B. Pardamean, J.W. Baurley, C.I. Pardamean, J.C. Figueiredo, Changing colorectal cancer trends in Asians, Int. J. Colorectal Dis. 31 (2016), 1537-1538. https://doi.org/10.1007/s00384-016-2564-z.

[8] V.H. Roos, C. Mangas-Sanjuan, M. Rodriguez-Girondo, et al. Effects of family history on relative and absolute risks for colorectal cancer: a systematic review and meta-analysis, Clinic. Gastroenterol. Hepatol. 17 (2019), 2657-2667.e9. https://doi.org/10.1016/j.cgh.2019.09.007.

[9] F. Kastrinos, N.J. Samadder, R.W. Burt, Use of family history and genetic testing to determine risk of colorectal cancer, Gastroenterology. 158 (2020), 389-403. https://doi.org/10.1053/j.gastro.2019.11.029.

[10] C.I. Pardamean, D. Sudigyo, A. Budiarto, et al. Changing colorectal cancer trends in Asians: Epidemiology and risk factors, Oncol. Rev. 17 (2023), 10576. https://doi.org/10.3389/or.2023.10576.

[11] T.W. Cenggoro, B. Mahesworo, A. Budiarto, et al. Features importance in classification models for colorectal cancer cases phenotype in Indonesia, Procedia Computer Sci. 157 (2019), 313-320. https://doi.org/10.1016/j.procs.2019.08.172.

[12] M.C. Wong, H. Ding, J. Wang, et al. Prevalence and risk factors of colorectal cancer in Asia, Intest. Res. 17 (2019), 317-329. https://doi.org/10.5217/ir.2019.00021.

[13] N. Keum, E. Giovannucci, Global burden of colorectal cancer: emerging trends, risk factors and prevention strategies, Nat. Rev. Gastroenterol. Hepatol. 16 (2019), 713-732. https://doi.org/10.1038/s41575-019-0189-8.

[14] V. Gausman, D. Dornblaser, S. Anand, et al. Risk factors associated with early-onset colorectal cancer, Clinic. Gastroenterol. Hepatol. 18 (2020), 2752-2759.e2. https://doi.org/10.1016/j.cgh.2019.10.009.

[15] I. Yusuf, B. Pardamean, J.W. Baurley, et al. Genetic risk factors for colorectal cancer in multiethnic Indonesians, Sci. Rep. 11 (2021), 9988. https://doi.org/10.1038/s41598-021-88805-4.

[16] C. Fernandez-Rozadilla, M. Timofeeva, Z. Chen, et al. Deciphering colorectal cancer genetics through multi-omic analysis of 100,204 cases and 154,587 controls of European and east Asian ancestries. Nat. Genet. 55 (2023), 89-

99. https://doi.org/10.1038/s41588-022-01222-9.

[17] R.E. Caraka, N.T. Nugroho, S.K. Tai, et al. Feature importance of the aortic anatomy on endovascular aneurysm repair (EVAR) using Boruta and Bayesian MCMC, Commun. Math. Biol. Neurosci. 2020 (2020), 22. https://doi.org/10.28919/cmbn/4584.

[18] R.E. Caraka, S. Shohaimi, I.D. Kurniawan, et al. Ecological show cave and wild cave: negative binomial Gllvm's arthropod community modelling, Procedia Computer Sci. 135 (2018), 377-384. https://doi.org/10.1016/j.procs.2018.08.188.

[19] I.D. Kurniawan, R.C.H. Soesilohadi, C. Rahmadi, et al. The difference on Arthropod communities' structure within show caves and wild caves in Gunungsewu Karst area, Indonesia, Ecol. Environ. Conserv. 24 (2018), 72-81.

[20] H.H. Muljo, A.S. Perbangsa, Y. Yulius, et al. Mobile learning for early detection of cancer, Int. J. Interact. Mob. Technol. 12 (2018), 39-53. https://doi.org/10.3991/ijim.v12i2.7814.

[21] P.E. Petersen, Oral cancer prevention and control - The approach of the World Health Organization, Oral Oncol. 45 (2009), 454-460. https://doi.org/10.1016/j.oraloncology.2008.05.023.

[22] Anindito, B. Pardamean, R. Christian, et al. Expert-system based medical stroke prevention, J. Computer Sci. 9 (2013), 1099-105. https://doi.org/10.3844/jcssp.2013.1099.1105.

[23] H.H. Muljo, B. Pardamean, A.S. Perbangsa, The implementation of online learning for early detection of cervical cancer, J. Computer Sci. 13 (2017), 600-607. https://doi.org/10.3844/jcssp.2017.600.607.

[24] H.H. Muljo, A.S. Perbangsa, Y. Lie, et al. Improving early cancer detection knowledge through mobile learning application, Int. J. Online Biomed. Eng. 15 (2019), 60. https://doi.org/10.3991/ijoe.v15i02.9678.

[25] A. Budiarto, T. Febriana, T. Suparyanto, et al. Health assistant wearable-based data science system model: A pilot study, in: 2018 International Conference on Information Management and Technology (ICIMTech), IEEE, Jakarta, 2018: pp. 438-442. https://doi.org/10.1109/ICIMTech.2018.8528102.

[26] M.F. Kacamarga, A. Budiarto, B. Pardamean, A platform for electronic health record sharing in environments with scarce resource using cloud computing, Int. J. Online Biomed. Eng. 16 (2020), 63. https://doi.org/10.3991/ijoe.v16i09.13187.

[27] R. Rahutomo, I. Nurlaila, A.S. Perbangsa, et al. Database management system design with time series

modification for child growth and malnutrition monitoring in the regency of Serdang Bedagai, in: 2020 International Conference on Information Management and Technology (ICIMTech), IEEE, Bandung, Indonesia, 2020: pp. 306-311. https://doi.org/10.1109/ICIMTech50083.2020.9211170.

[28] J. Baurley, A. Perbangsa, A. Subagyo, et al. A web application and database for agriculture genetic diversity and association studies, Int. J. Bio-Sci. Bio-Technol. 5 (2013), 33-42. https://doi.org/10.14257/ijbsbt.2013.5.6.04.

[29] M.F. Kacamarga, B. Pardamean, H. Wijaya, Lightweight virtualization in cloud computing for research, Commun. Computer Inform. Sci. (2015), 439-445. https://doi.org/10.1007/978-3-662-46742-8_40.

[30] P. Mohan, H.A. Lando, Cancer registries in oral cancer control in India, J. Cancer Policy. 4 (2015), 13-14. https://doi.org/10.1016/j.jcpo.2015.05.006.

[31] Sarjadi, P. Trihartini, Cancer registration in Indonesia, Asian Pac. J. Cancer Prevent. 2 (2001), 21-24.

[32] B. Pardamean, T. Suparyanto, D.R. Fadilah, CANREG 5 networks for Indonesia, in: 2015 2nd International Conference on Information Technology, Computer, and Electrical Engineering (ICITACEE), IEEE, Semarang, Indonesia, 2015: pp. 26–30. https://doi.org/10.1109/ICITACEE.2015.7437764.

[33] B. Pardamean, T. Suparyanto, Hospital-based cancer registry application, in: 2017 International Conference on Information Management and Technology (ICIMTech), IEEE, Yogyakarta, 2017: pp. 44-48. https://doi.org/10.1109/ICIMTech.2017.8273509.

[34] T.F. Imperiale, D.F. Ransohoff, S.H. Itzkowitz, et al. Multitarget stool DNA testing for colorectal-cancer screening, N. Engl. J. Med. 370 (2014), 1287-1297. https://doi.org/10.1056/NEJMoa1311194.

[35] S.M. Domchek, A. Bradbury, J.E. Garber, et al. Multiplex genetic testing for cancer susceptibility: Out on the high wire without a net?, J. Clinic. Oncol. 31 (2013), 1267-1270. https://doi.org/10.1200/jco.2012.46.9403.

[36] P. Abrahamsson, O. Salo, J. Ronkainen, et al. Agile software development methods: Review and analysis, Espoo, Finland, Technical Research Centre of Finland, VTT Publications 478, (2002).

[37] Y.B. Leau, W.K. Loo, W.Y. Tham, et al. Software Development life cycle AGILE vs traditional approaches, in: International Conference on Information and Network Technology, 2012, pp. 162-167.

[38] K. Muchtar, F. Rahman, T.W. Cenggoro, et al. An improved version of texture-based foreground segmentation: Block-based adaptive segmenter, Procedia Computer Sci. 135 (2018), 579-586. https://doi.org/10.1016/j.procs.2018.08.228.

[39] N.Dominic, Daniel, T.W. Cenggoro, et al. Transfer learning using inception-ResNet-v2 model to the augmented neuroimages data for autism spectrum disorder classification, Commun. Math. Biol. Neurosci. 2021 (2021), 39. https://doi.org/10.28919/cmbn/5565.

[40] R.E. Caraka, M. Tahmid, R.M. Putra, et al. Analysis of plant pattern using water balance and cimogram based on oldeman climate type, IOP Conf. Ser.: Earth Environ. Sci. 195 (2018), 012001. https://doi.org/10.1088/1755-1315/195/1/012001.

[41] K. Schwaber, M. Beedle, Agile software development with Scrum, Prentice Hall, Upper Saddle River, New Jersey, 2002.