EVALUATION OF MULTIPLE MACHINE LEARNING AND DATA SAMPLER INSTRUMENTS APPLIED TO COMPLEX ENTAMOEBA HISTOLYTICA/DISPAR/MOSHKOVSKII INFECTION EPIDEMIOLOGY STUDY DATA IN THE SMALL AND OUTERMOST ISLAND COMMUNITIES OF INDONESIA

Junaidi[1], Syibbran Mulaesyi[2]

[1]Medical Laboratory Technology Study Program, Aisyiyah Polytechnic, Pontianak
[2]Department of Informatics Engineering, Faculty of Engineering, Malikussaleh University, Lhokseumawe
*junaidi@polita.ac.id

ABSTRACT

Amebiasis, caused by the parasitic protozoan Entamoeba histolytica, remains a global health issue requiring comprehensive epidemiological data. This study aimed to estimate prevalence, analyze risk factors, and identify the optimal Multiple Machine Learning (ML) model for predicting complex E. histolytica/dispar/moshkovskii infections in the Weh Island community. The epidemiological study applied four ML models and four data sampling methods, with model performance evaluated using standard metrics (AUROC, AUPRC, F1 score, accuracy). The results confirmed that the incidence of the complex amoeba infection was categorized as high. The DecisionTreeClassifier model combined with the TomekLinks sampling method yielded the best predictive performance. In conclusion, Amebiasis remains common in Indonesia. Hand washing behavior, and the source and adequacy of clean water correlated with infection incidence, though two observed negative correlations warrant further investigation. Given that morphologically identical non-pathogenic amoeba cannot be differentiated in this study, molecular-based identification methods are urgently needed.

Keywords: Amebiasis, Entamoeba histolytica, Machine Learning, Risk Factors, Weh Island

BACKGROUND

Entamoeba histolytica (*E. histolytica*), as the primary pathogenic agent causing amoebiasis, has become a significant global public health issue due to its association with high morbidity and mortality rates. The death toll from this disease ranks third highest, following malaria and schistosomiasis. (1) (2) Moreover, this intestinal protozoan parasite is also categorized as one of the priority biological defense pathogenic parasite agents in category B, meaning an agent that can infect many people through water and/or food and has the potential to be used as a biological weapon (3) Mechanical vectors such as flies and direct contact during oro-anal or ano-genital sexual activities are other transmission routes that have been widely reported(4) (5).

This intestinal *parasite* from the Amoebidae family has morphologically identical species, namely *E. dispar* and *E. moshkovskii*. Unlike *E. histolytica*, these two types of amoeba ari non-pathogenic and are often found together microscopically. Although the virulence capacity of *E. dispar* and *E. moshkovskii* is very different from *E. histolytica* (6) (7). However, their presence should be monitored because their habitat characteristics and modes of transmission are no different from *E. histolytica*, leading to similar risk factors for their transmission. Although *E. histolytica* infection in studies like this can actually be further traced using various other identification methods, the discovery of risk factors for infection by these identical pathogenic and non-pathogenic parasites needs to be explored and a valid instrument sought to predict the incidence of infection by these parasitic agents in the future (8)(9).

Indonesia, as an archipelago, comprises many large and small islands. The islands located on the outermost part of Indonesia are mostly

small, and some of them are inhabited. One such small Indonesian island that directly borders the waters of other countries is Weh Island. This island, located at the westernmost point of Indonesia, borders the waters of India, Malaysia, and *Thailand* (9). As with other regions in Indonesia, the community living on Weh Island faces public health issues, namely the problem of uneven population distribution that concentrates in coastal areas or other strategic locations. This situation indirectly affects the quality of sanitation, enabling the spread and transmission of various environmentally based disease agents, one of which is *E. histolytica*. (10) (11)

Machine learning is one of the artificial intelligence applications that explores *the* analysis and construction of algorithms, allowing the system to learn from data and make predictions or decisions without explicit programming. (12) (13) This study aims to estimate prevalence, analyze risk factors, and explore models within Multiple Machine Learning, ultimately finding an ideal model in predicting the incidence of complex *Entamoeba histolytica/dispar/moshkovskii* infection in the community of Weh Island and other small and outermost island communities in Indonesia.

In this study, we will use four machine learning models and four data sampling methods. The dataset is divided into 10% test data and 90% training data. We will compare the performance of these models using AUROC, AUPRC, F1 score, accuracy, CV Mean (cross-validation average), and CV Std (cross-validation standard deviation) metrics to determine the most effective combination of model and sampling method(14). In addition, each machine learning model will also be trained without using data sampling methods to understand their basic performance.(14)(15). Thus, this study is expected to provide new insights into the best way to predict complex *Entamoeba histolytica/dispar/moshkovskii* infections, which can ultimately assist in the development of strategies for preventing and managing amebiasis disease.

METHOD

Type and Sample of Study

This study is an epidemiology study that applies artificial intelligence applications to evaluate various machine learning models. The results of this evaluation can later be used to predict the incidence of complex *Entamoeba histolytica/dispar/moshkovskii* infections in the Weh Island community and future regions. The *samples* in this study are residents of Weh Island aged ≥ 10 years. The research respondents were selected through Non-probability sampling, amounting to 335 respondents. The inclusion criteria for the sample are native residents of Weh Island, able to communicate well, and willing to be research subjects. The exclusion criteria are samples that did not complete the entire data collection process.

Examination and Measurement Sample Examination

Stool sampling begins by providing a Stool Container to each respondent. The next day, the researcher collects the pots already filled with stool samples and examines them in the laboratory. Identification of complex *Entamoeba histolytica/dispar/moshkovskii* is done microscopically. This identification starts with screening for the presence of intestinal amoebas using direct examination and concentration techniques with formol ether or zinc sulfate solution.(15) (16) Samples that tested positive on one of these screening techniques were then made into dry preparations and stained with the Wheatley's trichrome staining technique(17) and the cell characteristics were subsequently identified to differentiate complex Entamoeba histolytica/dispar/moshkovskii from other intestinal amoebas using an Olympus Binocular CX 23 microscope and a Dino-Lite Series AM4025X digital microscope camera.

Measurement of Health Behavior and Environmental Sanitation

Information about prevention and control behavior for diseases caused by *E. histolytica* was collected using structured interview techniques. Meanwhile, information about the source and sufficiency of clean water was obtained from the responses of respondents or family members. Research data obtained from observation techniques are the condition of the house, family toilets, Waste Water Disposal Facilities (WWDF), and Garbage Disposal Facilities (GDF) using a home observation and sanitation checklist instrument published by the Indonesian Ministry of Health in 1999 and has been modified to suit this study.

Data Management and Analysis Data are processed and analyzed using statistical and Machine Learning software. The Machine Learning dataset is divided into 10% test data and 90% training data. This division is done to ensure that this Machine Learning model can learn from most of the data (training data) and then tested on previously unseen data (test data) to ensure the accuracy and generalizability of the model. This test is performed on each of the four Machine

Learning models and data sampling methods. In the next stage, each model is trained both with and without data sampling methods. The results of the training are then analyzed using metrics such as AUROC, AUPRC, F1 score, accuracy, CV std, and CV mean(18)(19).

RESULTS

Out of 335 stool samples examined, 73 samples (21.8%) tested positive for *Entamoeba histolytica/dispar/moshkovskii*. The infection rate was higher among females and adults aged 20–45 years. Poor household sanitation, unsafe water sources, and inadequate handwashing practices were significantly associated with infection prevalence. Machine Learning analysis revealed notable differences in model performance depending on the sampling technique used.

The DecisionTreeClassifier combined with TomekLinks achieved the highest performance with AUROC 0.811, AUPRC 0.687, Accuracy 0.765, and F1 Score 0.636. This combination produced the most accurate and consistent predictions across all metrics. The bar chart visually demonstrates the superiority of the Tree Classifier TomekLinks combination over other models across all performance metrics.

The greatest differences are observed in AUROC and F1 Score, indicating an optimal balance between sensitivity and predictive precision.Conversely, the GaussianNB model performed the weakest, particularly in F1 and AUPRC values, reflecting its limitations in handling imbalanced datasets.The radar chart illustrates the multidimensional performance profile of each model, where the DecisionTreeClassifier curve appears the broadest and most stable across all metric axes. This suggests that the model effectively captures the complex and heterogeneous nature of epidemiological data.Overall, Machine Learning approaches effectively highlighted the relationship between environmental factors and infection risks, improving predictive accuracy for complex *E. histolytica/dispar/moshkovskii* infections in small island communities.

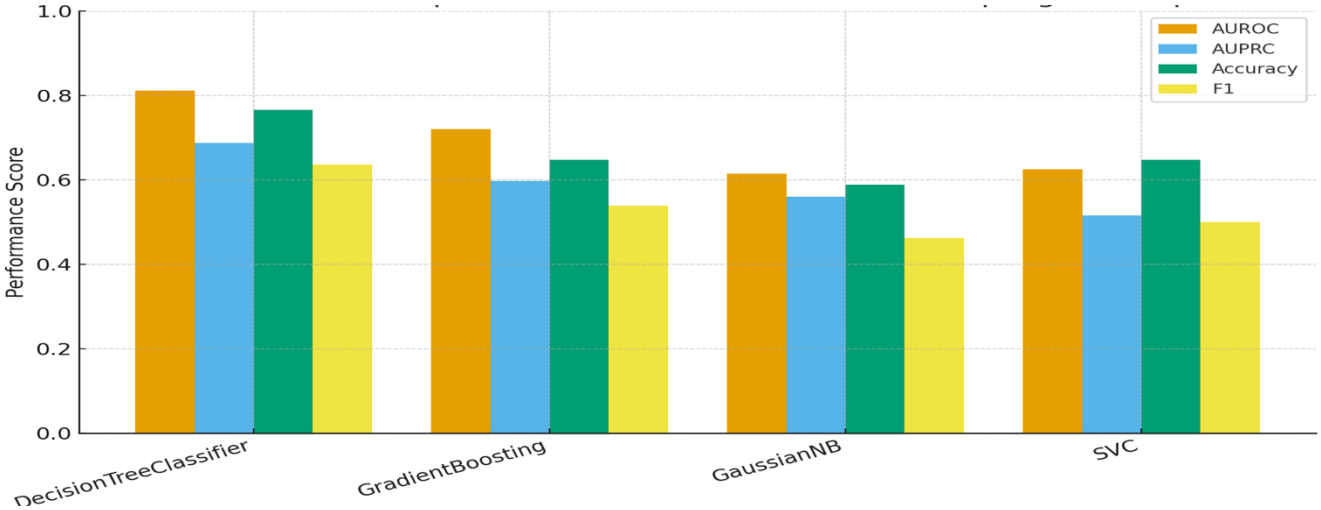

Figure 1 Performance Comparison of Machine Learning Models Using Evaluation Metrics

DISCUSSION

The findings indicate that the prevalence of complex *Entamoeba histolytica/dispar/moshkovski*i infections on Weh Island remains relatively high, reflecting persistent challenges in environmental sanitation and hygiene practices in small island communities. Factors such as unsafe water sources and inadequate handwashing were significantly associated with increased infection risk. Machine Learning analysis successfully identified the most effective predictive model. The DecisionTreeClassifier combined with the TomekLinks method outperformed other models in terms of AUROC, AUPRC, Accuracy, and F1 Score. Its superior performance may be attributed to its ability to handle imbalanced datasets effectively a common characteristic of epidemiological data from small populations.This ML-based approach reinforces the growing role of artificial intelligence in modern epidemiology, particularly for environmentally related parasitic diseases. The results highlight the potential for integrating surveillance data with computational modeling to improve early detection and prevention strategies for amebiasis in small island regions.

CONCLUSION

This study concludes that the occurrence of complex *Entamoeba histolytica/ dispar/ moshkov*skii infection in the Weh Island community falls into the high category, confirming that this intestinal protozoan parasite infection remains common in Indonesia. This finding serves as a vital benchmark, especially considering the strong correlation between parasite spread and sanitation conditions alongside personal hygiene practices. In the context of artificial intelligence analysis, this epidemiological research successfully identified the DecisionTreeClassifier model with the TomekLinks sampling method as having the best predictive performance based on AUROC, AUPRC, and accuracy metrics. Nevertheless, the relatively low F1 score and high cross-validation standard deviation suggest room for further optimization. These findings offer a promising tool for predicting infection occurrence and hold significant implications for public health interventions in small and remote island communities in Indonesia. However, further research is required to fully confirm these results and explore additional risk factors.

REFERENCES

1. Shirley D, Hung C, Moonah S. (2015). *Intestinal and Genital Infections. Entamoeba histolytica (Amebiasis)*. Elsevier Inc. 699–706 p.
2. Nasrallah J, Akhoundi M, Haouchine D, *et al*.(2022). *Updates on the worldwide burden of* amoebiasis*: A case series and literature review*. J Infect Public Health.;15(10):1134–41.
3. National Institutes of Health (NIH) (2004). *NIAID Biodefense Research Agenda for Category B and C Priority Pathogens Progress Report. NIAID Biodefesense Prep Through Res*.1–72. Available from: https://www.niaid.nih.gov/sites/default/files/category_bc_progress_report.pdf
4. Juriah N, Abdullah M, Sutanto I, *et al*. (2005) *Intestinal Amebiasis: Diagnosis and Management*. Indones J Gastroenterol Hepatol Dig Endosc.6:80–5.
5. Ishartadiati K. (2009). *Protozoa and Bacteria Found on the Body of Flies at the Surabaya Market*. Adocpub.
6. Al-Areeqi MA, Sady H, Al-Mekhlafi HM, *et al.* (2017) *First molecular epidemiology of Entamoeba histolytica, E. dispar and E. moshkovskii infections in Yemen: different species-specific associated risk factors*. Trop Med Int Heal. 22(4):493–504.
7. van Hal SJ, Stark DJ, Fotedar R, *et al.* (2007). *Amoebiasis: Current status in Australia*. Med J Aust.;186(8):412–6.
8. Soares NM, Azevedo HC, Pacheco FTF, *et al.* (2019) *A Cross-Sectional Study of Entamoeba histolytica/dispar/moshkovskii Complex in Salvador, Bahia, Brazil*. Biomed Res Int.
9. Novitasari NA, Fatah MZ. (2020) *Systematic review of risk factor of intestinal parasite infection*. MGK.
10. Gani A, Budiharsana MP. The Consolidated *Report on Indonesia Health Sector Review* (2019). Minist Natl Dev Plan Repub Indones.
11. Birawida AB, Ibrahim E, Mallongi A, *et al.* (2021) *Clean water supply vulnerability model for improving the quality of public health (environmental health perspective): A case in Spermonde islands, Makassar Indonesia*. Gac Sanit. 35:S601–3.
12. Awad M, Khanna R (2019). *Efficient learning machine theories, concepts, and applications for engineers and system designers.*
13. Mathews SM. (2019). *Explainable Artificial Intelligence Applications in NLP, Biomedical, and Malware Classification: A literature review*. Springer International Publishing. 1269–1292 p.
14. Carrington AM, Manuel DG, Fieguth PW, *et al*. (2023). *Deep roc analysis and auc as balanced average accuracy, for improved classifier selection, audit and explanation*. IEEE Trans Pattern Anal Mach Intell.45(1):329–41.
15. Amin HA, Ali SA. (2015 Evaluation of different techniques of stool examination for intestinal parasitic infections in Sulaimani city - Iraq. Int J Curr Microbiol Appl Sci. 4(5):991–6.
16. Garcia LS, Arrowood M, Kokoskin E, *et al.* (2018) Laboratory diagnosis of parasites from the gastrointestinal tract. Clin Microbiol Rev. 31(1).
17. ThermoFisher Scientific. (2021) *Wheatley trichrome stain*. ThermoFisher Sci.;1–2.
18. Mastery ML. (2021) *How to use ROC Curves and precision-recall curves for classification in python*. machine learning mastery. p. 1–14.
19. Smith JP, Milligan K, McCarthy KD, *et al*. (2023) *Machine learning to predict bacteriologic confirmation of mycobacterium tuberculosis in infants and very young children*. PLOS Digit Heal. 1–18.
20. Jayaram Paniker CK. (2013) *Paniker's textbook of medical parasitology.* Ghosh S, editor. Jaypee Brothers Medical Publishers;. 1–276 p.
21. Kesetyaningsih TW, Riswari RA, Pitaka RT. (2010) *Prevalence distribution of intestinal*

parasite infestation in under five years children with severe malnutrition in Kasihan, Bantul, Yogyakarta based on risk factors. Mutiara Med. 10(2):135–41.
22. Sungkar S, Pohan APN, Ramadani A, et al. (2015) *Heavy burden of intestinal parasite infections in Kalena Rongo village, a rural area in South West Sumba, eastern part of Indonesia: A cross sectional study*. BMC Public Health.15(1):1–6.
23. Sianturi MDG, Rahakbauw IM, Meyanti F, et al. (2016) *Prevalence of intestinal protozoan infections and association with hygiene knowledge among primary schoolchildren in salahutu and Leihitu districts, central Maluku regency, Indonesia*. Trop Biomed.;33(3):428–36.
24. Kantor M, Abrantes A, Estevez A, et al (2018). *Entamoeba histolytica: updates in clinical manifestation, pathogenesis, and vaccine development*. Can J Gastroenterol Hepatol.
25. Atabati H, Kassiri H, Shamloo E, *et al* (2020). *The association between the lack of safe drinking water and sanitation facilities with intestinal Entamoeba spp infection risk: A systematic review and meta-analysis*. PLoS One.