

Database System for Storing Tuberculosis Sputum Sample Images as an AI Training Dataset

<https://doi.org/10.3991/ijoe.v18i15.28245>

Hery Harjono Muljo^(✉), Anzaludin Samsinga Perbangsa, Tjeng Wawan Cenggoro,
Kartika Purwandari, Digdo Sudigyo, Bens Pardamean
Bina Nusantara University, Jakarta, Indonesia
heryhm@binus.edu

Abstract—The high prevalence of tuberculosis (TB) in Indonesia puts Indonesia in the second-highest national TB prevalence in the world after India. This high prevalence can cause a failure to deliver medical treatments to TB patients, which is exacerbated by the disproportionate distribution of doctors in Indonesia. To address this issue, an artificial intelligence (AI) system is necessary to help doctors in screening a large number of patients in a short time. However, to develop a robust AI for this purpose, we need a large dataset. This study aims to develop a database system for storing TB sputum sample images, which can be used as the dataset to train an AI system for TB detection. The developed system can help doctors and health workers to manage the images during their daily job.

Keywords—tuberculosis, artificial intelligence, dataset, database, sputum sample image

1 Introduction

Tuberculosis (TB) cases in Indonesia occupy the second-highest incident in the world after India. Based on WHO in 2019, the estimation of TB patients in the Indonesian population was 845 thousand as the best estimate and estimated interval of 770 to 923 thousand. According to the Tuberculosis Report, Indonesia is one of 30 nations having a High Burden of Tuberculosis (HBC) in terms of global TB incidence [1]. As a result, tuberculosis (TB) has become a significant public health problem in Indonesia. The Indonesian Ministry of Health highlighted the problem by reporting a reduction in the rate of TB recovery with successful treatment every year, especially from 2008 to 2017 [2]. One of the main factors affecting the high TB cases in the Indonesian population is the country's unequal access to infrastructure. Furthermore, in Indonesia, patients' unwillingness to check for tuberculosis symptoms and the waiting list for a diagnosis from a few lung specialists are crucial issues. Indonesia's vast geographical area and the distribution of doctors and pulmonary specialists are not proportional to the population. Consequently, to reduce TB illness incidence and early detection, more facilities and infrastructure are required, including the use of advanced technology, such as an artificial intelligence system [3].

In the health industry, artificial intelligence is rapidly improving in terms of recognizing and detecting diseases. Medical photographs on x-ray or rontgen [4]–[6], endoscopy [7], [8], ECG [9], and sample preparations from the patient’s body [10]–[12] have all been used to develop and study disease recognition through medical diagnostic images. Along with the photo of a sputum sample preparation in patients with suspected TB, the Artificial Intelligence (AI) system can assist medical workers, particularly doctors, in the screening and early diagnosis of tuberculosis. However, a large amount of data is required to test and measure the accuracy of the artificial intelligence system. At least thousands of data samples are required for high accuracy when using artificial intelligence systems [13], [14]. Otherwise, the system’s application can be employed conveniently by doctors in the form of applications. A database application is required to gather and store images of sputum samples in TB patients before advancing on to the development stage of the TB screening application. Due to the laborious work needed to manage the images manually, it is necessary to develop a database system that can help doctors and health workers with image file management. Therefore, we developed a database system that is aimed to store sputum sample images of TB patients. The application is developed using Laravel by implementing the concept of Create Update Delete (CRUD), which enables a quick and easy process of image file management [5], [15], [16].

2 Previous studies

A standardized system for recording and reporting TB patient data from diagnosis to treatment outcome was implemented by the World Health Organization (WHO) in the mid-1990s [17]. WHO DOTS (directly observed treatment, short-course) strategy, as well as its successor, the Stop TB Strategy [18], includes this information system, which contains geographic, administrative, and quality control details. The WHO-recommended reporting system was being used in 90% of nations by the mid-2000s, so virtually all TB cases were documented and reported there. This system has allowed TB data to be compared across thousands of treatment facilities around the world. Although it requires a significant amount of data to collect and report data about TB symptoms and diagnosis, this process remains a costly data-intensive process. A patient must maintain a few times a week (sometimes daily) regimen of taking anti-TB medicines. The duration of treatment can last several months (or even years). Applications utilize database servers to provide data to users without the need for users to access database tables or queries even if they have direct access to the data files. Reporting and recording systems made for the web require both types of servers. With proper configuration and professional management, servers can provide a secure foundation for running information systems used by thousands of people at once [19].

Not only for TB but database systems are also utilized in many cases in the health sector. For instance, it serves as the backend system of mobile learning applications to learn early cancer detection [20], [21], health education [22], and child growth

monitoring [23]. In particular, for early cancer detection learning, the database system has been proved to be useful for the end-users [24]. With a similar purpose to our study, Cenggoro and Pardamean have developed a database system to collect pap smear images as AI training data [25].

Meanwhile, for AI to diagnose tuberculosis, various methods have been used, including the use of blood tests, skin tests, interferon-gamma release assays, fluorescent microscopy, culturing bacteria, polymerase chain reaction, GeneXpert test, nucleic acid amplification, chest X-ray, and sputum smear test [26]. Conventional methods take a long time and are less accurate, so they require the support of Artificial Intelligence (AI) technology to overcome them [27]. With the increasing speed of computing technology, it is possible to identify abnormalities caused by tuberculosis from chest x-rays (CXRs) images using AI called Computer-Aided Detection (CAD). However, most CAD studies tended to develop rather than clinical evaluations [28].

AI can identify tuberculosis using machine learning (ML) or deep learning (DL) [29]. ML uses an algorithm that does not depend on human decisions to analyze dominant variables [30]. While DL models the brain architecture that can learn more precise data representations as the data used increases [31]. Many studies have found algorithms that strive to improve the accuracy of TB detection results even with all their limitations.

The two-stage classification method (TSCM) using a convolutional neural network (CNN) with transfer learning can detect TB with an accuracy of 98% [32]. A study in Pakistan using an artificial neural network (ANN) approach can detect TB with 94% accuracy [33].

3 System design

We used diagrams from unified modeling language (UML) to design the TBC image recording system and an entity-relationship diagram to design the database. Subsections 3.1, 3.2, 3.3, and 3.4 of the documents elaborate on the use case, class, activity, and entity-relationship diagram of the system.

3.1 Use case diagram

In Figure 1, the four development use cases are illustrated: Show Image Data, Upload Image Data, Delete Image Data, and Edit Image Data. Each use case for the system is represented by a web page of the system since the system is developed as a web application. Specifically, the Uploaders are the only members of the system who are able to access these four use cases to provide annotations of the TBC images.

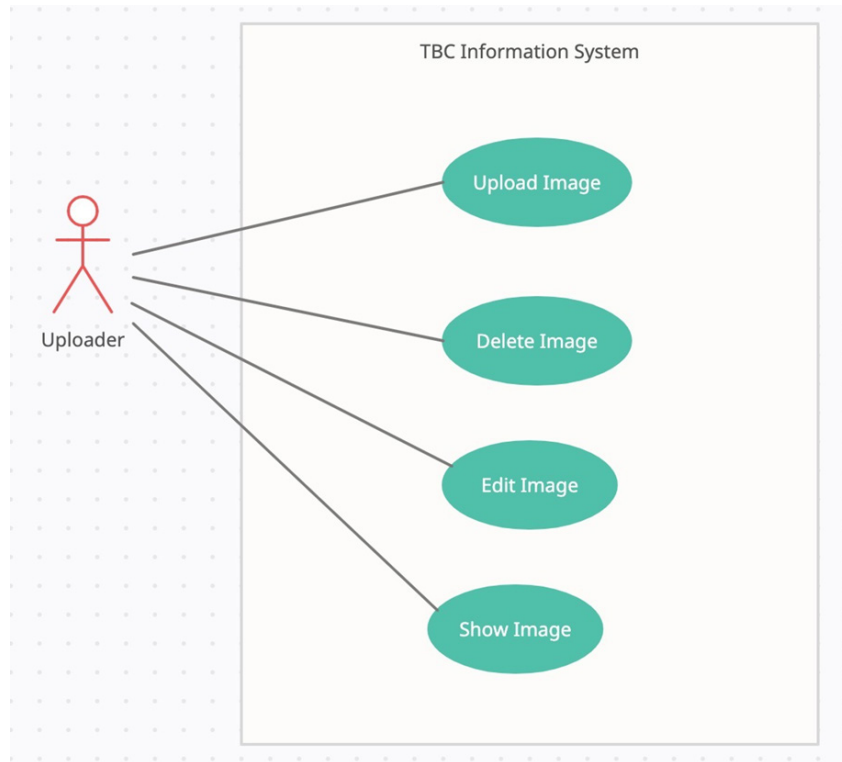


Fig. 1. Use case diagram

3.2 Class diagram

Figure 2 illustrates the class diagram for the system design. The class diagram for this TB information system has four instances, namely TB images, detailed information from images, image storage hospital information, and uploading TB image data. There is also a relation for each class that shows that one instance is connected to another instance. Each TB picture can be obtained from various hospitals, and vice versa, each hospital has many TB samples. Therefore, the relationship can be interpreted as many to many. Then each picture of TBC has different detailed information, so the relationship can be interpreted as many to one. while uploading images can be operated only once for several images, so the relationship can be interpreted as one to many.

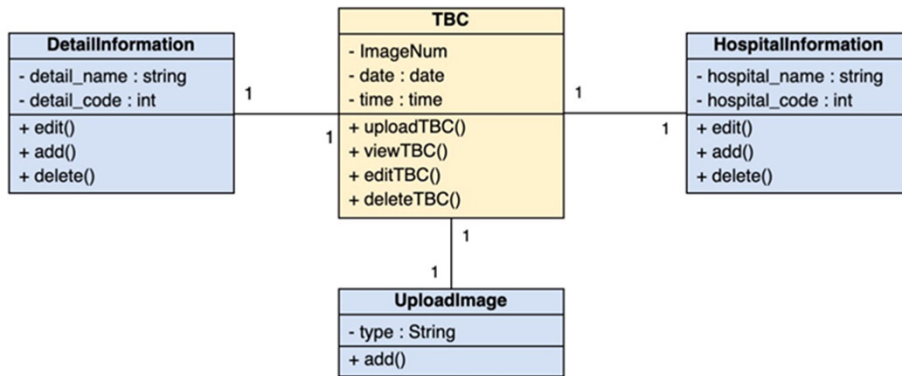


Fig. 2. Class diagram

3.3 Activity diagram

This activity diagram is divided into three diagrams for a more concise explanation. Each diagram describes the logical flow of the Upload Image Data, Edit Image Data, and Delete Image Data use cases. They are illustrated in Figures 3a, 3b, and 3c. According to Figure 3a, the uploading of images starts the flow of the Upload Image Data phase.

The main page displays all the recorded images available in the system. An uploader is encouraged to upload TBC images and provide information on the data of the images whose sample is captured. In the process of uploading images and entering data, the images and data are stored in the database, and the system is redirected to the home page afterward.

Upon selecting the image to edit, the uploader can view the edit page. An uploader can manage images by selecting one, uploading new images, deleting old images, and editing image data. Before the user is redirected to the main page, any updates to images or information data will be saved to the database.

In order to delete an image, the uploader must click the delete button next to the image. After this process is complete, the image is deleted from the database, and the system is redirected to the main page.

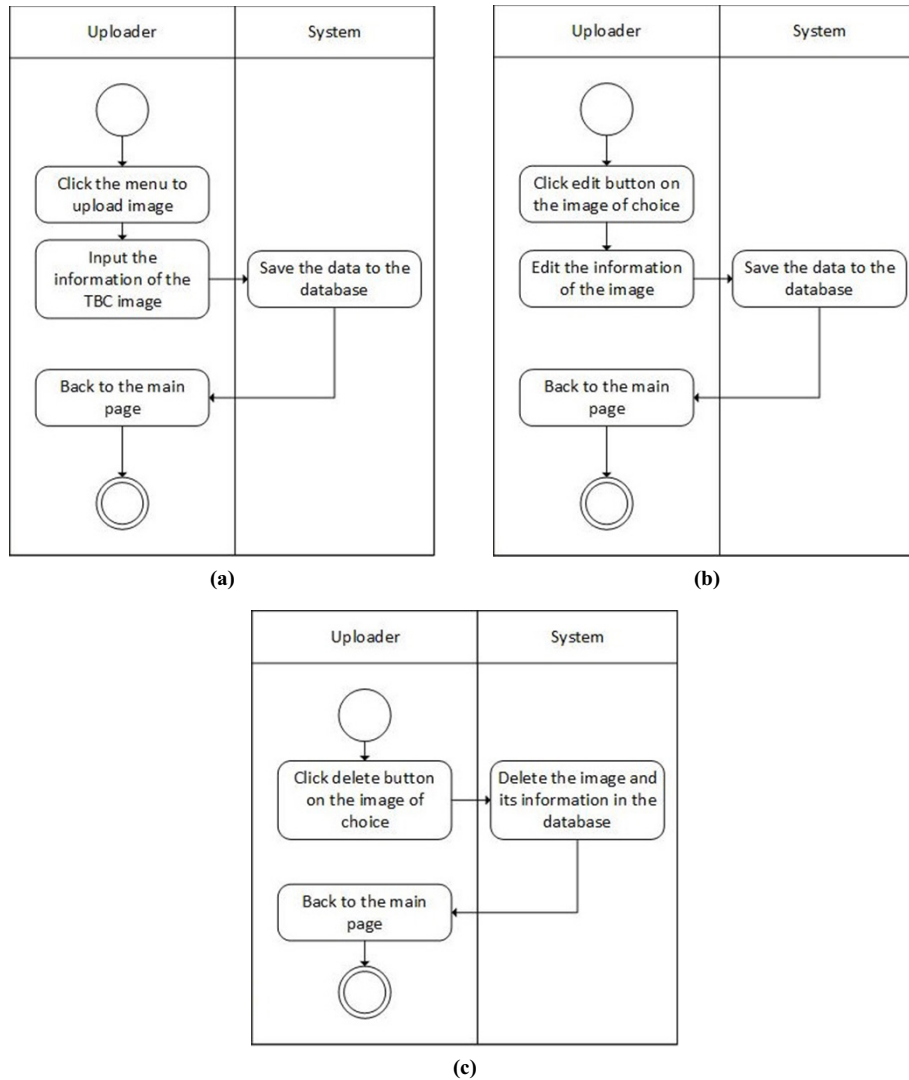


Fig. 3. Activity diagram

3.4 Entity-relationship diagram

The developed system uses four tables: TBC image data, image location, image uploader user data, and TBC image album. The TB image table is used to store uploaded TB images. While the location table is used to store images of the TB sample results. All collections of images are grouped into album tables. And the picture of TBC itself is a table containing each ID of the three related tables, namely album ID, location ID, and

user/member ID. As depicted in Figure 4, one album, one location, and one user can have zero or more images, whereas a single image must have one and only one album, location, and uploading user.

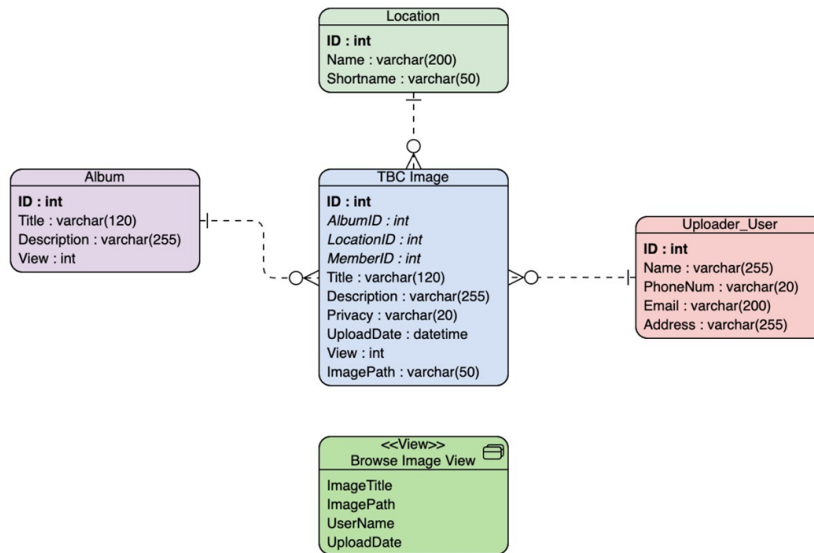


Fig. 4. Entity-relationship diagram

4 Implementation results

When the application runs for the first time, the display in the browser shows a display as shown in Figure 5. In that view, there is an “Upload Project Image” button with a camera icon that serves to open a page to upload a photo image of a TB sputum sample which will be entered into the database. When the button is pressed, a page is displayed as shown in Figure 6.

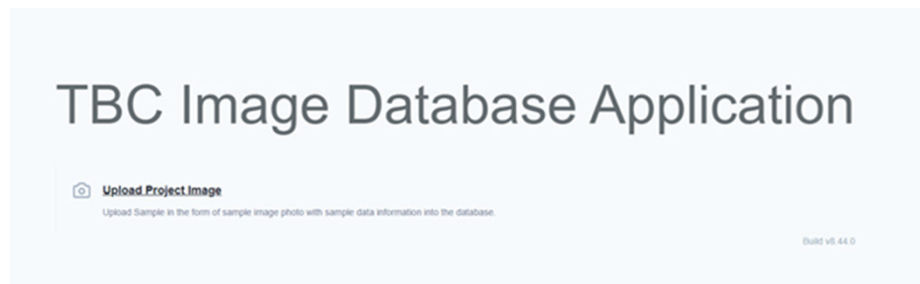


Fig. 5. Main page

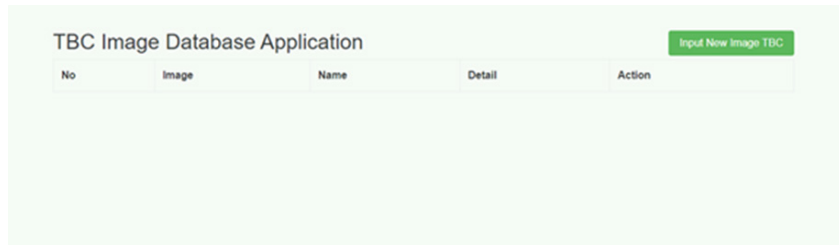


Fig. 6. List of images page (blank)

Opening the advanced page from the “Upload Project Image” option in the application and then pressing the “Input New Image Sample” button will open the file upload page that is displayed in Figure 7. On this page, we can fill in the information of the name, the detailed information of the sputum sample, and upload the image of the sputum sample. When the “Browse” button is pressed on the page, a window for selecting still images is displayed as shown in Figure 8. When the image has been selected, the data can be saved in the system by clicking the “Submit” button.

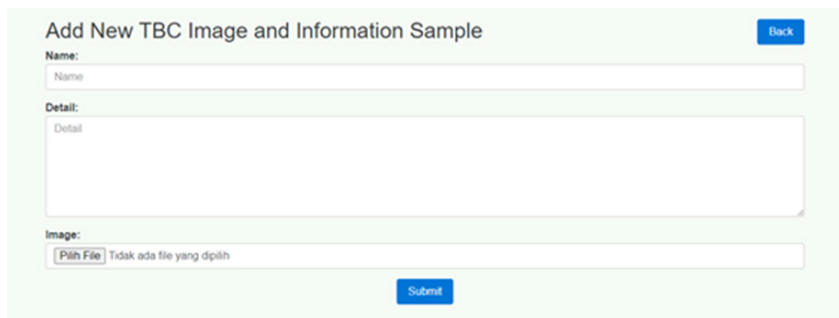


Fig. 7. Upload image page

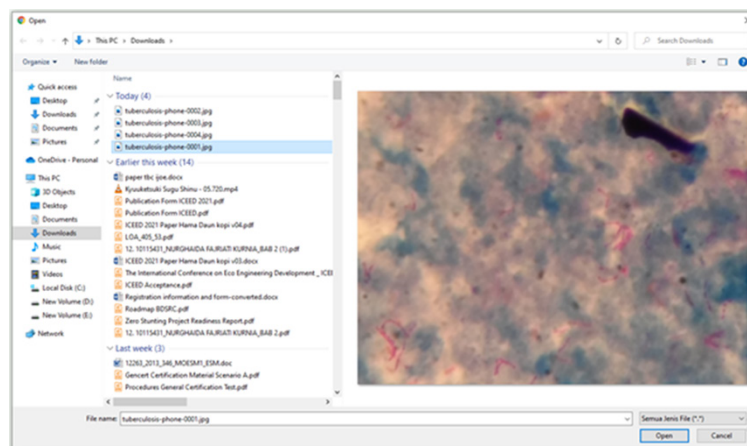


Fig. 8. Image selection window

After pressing the submit button, the file upload results page is displayed as seen in Figure 9. The uploaded results can be deleted via the red Erase button. After the deletion process, a notification will appear as displayed in Figure 10. The data can also be edited by clicking the dark blue Edit button. The edit page that appears after clicking Edit has the same display as in Figure 7. Similar to the deletion process, a notification also appears after the editing process, as displayed in Figure 11.



The screenshot shows the 'TBC Image Database Application' interface. At the top right is a green button labeled 'Input New Image TBC'. Below it is a green notification bar that says 'Data created successfully'. The main content is a table with the following data:



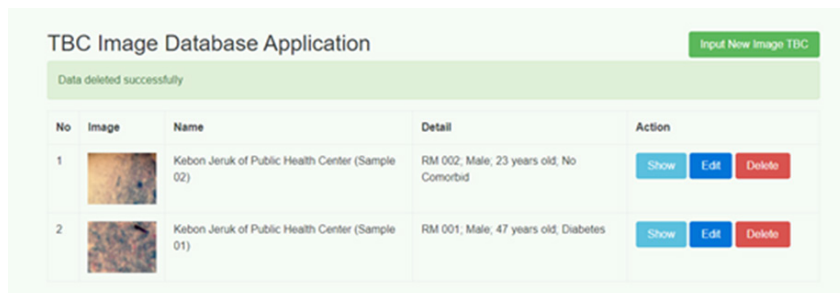
No	Image	Name	Detail	Action
1		Kebon Jeruk of Public Health Center (Sample 03)	RM 003; Male; 35 years old; Psoriasis arthritis	Show Edit Delete
2		Kebon Jeruk of Public Health Center (Sample 02)	RM 002; Male; 23 years old; No Comorbid	Show Edit Delete
3		Kebon Jeruk of Public Health Center (Sample 01)	RM 001; Male; 47 years old; Diabetes	Show Edit Delete

Fig. 9. List of images page after image uploading



The screenshot shows the 'TBC Image Database Application' interface. At the top right is a green button labeled 'Input New Image TBC'. Below it is a green notification bar that says 'Data deleted successfully'. The main content is a table with the following data:



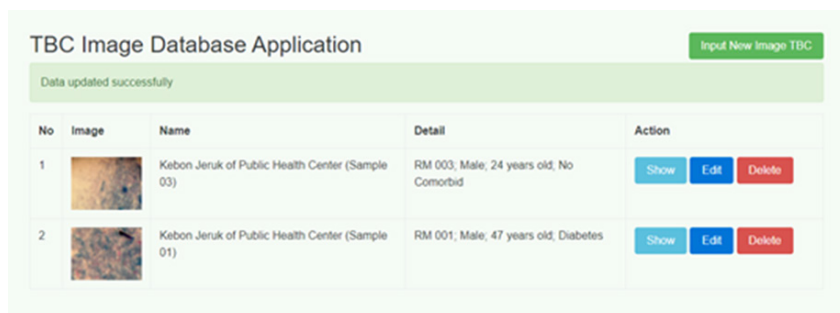
No	Image	Name	Detail	Action
1		Kebon Jeruk of Public Health Center (Sample 02)	RM 002; Male; 23 years old; No Comorbid	Show Edit Delete
2		Kebon Jeruk of Public Health Center (Sample 01)	RM 001; Male; 47 years old; Diabetes	Show Edit Delete

Fig. 10. List of images page after deletion



The screenshot shows the 'TBC Image Database Application' interface. At the top right is a green button labeled 'Input New Image TBC'. Below it is a green notification bar that says 'Data updated successfully'. The main content is a table with the following data:



No	Image	Name	Detail	Action
1		Kebon Jeruk of Public Health Center (Sample 03)	RM 003; Male; 24 years old; No Comorbid	Show Edit Delete
2		Kebon Jeruk of Public Health Center (Sample 01)	RM 001; Male; 47 years old; Diabetes	Show Edit Delete

Fig. 11. List of images page after data editing

5 Conclusion

In this paper, the development process for a database system for storing TB sputum sample images is presented. This system allows doctors and health workers to efficiently manage the images of TB sputum samples, which in effect creates a large dataset that can be used to train AI system to speed up the diagnostic process of TB. The future study following this study can use the training data stored in this system to train a robust AI.

6 Acknowledgement

This study was funded by a Competitive Grant of the Directorate of Research and Community Services of the Republic of Indonesia's Ministry of Research, Technology, and Higher Education, grant agreement number: 309/E4.1/AK.04.PT/2021.

7 References

- [1] World Health Organization. (2020). Global tuberculosis report 2020: Executive summary.
- [2] Lega, A. D., Megatsari, H., & Devy, S. R. (2021). Ecological analysis of success of TB treatment and its related factors in Indonesia in 2019. *Medico-Legal Update*, 21(4).
- [3] Kurniawan, R. (Ed.). (2019). Profil kesehatan Indonesia tahun 2018. Kementerian Kesehatan RI.
- [4] Owais, M., Arsalan, M., Choi, J., & Park, K. R. (2019). Effective diagnosis and treatment through content-based medical image retrieval (CBMIR) by using artificial intelligence. *Journal of Clinical Medicine*, 8(4): 462. <https://doi.org/10.3390/jcm8040462>
- [5] Orozco, C. I., Xamena, E., Martínez, C. A., & Rodríguez, D. A. (2021). COVID-XR: A web management platform for coronavirus detection on X-ray chest images. *IEEE Latin America Transactions*, 19(6): 1033–1040. <https://doi.org/10.1109/TLA.2021.9451249>
- [6] Punn, N. S., & Agarwal, S. (2021). Automated diagnosis of COVID-19 with limited posteroanterior chest X-ray images using fine-tuned deep neural networks. *Applied Intelligence*, 51(5): 2689–2702. <https://doi.org/10.1007/s10489-020-01900-3>
- [7] Choi, J., Shin, K., Jung, J., Bae, H. J., Kim, D. H., Byeon, J. S., & Kim, N. (2020). Convolutional neural network technology in endoscopic imaging: Artificial intelligence for endoscopy. *Clinical Endoscopy*, 53(2): 117. <https://doi.org/10.5946/ce.2020.054>
- [8] Wang, Y., Feng, Z., Song, L., Liu, X., & Liu, S. (2021). Multiclassification of endoscopic colonoscopy images based on deep transfer learning. *Computational and Mathematical Methods in Medicine*. <https://doi.org/10.1155/2021/2485934>
- [9] Attia, Z. I., Noseworthy, P. A., Lopez-Jimenez, F., Asirvatham, S. J., Deshmukh, A. J., Gersh, B. J., ... & Friedman, P. A. (2019). An artificial intelligence-enabled ECG algorithm for the identification of patients with atrial fibrillation during sinus rhythm: A retrospective analysis of outcome prediction. *The Lancet*, 394(10201): 861–867. [https://doi.org/10.1016/S0140-6736\(19\)31721-0](https://doi.org/10.1016/S0140-6736(19)31721-0)
- [10] Khutlang, R., Krishnan, S., Dendere, R., Whitelaw, A., Veropoulos, K., Learmonth, G., & Douglas, T. S. (2009). Classification of mycobacterium tuberculosis in images of ZN-stained sputum smears. *IEEE Transactions on Information Technology in Biomedicine*, 14(4): 949–957. <https://doi.org/10.1109/TITB.2009.2028339>

- [11] Panicker, R. O., Kalmady, K. S., Rajan, J., & Sabu, M. K. (2018). Automatic detection of tuberculosis bacilli from microscopic sputum smear images using deep learning methods. *Biocybernetics and Biomedical Engineering*, 38(3): 691–699. <https://doi.org/10.1016/j.bbe.2018.05.007>
- [12] Devi, K. R., Pradhan, J., Bhutia, R., Dadul, P., Sarkar, A., Gohain, N., & Narain, K. (2021). Molecular diversity of mycobacterium tuberculosis complex in Sikkim, India and prediction of dominant spoligotypes using artificial intelligence. *Scientific Reports*, 11(1): 1–16. <https://doi.org/10.1038/s41598-021-86626-z>
- [13] Benke, K., & Benke, G. (2018). Artificial intelligence and big data in public health. *International Journal of Environmental Research and Public Health*, 15(12): 2796. <https://doi.org/10.3390/ijerph15122796>
- [14] Panch, T., Pearson-Stuttard, J., Greaves, F., & Atun, R. (2019). Artificial intelligence: Opportunities and risks for public health. *The Lancet Digital Health*, 1(1): e13–e14. [https://doi.org/10.1016/S2589-7500\(19\)30002-0](https://doi.org/10.1016/S2589-7500(19)30002-0)
- [15] Dangar, H. (2013). *Learning laravel 4 application development*. Packt Publishing.
- [16] Armel, J. (2014). *Web application development with Laravel PHP Framework version 4*.
- [17] Raviglione, M. C. (2003). The TB epidemic from 1992 to 2002. *Tuberculosis*, 83(1–3): 4–14. [https://doi.org/10.1016/S1472-9792\(02\)00071-9](https://doi.org/10.1016/S1472-9792(02)00071-9)
- [18] Piot, A., & Chaulet, P. (2008). *Implementing the WHO stop TB strategy: A handbook for national TB control programmes*. World Health Organization.
- [19] World Health Organization. (2021). *National tuberculosis prevalence surveys 2007–2016*.
- [20] Muljo, H. H., Perbangsa, A. S., Yulius, & Pardamean, B. (2018). Mobile Learning for Early Detection of Cancer. *International Journal of Interactive Mobile Technology*, 12(2): 39–53. <https://doi.org/10.3991/ijim.v12i2.7814>
- [21] Muljo, H. H., Perbangsa, A. S., Yulius, & Pardamean, B. (2019). Improving early cancer detection knowledge through mobile learning application. *International Journal of Online and Biomedical Engineering*, 15(2): 60–70. <https://doi.org/10.3991/ijoe.v15i02.9678>
- [22] Muljo, H. H., Perbangsa, A. S., & Pardamean, B. (2019). Assessment of online learning application for health education. *International Journal of Online and Biomedical Engineering*, 15(12): 69–80. <https://doi.org/10.3991/ijoe.v15i12.11157>
- [23] Rahutomo, R., Nurlaila, I., Perbangsa, A. S., & Pardamean, B. (2020). Database management system design with time series modification for child growth and malnutrition monitoring in the regency of serdang bedagai. *International Conference on Information Management and Technology*, 306–311. <https://doi.org/10.1109/ICIMTech50083.2020.9211170>
- [24] Muljo, H. H., Pardamean, B., Perbangsa, A. S., Purwandari, K., Mahesworo, B., Hidayat, A. A., & Cenggoro, T. W. (2020). TAM as a model to understand the intention of using a mobile-based cancer early detection learning application. *International Journal of Online and Biomedical Engineering*, 16(2): 80–93. <https://doi.org/10.3991/ijoe.v16i02.12609>
- [25] Cenggoro, T. W., & Pardamean, B. (2021). PapSmear image recording system for artificial intelligence data collection. *IOP Conference Series: Earth and Environmental Science*, 794(1): 12109. <https://doi.org/10.1088/1755-1315/794/1/012109>
- [26] Xiong, Y., Ba, X., Hou, A., Zhang, K., Chen, L., & Li, T. (2018). Automatic detection of mycobacterium tuberculosis using artificial intelligence. *Journal of Thoracic Disease*, 10(3): 1936. <https://doi.org/10.21037/jtd.2018.01.91>
- [27] Qin, Z. Z., Sander, M. S., Rai, B., Titahong, C. N., Sudrungrot, S., Laah, S. N., ... & Creswell, J. (2019). Using artificial intelligence to read chest radiographs for tuberculosis detection: A multi-site evaluation of the diagnostic accuracy of three deep learning systems. *Scientific Reports*, 9(1): 1–10. <https://doi.org/10.1038/s41598-019-51503-3>

- [28] Harris, M., Qi, A., Jeagal, L., Torabi, N., Menzies, D., Korobitsyn, A., ... & Ahmad Khan, F. (2019). A systematic review of the diagnostic accuracy of artificial intelligence-based computer programs to analyze chest x-rays for pulmonary tuberculosis. *PloS One*, 14(9): e0221339. <https://doi.org/10.1371/journal.pone.0221339>
- [29] Meraj, S. S., Yaakob, R., Azman, A., Rum, S. N., Shahrel, A., Nazri, A., & Zakaria, N. F. (2019). Detection of pulmonary tuberculosis manifestation in chest x-rays using different convolutional neural network (CNN) models. *Int. J. Eng. Adv. Technol.(IJEAT)*, 9(1): 2270–2275. <https://doi.org/10.35940/ijeat.A2632.109119>
- [30] Lino, F. D. S. B. M., Alves, G., Morais, F. S. L., Da Silva, R. E., Lorenzato, D. O. J., Endo, P., ... & Sampaio, V. (2021). Benchmarking machine learning models to assist in the prognosis of tuberculosis. *Informatics Учредители: Multidisciplinary Digital Publishing Institute (Basel)*, 8(2). <https://doi.org/10.3390/informatics8020027>
- [31] Yoo, S. H., Geng, H., Chiu, T. L., Yu, S. K., Cho, D. C., Heo, J., ... & Min, B. J. (2020). Study on the TB and non-TB diagnosis using two-step deep learning-based binary classifier. *Journal of Instrumentation*, 15(10): P10011. <https://doi.org/10.1088/1748-0221/15/10/P10011>
- [32] Chang, R. I., Chiu, Y. H., & Lin, J. W. (2020). Two-stage classification of tuberculosis culture diagnosis using convolutional neural network with transfer learning. *The Journal of Supercomputing*, 1–16. <https://doi.org/10.1007/s11227-020-03152-x>
- [33] Khan, M. T., Kaushik, A. C., Ji, L., Malik, S. I., Ali, S., & Wei, D. Q. (2019). Artificial neural networks for prediction of tuberculosis disease. *Frontiers in Microbiology*, 10: 395. <https://doi.org/10.3389/fmicb.2019.00395>

8 Authors

Hery Harjono Muljo is a researcher at Bioinformatics & Data Science Research Center and a lecturer at Accounting Department Faculty of Economic and Communication, Bina Nusantara University, Jakarta, Indonesia. His research expertise is in developing management information system of health institutions such as hospitals and clinics.

Anzaludin Samsinga Perbansa is a researcher at Bioinformatics & Data Science Research Center and a lecturer at School of Information Systems, Bina Nusantara University, Jakarta, Indonesia. His research expertise is in developing tools to investigate the interplay of genetic and environmental factors in agriculture and has developed agricultural germplasm database.

Tjeng Wawan Cenggoro is a researcher of the Bioinformatics & Data Science Research Center (BDSRC), as well as a faculty member of School of Computer Science, Bina Nusantara University. His research focus is the application and development of artificial intelligence for bioinformatics.

Kartika Purwandari is a research assistant at Bioinformatics & Data Science Research Center. Her current research interest is focused on implementing and developing system based on artificial intelligence.

Digdo Sudigyo is a researcher at Bioinformatics & Data Science Research Center. He received a bachelor's degree in Biology from Gadjah Mada University in 2016 and a master's degree in Biotechnology from the Graduated School of Gadjah Mada University in 2020. Currently, he is involved in bioinformatics research, especially in the molecular and cancer fields and data analysis projects.

Bens Pardamean is Director of Bioinformatics & Data Science Research Center and Professor of Computer Science, Bina Nusantara University, Jakarta, Indonesia. His research expertise is in information technology, bioinformatics, and education, including a strong background in database systems, computer networks, and quantitative research.

Article submitted 2021-11-11. Resubmitted 2022-01-06. Final acceptance 2022-01-07. Final version published as submitted by the authors.