VOLUME 12 NOMOR 1 JUNE 2025 ISSN 2548 – 611X

JURNAL
BIOTEKNOLOGI & BIOSAINS INDONESIA

Homepage Jurnal: http://ejournal.brin.go.id/JBBI/index

THE COMPARISON BETWEEN LOGISTIC REGRESSION AND CONVOLUTIONAL NEURAL NETWORK FOR MULTI-DRUG RESISTANT TUBERCULOSIS PREDICTION

Perbandingan antara Regresi Logistik dan Jaringan Saraf Konvolusional untuk Prediksi Tuberkulosis Resistan terhadap Berbagai Obat

Albert Widjaja[1], Satrio Wibowo[2], Arli Aditya Parikesit[3]*

[1]Department of Biomedicine, School of Life Sciences, Indonesia International Institute for Life Sciences, Jakarta, Indonesia

[2]Digital Tribe Organization, Indonesian Ministry of Health

[3]Department of Bioinformatics, School of Life Sciences, Indonesia International Institute for Life Sciences, Jakarta, Indonesia

*Email: arli.parikesit@i3l.ac.id

ABSTRACT

Multi-drug resistant tuberculosis (MDR-TB) is caused by *Mycobacterium tuberculosis* strains that resist at least two first-line anti-TB drugs. This disease presents a major global health challenge, particularly affecting middle to lower income countries where affordable and rapid diagnostic tools are urgently needed. To address this, researchers are exploring the combination of whole genome sequencing and machine learning for drug resistance predictions. Using *Mycobacterium tuberculosis* genomic data from databases, both Logistic Regression (LR) and Convolutional Neural Network (CNN) models were trained to predict drug resistance. Performance evaluation revealed that CNN slightly outperformed LR in accuracy and specificity for Rifampicin and Pyrazinamide predictions, while LR showed better results for Isoniazid and Ethambutol. In terms of sensitivity, LR demonstrated superior performance for most drugs, except Ethambutol where CNN excelled. Though computational complexity assessment was incomplete due to hardware limitations, both models showed distinct advantages in predicting first-line anti-TB drug resistance.

Keywords**:** *Convolutional Neural Network, Machine Learning, MDR-TB, Mycobacterium tuberculosis, Logistic Regression*

ABSTRAK

Tuberkulosis yang resistan terhadap berbagai obat (MDR-TB) disebabkan oleh strain *Mycobacterium tuberculosis* yang resisten terhadap minimal dua obat anti-TB lini pertama. Penyakit ini menjadi tantangan kesehatan global yang signifikan, terutama di negara-negara berpenghasilan menengah hingga rendah yang sangat membutuhkan alat diagnostik yang terjangkau dan cepat. Untuk mengatasi hal ini, para peneliti mengeksplorasi kombinasi sekuensing genom lengkap dan pembelajaran mesin untuk memprediksi resistensi obat. Menggunakan data genomik *Mycobacterium tuberculosis* dari database, model Regresi Logistik (RL) dan Jaringan Saraf Konvolusional (JSK) dilatih untuk memprediksi resistansi obat. Evaluasi kinerja menunjukkan bahwa JSK sedikit mengungguli RL dalam akurasi dan spesifisitas untuk prediksi Rifampisin dan Pirazinamid, sementara RL menunjukkan hasil yang lebih baik untuk Isoniazid dan Etambutol. Dalam hal sensitivitas, RL menunjukkan kinerja yang lebih unggul untuk sebagian besar obat, kecuali Etambutol di mana JSK lebih unggul. Meskipun penilaian kompleksitas komputasi tidak lengkap karena keterbatasan perangkat keras, kedua model menunjukkan keunggulan tersendiri dalam memprediksi resistansi obat anti-TB lini pertama.

Kata kunci: *Convolutional Neural Network, Machine Learning, MDR-TB, Mycobacterium tuberculosis, Logistic Regression*

Received*: 27 February 2025* ***Accepted****: 28 April 2025* ***Published****: 1 June 2025*

INTRODUCTION

Tuberculosis (TB) is a disease caused by a bacteria called *Mycobacterium tuberculosis* from the family Mycobacteriaceae. TB has been the second leading cause of death from an infectious agent (World Health Organization, 2021). TB usually only affects the lungs, but it could also infect other sites of the body. Severe symptoms that can be caused by TB include respiratory hemoptysis, cardiovascular disease, high blood pressure, and also cirrhosis (Simonovska et al., 2015).

According to the World Health Organization (WHO), approximately 10 million people were infected by TB with 800 thousand of them coming from Indonesia in 2020 (WHO, 2021). Indonesia is also the second highest TB burden in the world with estimated incidence in three regions in Indonesia were between 201 to 2,485 cases in 100 thousand people per year (Pelletreau, 2022; Parwati et al., 2020).

Currently, diagnosis of MDR-TB was usually done using the GeneXpert MTB/RIF test that utilizes multiplex real-time PCR (Nguyen et al. 2019). However, GeneXpert is relatively expensive and burdens low to middle income countries including Indonesia (Nadjib et al., 2022). Drug susceptibility testing can also be done to diagnose drug resistance; but needs around 3-4 weeks long to culture the TB bacteria (Maksum et al., 2018). To address this, sequencing technologies combined with machine learning models can be used to predict MDR-TB from whole genome sequences (WGS) or area profiling of MDR-TB to create the best drug regimen for each area.

The CRyPTIC Consortium and the 100,000 Genomes Project found that WGS is more scalable, faster, and potentially cheaper than drug-susceptibility testing (The CRyPTIC Consortium and the 100,000 Genomes Project, 2018). Utilizing a database of resistance conferring mutations as a reference to create machine learning models to predict MDR-TB could be a better alternative for diagnosing MDR-TB or resistance profiling. Notably, sequencing outperforms traditional methods like Molecular Rapid Tests and Culture in both sensitivity, specificity, and speed, particularly for detecting TB strain variants (Miotto et al., 2017).

Machine learning, a branch of artificial intelligence, uses statistical methods and algorithms to improve prediction accuracy (Koteluk et al., 2021). There are different types of machine learning models to predict different outcomes depending on what is the desired output. Logistic regression (LR) models, which output categorical variables, are favorable for drug resistance diagnosis. Another approach to predict MDR-TB is using deep learning models, a subfield of machine learning whose methods are inspired by the function and structure of the brain (Koteluk et al., 2021). For prediction of MDR-TB, the convolutional neural network (CNN) would be suitable as it is able to do classification and come up with their own features from detected patterns, unlike traditional ML that needs to be fed with the appropriate features. Though typically used for image classification, CNNs can process genomic data using one-dimensional convolutional algorithms (Kaushik and Kumar, 2019). Machine learning and deep learning have been applied in various healthcare industry such as research, clinical trials, personalized treatment, medical imaging diagnostics, etc (Verma and Verma, 2022).

Current advancements in TB diagnostics implementation for drug resistance have prioritized accuracy and scalability, yet molecular diagnostics in low-income countries continue to face significant logistical and infrastructure challenges (Nadjib et al., 2022; Nalugwa et al. 2020; Pai & Schito, 2015). This study focuses on optimizing predictive performance for first-line anti-tuberculosis drugs while emphasizing computational efficiency. This is particularly relevant given that low socioeconomic countries like Indonesia might lack sufficient computational resources for running computationally intensive models (Vasiliu et al., 2022). Thus, the emphasis is on developing a model that requires less time and space complexity so that it can be practically implemented in resource-constrained environments such as Indonesia.

So, the development of convolutional neural network and logistic regression to predict the drug resistance from the first-line antituberculosis drugs of TB samples were

performed in this project. The accuracy, sensitivity, and specificity of both of the algorithms were compared to assess which algorithm is suitable for the MDR-TB prediction for the four first-line drugs, while tracking their computational utilization and time. Furthermore, the performance metrics were compared against WHO benchmarks to assess their conformity with international standards (WHO, 2018).

Studies have demonstrated that CNNs achieve superior performance compared to logistic regression in omics-based studies, particularly for diagnostic applications (Albaradei et al., 2021; Jin et al., 2023; Shoaib et al., 2023). CNNs were hypothesized to outperform LR in accuracy, sensitivity, and specificity for MDR-TB prediction as they are able to process more complex data and classify with their own features. Thus, these algorithms are expected to give promising results in advancing the diagnosis of MDR-TB in Indonesia.

MATERIALS AND METHODS

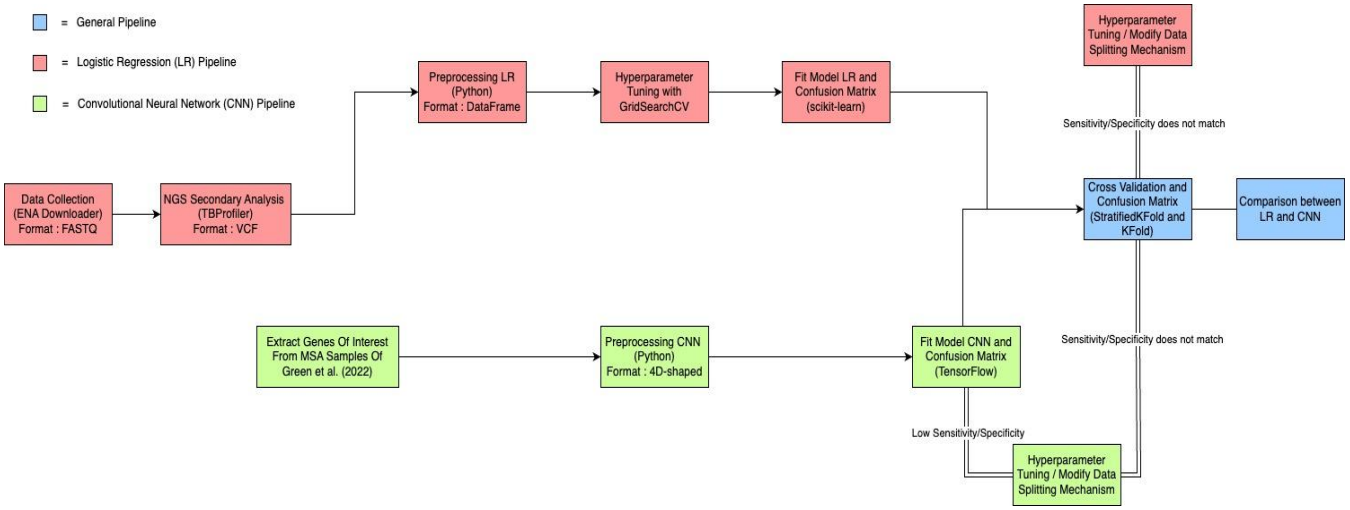

Figure 1. Methodology Overview for Predictive Models Development

Dataset Description

The dataset obtained from Green et al. (2022) consist of the SRR accessions, Multiple Sequence Alignment (MSA) file, and the phenotype data where each entry has been selected based on the presence of all four first-line drugs: Rifampicin, Isoniazid, Ethambutol, and Pyrazinamide. Only data points that include information on all four drugs were included in this dataset, ensuring consistency and completeness for the training process. The number of total samples that was collected for the whole dataset were 12,179 samples. Each samples had the drug susceptibility testing done by PathoSystems Resource Integration Center (PATRIC), WHO Supranational Reference Laboratory Network, ReSeqTB, and various other published literatures (Green et al., 2022). The dataset was divided into 80% train dataset, 10% validation dataset, and 10% test dataset.

Bioinformatics Pipeline to Create MDR-TB Predictive Models

The methodology for creating MDR-TB predictive models is outlined in Figure 1. All tools used in the pipeline were free and open-source. The process begins with collecting FASTQ files from European Nucleotide Archive (ENA) database. These files undergo secondary analysis (trimming, alignment, and variant calling) to produce VCF files. For logistic regression (LR), VCF files and a CSV of mutation positions were preprocessed into a DataFrame, and LR was optimized using cross-validation and evaluated with a confusion matrix. For convolutional neural networks (CNN), gene sequences were extracted from MSA files (Green et al., 2022), one-hot encoded into a 4D numpy array, and fitted to the CNN model, which was also cross-validated and assessed using a confusion matrix. Model performance was evaluated based on

computational complexity (training time, CPU usage), accuracy, sensitivity, and specificity. All processes were executed on a Jupyter Notebook server provided by DTO Kemenkes.

Logistic Regression (LR)

Data Collection and Preprocessing

The data collection process was done using the ENA Downloader tool. NGS secondary analysis was performed using TBProfiler v4.4.1, a pipeline developed by the London School of Hygiene & Tropical Medicine (LSHTM, UK) for M. tuberculosis drug resistance prediction (Phelan et al., 2019). The integrated pipeline incorporates:

- Trimmomatic v0.39 developed by Usadel Lab, Germany for read trimming (Bolger et al., 2014),
- Bowtie2 v2.4.4 developed by Johns Hopkins University, USA for alignment to the H37Rv genome (Langmead & Salzberg, 2012), and
- BCFtools v1.12 developed by Wellcome Sanger Institute, UK for variant calling (Danecek et al., 2021).

An algorithm developed to detect the presence ("1") or absence ("0") of mutations in each VCF sample, with rows represents each sample with labels printed on the "Sample" column. The listed features in the table are non-phylogenetic mutations obtained from Coll et al. (2015) as phylogenetic mutations have caused multiple false positives before in other predictive models (Bolger et al., 2014). Lastly, the final result of the DataFrame is depicted in Table 1.

Table 1. Preprocessed Input Data for LR Model. X represents the number of features and the sample column and Y represents the number of samples. The features selected are written as "mutation coordinate + nucleotide substitution" and the "Sample" column is filled with the labels for each sample. The column of each feature is filled with 0 and 1, which represents the absence and presence of the mutations, respectively.

Sample	760314 G/T	760663 C/T	X = 1,045
ERR2514773	0	1	
ERR2514776	1	0	

Y = 12,179

Convolutional Neural Network (CNN)

Gene Sequences Extraction

The genes of interest were extracted from the MSA samples obtained from Green et al. (2022). An algorithm was designed to cut the genes according to the desired length and positions. The desired position was obtained from Mycobrowser as depicted in Table 2.

Table 2. The 7 clinically relevant genes with its positions according to H37Rv obtained from Mycobrowser

Gene	Position
rpoB	759,609 - 763,369
fabG – inhA	1,672,457 - 1,675,011
katG	2,153,235 - 2,156,706
pncA	2,287,883 - 2,289,599
ahpC	2,726,087 - 2,726,780
embB	4,246,514 - 4,249,810

CNN Data Preprocessing

The gene sequences were one-hot encoded based on their nucleotides (adenine, guanine, cytosine, and thymine) with one gap character. All the samples were later compiled into a 4D Numpy Array as the input for the CNN model fitting. The dimensions in the array would consist of the number of total samples, one-hot encoded nucleotides, length of the longest gene, and also the number of gene sequences ex-

tracted from the sample whole genome sequence.

CNN Architecture

The architecture of the CNN model utilized was obtained from Green et al. with a total of 10 different layers (Green et al., 2022). The input layer was the input dimension for the CNN, which is 12,179 x 5 x 3,888 x 6 where 12,179 represented total number of samples, 5 represented one-hot encoded nucleotides and a gap character, 3,888 represented the longest gene length including its gaps, and 6 represented the clinically relevant gene sequences used. The first hidden layer was a 1D convolution layer with Rectified Linear Unit (ReLU) activation function consisted of 64 filters with a size of 5 x 12 and 1 x 1 stride. The output dimension for the first hidden layer was 12,179 x 1 x 3,877 x 64. The second hidden layer was also a 1D convolution layer with ReLU activation function consisted of 64 filters with a size of 1 x 12 and 1 x 1 stride. The output dimension for the second hidden layer was 12,179 x 1 x 3,866 x 64. The third hidden layer was a pooling layer with max pooling operation consisted of 1 filter with a size of 1 x 3 and 1 x 1 stride. The output dimension for the third hidden layer was 12,179 x 1 x 1,288 x 64. The fourth hidden layer was a 1D convolution layer with ReLU activation function consisted of 32 filters with a size of 1 x 3 and 1 x 1 stride. The output dimension for the fourth hidden layer was 12,179 x 1 x 1,286 x 32. The fifth hidden layer was a 1D convolution layer with ReLU activation function consisted of 32 filters with a size of 1 x 3 and 1 x 1 stride. The output dimension for the fifth hidden layer was 12,179 x 1 x 1,284 x 32. The sixth hidden layer was a pooling layer with max pooling operation consisted of 1 filter with a size of 1 x 3 and 1 x 1 stride. The output dimension for the sixth hidden layer was 12,179 x 1 x 428 x 32. The seventh and eighth hidden layer was both a fully connected layer with ReLU activation function and 256 output nodes each. Lastly, the output layer had 4 output nodes that used the Sigmoid activation function.

Fitting CNN and LR Models

The dataset was divided into 80% train dataset, 10% validation dataset, and 10% test dataset. For the LR model, logistic regression with L2 regularization and parameters obtained from the cross validation was later fitted to the training dataset and trained for a maximum of 250 epochs using the scikit-learn library developed by French Institute for Research in Computer Science and Automation, France (Pedregosa et al., 2012). Logistic regression with L2 regularization was used as it was able to outperform random forest classifier (Green et al. 2022). For the CNN model, the architecture was built using TensorFlow that was developed by Google, USA (Abadi et al., 2016). The model was trained with Adam optimizer applied to stochastic gradient descent with a learning rate of e–9. The model would also be trained for a maximum of 250 epochs with early stopping and patience of 20 epochs. To address the dataset imbalance as shown in Table 3, the LR model used the built-in feature from scikit-learn library that could adjust different weights to each drug according to the proportions, while the CNN model used a modified binary cross entropy and accuracy function from Green et al. that served a similar purpose (Green et al. 2022; Pedregosa et al., 2012). The validation set was utilized to obtain the best threshold using Youden's index. The model was tested on the test dataset and results were implemented into a confusion matrix to calculate the accuracy, sensitivity, and specificity of each drug.

Table 3. Resistant proportions of each drug

Drug	Resistant	Susceptible	Total	Resistant proportion
Ethambutol	1823	10356	12179	0.150
Isoniazid	3665	8514	12179	0.301
Pyrazinamide	1739	10440	12179	0.143
Rifampicin	2896	9283	12179	0.238

Cross Validation

Both models undergo 10-fold cross validation and were assessed using confusion matrices that were concatenated to get the overall accuracy, sensitivity, and specificity. However, for LR model, the cross validation was done prior to model fitting to find the best hyperparameters for the model, which it searched for the ideal solver and the L2 regularization constant. For the CNN model, cross validation was only done after the model fitting and if the accuracy, sensitivity, and specificity had more than 10% difference with the model, fine tuning would also be done to find the best training epochs, learning rate, optimization algorithm, or activation functions.

Comparison Between LR and CNN

For each drug from both models, receiver operating characteristic (ROC) curve was plotted to find the sensitivity and specificity from each cut off value. The Youden's index was used to find the best cut off that obtained the highest sensitivity and specificity with minimum differences between both values (most balanced). The accuracy was then calculated with the optimum sensitivity and specificity. The obtained accuracy, sensitivity, and specificity from both models for each first-line drug were used to benchmark and determine which model performs better in predicting first-line antituberculosis drug resistance. The accuracy would be the main parameter to determine the better model, followed by the sensitivity parameter. The sensitivity parameter was chosen to be the more significant parameter over specificity for the assessment because the risk of further tests is insignificant and TB itself is curable in the preclinical phase (Connolly et al., 2007; Gupta, 2013). Prior machine learning studies have established frameworks for model evaluation without reliance on traditional statistical significance testing (Arango-Argoty et al., 2018; Gröschel et al., 2021). Cross-validation was implemented to validate the stability of the results, assessing performance consistency across distinct data subsets. Moreover, cross-validation ensures that a model's predictive capability is consistent rather than fortuitous, making it a robust method for assessing generalizability and preventing overfitting where traditional hypothesis testing might be inadequate (De Rooij & Weeda, 2020). According to Wilimitis & Walsh (2023), cross-validation is a resilient method for model evaluation and selection in healthcare predictive modeling. The differences in metrics between valid, test, and cross-validation sets would be evaluated to ensure no significant difference (more than 10%) between the datasets. This 10% threshold was chosen as a standard criterion in the field to ensure that the model's performance is robust and consistent across different data subsets. The choice of a 10% threshold is supported by the fact that even a stringent p-value such as $p < 0.001$ might still result in a false discovery rate of at least 10%, making it important to ensure that the model's performance is stable and reliable across different data subsets (Colquhoun, 2014; Korthauer et al., 2019; Concato & Hartigan, 2016).

The computational power and computational time parameter would also be taken into consideration to decide the better model. The computational details of the Jupyter Notebook server provided by Digital Transformation Office (DTO) Kemenkes were the E2 server series from Google Cloud that provides 32 cores of CPU and 128 GB of RAM.

RESULTS AND DISCUSSION

Logistic Regression (LR)

The results of the accuracy, sensitivity, and specificity of the Test Set, Valid Set, and Cross Validation for Logistic Regression (LR) model was shown in Table 4. The results showed that there were no significant differences (above 10%) in the three parameters measured between all of the sets. No significant differences means that there was no biases that happened during the learning process on the training set and no signs of the model to overfit to the training data.

Table 4. Accuracy, Sensitivity, and Specificity Results of LR Model on Different Datasets

		Rifampicin	Isoniazid	Ethambutol	Pyrazinamide
Accuracy	Test Set	91.22%	97.04%	91.54%	88.34%
	Valid Set	93.02%	96.39%	91.63%	89.49%
	Cross Validation	92.42%	95.71%	90.89%	88.00%
Sensitivity	Test Set	95.33%	94.32%	91.00%	88.04%
	Valid Set	90.22%	91.45%	93.33%	86.90%
	Cross Validation	87.92%	89.58%	89.97%	86.72%
Specificity	Test Set	89.87%	98.32%	91.65%	88.39%
	Valid Set	93.84%	98.38%	91.36%	89.90%
	Cross Validation	93.83%	98.36%	91.05%	88.22%

Convolutional Neural Network (CNN)

The results of the accuracy, sensitivity, and specificity of the Test Set, Valid Set, and Cross Validation for Convolutional Neural Network (CNN) model was shown in Table 5**.** The results showed that there were no significant differences (above 10%) in the three parameters measured between all of the sets, similar to results from LR. The results show no biases that happen during the learning process on the training set and also no signs of model overfitting.

Table 5. Accuracy, Sensitivity, and Specificity Results of CNN Model on Different Datasets

		Rifampicin	Isoniazid	Ethambutol	Pyrazinamide
Accuracy	Test Set	97.04%	96.80%	90.48%	90.31%
	Valid Set	97.78%	97.95%	91.13%	91.79%
	Cross Validation	97.41%	96.81%	89.37%	88.60%
Sensitivity	Test Set	93.75%	94.09%	94.12%	84.85%
	Valid Set	95.89%	95.44%	96.77%	92.86%
	Cross Validation	94.92%	93.78%	96.22%	91.09%
Specificity	Test Set	98.06%	97.99%	89.89%	91.17%
	Valid Set	98.38%	99.05%	90.12%	91.60%
	Cross Validation	98.18%	98.12%	88.16%	88.19%

Accuracy Comparison Between CNN and LR

The accuracy of each model on each drug for the test dataset is depicted in Figure 2(a). For the drug, Rifampicin, CNN model performed better with 97.04% accuracy, while the LR model had 91.22% accuracy. CNN also performed better on Pyrazinamide with 90.31% and the LR model had 88.34% accuracy. However, the LR model was able to outperform the CNN model on the other two drugs, which are the Isoniazid and Ethambutol. For Isoniazid, the LR model could outperform CNN by a thin margin with 97.04% compared to 96.8%. While for Ethambutol, the LR model had 91.54% accuracy, while CNN had 90.48%.

Sensitivity Comparison Between CNN and LR

The Sensitivity of each model on each drug for the test dataset is depicted in Figure 2(b). The LR model was able to outperform the sensitivity of the CNN model on three different drugs, which are the Rifampicin, Isoniazid, and Pyrazinamide with 95.33%, 94.32%, and 88.04%, respectively. On the other hand, sensitivity from CNN for the Rifampicin drug was 93.75%, Isoniazid with 94.09%, and Pyrazinamide with 84.85%. However, CNN was able to outperform LR in Ethambutol with 94.12%, while LR had 91% sensitivity.

Specificity Comparison Between CNN and LR

The Specificity of each model on each drug for the test dataset is depicted in Figure 2(c). CNN was able to outperform LR in Rifampicin and Pyrazinamide with a bigger margin than how LR outperformed CNN in Isoniazid and Ethambutol. CNN scored 98.06% for Rifampicin and 91.17% for Pyrazinamide, while LR scored 89.97% for Rifampicin and 88.04% for Pyrazinamide. However, LR performs better in Isoniazid with 98.32% and Ethambutol with 91.65% specificity, while CNN had 97.99% for Isoniazid and 89.89% for Ethambutol.

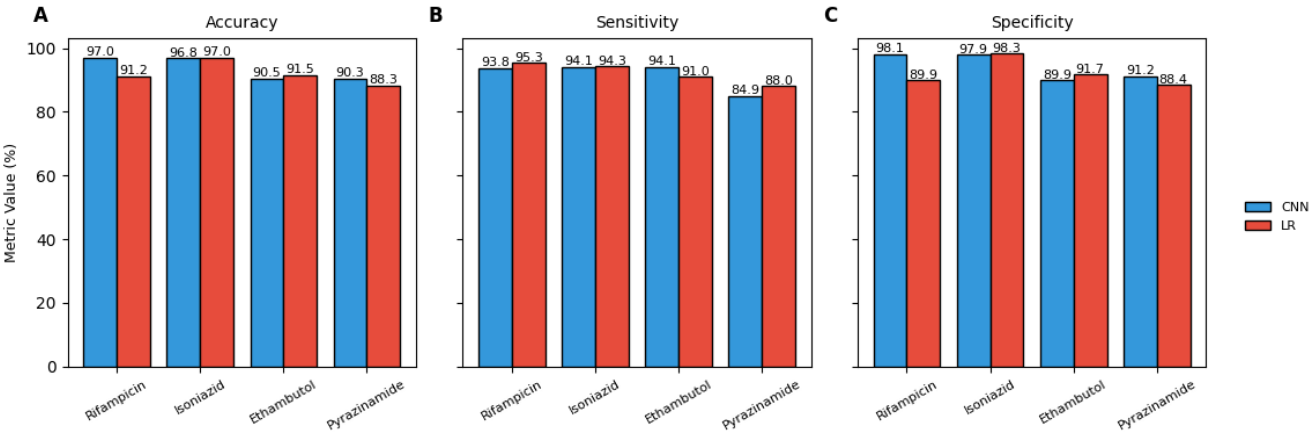

Figure 2. (a) Accuracy, (b) Sensitivity, and (c) Specificity Comparison of CNN and LR Model Bar Chart

Training Time and Prediction Time Between CNN and LR

The Training time and Prediction time of each model on each drug is depicted in Table 6. The amount of time it took for the LR model to train for predicting resistance against Rifampicin, Isoniazid, Ethambutol, and Pyrazinamide was 0.75 s, 0.36 s, 1.43 s, and 1.13 s, respectively. For the CNN model, it took 14,269.71 s for the model to train on the data to predict all of the first-line drugs. For prediction time, the LR model took 0.03 s to predict resistance against Rifampicin or Isoniazid. While for Ethambutol and Pyrazinamide, it took the LR model 0.02 s to predict either one of the drugs mentioned. On the other hand, the CNN model needed 2.57 seconds to predict all four of the drugs simultaneously.

Table 6. Training Time and Prediction Time Results of LR and CNN Models (In Seconds)

		Rifampicin	Isoniazid	Ethambutol	Pyrazinamide
Training	LR	0.75	0.36	1.43	1.13
	CNN	14,269.71			
Prediction	LR	0.03	0.03	0.02	0.02
	CNN	2.57			

Abbreviations: CNN convolutional neural network; LR logistic regression

CPU Utilization During Training and Predicting Between CNN and LR

The CPU Utilization during training and predicting from each model on each drug is depicted in Table 7. The LR model utilized 5.6% of the CPU to train for predicting resistance against Rifampicin, 1.8% for Isoniazid, 6% for Ethambutol, and 5.6% for Pyrazinamide. While for the CNN model, the model only utilized 0.2% of the CPU to train on the data for all of the four drugs. During prediction, the LR model that predicts resistance against Rifampicin utilized 9.5% of CPU, 11.1% for Isoniazid, 10% for Ethambutol, and 9.4% for Pyrazinamide. However, the CNN model only utilized 0.2% of the CPU, the same percentage that was utilized for training the model.

Table 7. CPU Utilization Percentage During Training and Prediction of LR and CNN Models

		Rifampicin	Isoniazid	Ethambutol	Pyrazinamide
Training	LR	5.60%	1.80%	6.00%	5.60%
	CNN	0.20%			
Prediction	LR	9.50%	11.10%	10.00%	9.40%
	CNN	0.20%			

Abbreviations: CNN convolutional neural network; LR logistic regression

Results Summary and WHO Benchmarking

The CNN model demonstrated superior accuracy for Rifampicin and Pyrazinamide, while LR showed a slight edge for Isoniazid and Ethambutol. However, LR consistently achieved higher sensitivity across three of the four drugs, excelling at identifying true positive cases, whereas CNN outperformed LR in specificity for Rifampicin and Pyrazinamide, minimizing false positives.

The performance of the models from the test set were compared to benchmarks published by the WHO (World Health Organization, 2018). The sensitivity benchmark set by WHO (2018) for Pyrazinamide, Isoniazid, and Rifampicin were 85%, 90%, and 95%. However, there is no benchmark set for Ethambutol as the mechanism of resistance were not well-defined or well-documented (World Health Organization, 2018). On the other hand, the benchmark set for specificity was 95% for all of the four drugs (World Health Organization, 2018).

In comparison, Logistic Regression (LR) met or exceeded the WHO sensitivity benchmarks for Pyrazinamide (88.04%), Isoniazid (94.32%), and Rifampicin (95.33%). However, only Isoniazid (98.32%) that exceeded the benchmark in specificity, while it slightly underperforms for Pyrazinamide (88.39%), Ethambutol (91.65%), and Rifampicin (89.87%). CNN achieved higher sensitivity for Isoniazid (94.09%). Moreover, CNN also reached near-benchmark sensitivity for Rifampicin (93.75%), but lagged slightly for Pyrazinamide (84.85%). Meanwhile, CNN achieved superior specificity on Isoniazid (97.99%) and Rifampicin (98.06%), but underperform on Ethambutol (89.89%) and Pyrazinamide (91.17%). Nonetheless, both models performance were comparable to the WHO benchmarks, which shows the models' potential for clinical application.

The most significant difference from the results of the LR model was the sensitivity value from Rifampicin on the test set and cross validation, which was 95.33% and 87.92%. The 7.41% difference might be caused by the utilization of the decision threshold obtained from the Youden's Index on the test set results. Youden's Index itself was utilized to search for the best threshold that provided optimal sensitivity and specificity (Schisterman et al., 2008). However, in certain cases, there could be a value tradeoff between both parameters to achieve the most balanced decision threshold (Habibzadeh et al., 2016). The valid set and cross validation in the LR model were not implemented with the optimized threshold from Youden's Index. Moreover, the specificity value from Rifampicin the test set seems to be lower than the other datasets, which emphasizes the reason why the sensitivity on the test set seems to be abnormally high compared to the other datasets. Furthermore, the results from CNN on the same drug does not indicate any abnormal differences between the datasets, which means data bias was not the reason for the differences as the data splitting mechanism was the same.

The most significant difference from the results of the CNN model was the sensitivity value from Pyrazinamide on the test set and valid set, which was 84.85% and 92.86%. The 8.01% difference might be caused by the improper relationships with other genes that should not influence the prediction. Moreover, phylogenetic mutations were included in the input, which could cause prediction errors (Coll et al., 2015). Furthermore, the results from the LR model does not show any abnormal differences between the datasets, which eliminates data

bias as a reason for the underperformed sensitivity. Lastly, the CNN model itself was not hyperparameter tuned for the specific input and could result in suboptimal performance (Li et al., 2018).

The time taken and CPU utilization between both models to train and predict was exceptionally distant. The huge difference in the time taken and CPU utilization was due to the hardware used for the model training and prediction. Logistic regression is a linear model, which is still suitable to train using CPU as CPU works linearly (Hennessy & Patterson, 2011). However, convolutional neural network is not a linear model and CPU is unable to train the model efficiently (LeCun et al., 2015). As a result, the LR model seems to have high CPU utilization and lower time needed for training and prediction compared to the exceptionally low CPU utilization and exceptionally higher time needed for the CNN model to train and predict. To fix the issue, GPU should be utilized to do model training and prediction as GPU is able to run tasks parallelly, which is suitable for deep learning models or even simple machine learning models (LeCun et al., 2015). However, considering the lack of established computational infrastructure in low socioeconomic countries like Indonesia, access to GPUs is likely to remain limited (Vasiliu et al., 2022). Thus, LR is better suited for resource-constrained environments due to its rapid training and inference.

The comparison between both models should be determined based on the results from the test sets. From the accuracy results, CNN seems to be able to outperform the LR model slightly as it was able to perform better on a bigger margin on Rifampicin and Pyrazinamide, but perform slightly worse on Ethambutol and Isoniazid. However, judging from the sensitivity performance, LR was able to outperform CNN slightly on three different drugs, which are the Rifampicin, Isoniazid, and Pyrazinamide. From the specificity side, both models had two different drugs that outperform each other. Prioritizing accuracy as the primary metric, CNN emerges as marginally stronger overall due to its higher accuracy for two drugs and robust specificity, though LR's superior sensitivity suggests its utility in clinical settings where detecting true positives are critical (Connolly et al., 2007; Gupta, 2013). Moreover, both models' performance was comparable to WHO benchmarks, demonstrating their potential to meet international standards for clinical applicability.

According to Green et al. (2022), the CNN model was expected to outperform the LR model specifically in terms of sensitivity. However, the LR model seems to be able to slightly outperform CNN on sensitivity, but slightly fell short on specificity, based on the margin discrepancy at the drug it underperformed (Green et al., 2022). Difference in results might be due to different input data where in this LR model, the input data were non-phylogenetic mutations obtained from Phelan et al. utilized in TBProfiler, whereas Green et al. utilized specific sites of the gene to extract the features (Green et al., 2022; Phelan et al., 2019). Furthermore, another possible explanation for the CNN's inability to outperform LR could be attributed to the absence of hyperparameter tuning specifically tailored for the given input. Hyperparameter tuning could optimize the model to perform the best it could in predicting the resistances as the right parameters could improve the model's performance (Li et al., 2018). Hyperparameter tuning, such as adding regularization or initializer, experimenting on different filter sizes and number of filters, or changing the learning rate and batch sizes could be done to optimize the model better (Hinz et al. 2018).

Despite the CNN's potential for higher performance with appropriate tuning and hardware, the LR model stands out as a more viable option for deployment in Indonesia due to its ability to operate efficiently on CPUs. Given the limited availability of GPUs in the country, the LR model offers a practical and cost-effective solution. LR ensures reliable performance without the need for advanced hardware that is often unattainable or prohibitively expensive in many Indonesian regions. Moreover, the model can also be trained with clinical data that could help identify new variant/strain of TB.

This research had several limitations that included uncertain genomic variants, possibility of suboptimal performance from

CNN, and also the origin of the sample sequence that was used. The genomic variants that were detected by the variant callers might not be 100% accurate. However, most variant callers had 95% sensitivity and 99% specificity on average, which means the risk would be very small for an error that could alter the model's prediction (Liu et al., 2013). Nonetheless, the possibility would still be a limitation in this research. Next, the CNN model was not hyperparameter tuned according to the specific input, which could cause suboptimal performance. Lastly, the origin of the sample sequences that were used in this research does not come from Indonesia. Due to not originating from Indonesia, utilizing the model on Indonesian samples might cause false negatives as there could be important features that specifically occur on Indonesian samples.

CONCLUSION

The results from the accuracy, sensitivity, and specificity parameters on both models implied that each model had its own advantages and disadvantages. CNN exhibits a slight edge in overall performance, driven by its accuracy and specificity. However, LR's better sensitivity supports its utility in clinical settings. Nonetheless, the LR model is particularly suitable to be used in Indonesia because of its efficiency when operating on CPUs, especially given the scarcity of GPUs in the country. This makes the LR model a practical and cost-effective choice, as it can deliver reliable performance without relying on high-end hardware, which is often inaccessible or expensive in many regions of Indonesia.

ACKNOWLEDGEMENT

The authors would like to express their gratitude to the Digital Transformation Office from the Ministry of Health (MOH) of the Republic of Indonesia who has provided support and funding for the cloud server.

REFERENCES

Abadi M, Agarwal A, Barham P, Brevdo E, Chen Z, Citro C, Corrado GS, Davis A, Dean J, Devin M, Ghemawat S, Goodfellow I, Harp A, Irving G, Isard M, Jia Y, Jozefowicz R, Kaiser L, Kudlur M, Levenberg J, Mane D, Monga R, Moore S, Murray D, Olah C, Schuster M, Shlens J, Steiner B, Sutskever I, Talwar K, Tucker P, Vanhoucke V, Vasudevan V, Viegas F, Vinyals O, Warden P, Wattenberg M, Wicke M, Yu Y, Zheng X (2016) TensorFlow: Large-Scale Machine Learning on Heterogeneous Distributed Systems

Albaradei, S., Thafar, M., Alsaedi, A., Van Neste, C., Gojobori, T., Essack, M., & Gao, X. (2021). Machine learning and deep learning methods that use omics data for metastasis prediction. *Computational and Structural Biotechnology Journal*, *19*, 5008–5018. https://doi.org/10.1016/j.csbj.2021.09.001

Arango-Argoty, G., Garner, E., Pruden, A., Heath, L. S., Vikesland, P., & Zhang, L. (2018). DeepARG: a deep learning approach for predicting antibiotic resistance genes from metagenomic data. Microbiome, 6(1). https://doi.org/10.1186/s40168-018-0401-z

Bolger AM, Lohse M, Usadel B (2014) Trimmomatic: a flexible trimmer for Illumina sequence data. Bioinformatics 30:2114–2120. https://doi.org/10.1093/bioinformatics/btu170

Coll F, McNerney R, Preston MD, Guerra-Assunção JA, Warry A, Hill-Cawthorne G, Mallard K, Nair M, Miranda A, Alves A, Perdigão J, Viveiros M, Portugal I, Hasan Z, Hasan R, Glynn JR, Martin N, Pain A, Clark TG (2015) Rapid determination of anti-tuberculosis drug resistance from whole-genome sequences. Genome Medicine 7:51. https://doi.org/10.1186/s13073-015-0164-0

Colquhoun, D. (2014). An investigation of the false discovery rate and the misinterpretation of p -values. *Royal Society Open Science*, *1*(3), 140216. https://doi.org/10.1098/rsos.140216

Concato, J., & Hartigan, J. A. (2016). P values: From suggestion to superstition. *Journal of Investigative Medicine*, *64*(7), 1166–1171. https://doi.org/10.1136/jim-2016-000206

Connolly LE, Edelstein PH, Ramakrishnan L (2007) Why Is Long-Term Therapy Required to Cure Tuberculosis? PLoS Med 4:e120. https://doi.org/10.1371/journal.pmed.0040120

Danecek P, Bonfield JK, Liddle J, Marshall J, Ohan V, Pollard MO, Whitwham A, Keane T, McCarthy SA, Davies RM, Li H (2021) Twelve years of SAMtools and BCFtools. Gigascience 10:giab008. https://doi.org/10.1093/gigascience/giab008

De Rooij, M., & Weeda, W. (2020). Cross-Validation: a method every psychologist should know. *Advances in Methods and Practices in Psychological Science*, *3*(2), 248–263. https://doi.org/10.1177/2515245919898466

Green AG, Yoon CH, Chen ML, Ektefaie Y, Fina M, Freschi L, Gröschel MI, Kohane I, Beam A, Farhat M (2022) A convolutional neural network highlights mutations relevant to antimicrobial resistance in Mycobacterium tuberculosis. Nat Commun 13:3817. https://doi.org/10.1038/s41467-022-31236-0

Gröschel, M. I., Owens, M., Freschi, L., Vargas, R., Marin, M. G., Phelan, J., Iqbal, Z., Dixit, A., & Farhat, M. R. (2021). GenTB: A user-friendly genome-based predictor for tuberculosis resistance powered by machine learning. *Genome Medicine*, *13*(1). https://doi.org/10.1186/s13073-021-00953-4

Gupta N (2013) Accuracy, Sensitivity and Specificity Measurement of Various Classification Techniques on Healthcare Data. IOSR-JCE 11:70–73. https://doi.org/10.9790/0661-1157073

Habibzadeh F, Habibzadeh P, Yadollahie M (2016) On determining the most appropriate test cut-off value: the case of tests with continuous results. Biochem Med (Zagreb) 26:297–307. https://doi.org/10.11613/BM.2016.034

Hennessy JL, Patterson DA (2011) Computer Architecture, Fifth Edition: A Quantitative Approach, 5th edn. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA

Hinz T, Navarro-Guerrero N, Magg S, Wermter S (2018) Speeding up the Hyperparameter Optimization of Deep Convolutional Neural Networks. Int J Comp Intel Appl 17:1850008. https://doi.org/10.1142/S1469026818500086

Jin, C., Jia, C., Hu, W., Xu, H., Shen, Y., & Yue, M. (2023). Predicting antimicrobial resistance in E. coli with discriminative position fused deep learning classifier. *Computational and Structural Biotechnology Journal*, *23*, 559–565. https://doi.org/10.1016/j.csbj.2023.12.041

Kaushik R, Kumar S (2019) Image Segmentation Using Convolutional Neural Network. International Journal of Scientific & Technology Research

Korthauer, K., Kimes, P. K., Duvallet, C., Reyes, A., Subramanian, A., Teng, M., Shukla, C., Alm, E. J., & Hicks, S. C. (2019). A practical guide to methods controlling false discoveries in computational biology. *Genome Biology*, *20*(1). https://doi.org/10.1186/s13059-019-1716-1

Koteluk O, Wartecki A, Mazurek S, Kołodziejczak I, Mackiewicz A (2021) How Do Machines Learn? Artificial Intelligence as a New Era in Medicine. J Pers Med 11:32. https://doi.org/10.3390/jpm11010032

Langmead B, Salzberg SL (2012) Fast gapped-read alignment with Bowtie 2. Nat Methods 9:357–359. https://doi.org/10.1038/nmeth.1923

LeCun Y, Bengio Y, Hinton G (2015) Deep learning. Nature 521:436–444. https://doi.org/10.1038/nature14539

Li L, Jamieson K, DeSalvo G, Rostamizadeh A, Talwalkar A (2018) Hyperband: A Novel Bandit-Based Approach to Hyperparameter Optimization

Liu X, Han S, Wang Z, Gelernter J, Yang B-Z (2013) Variant Callers for Next-Gen-

eration Sequencing Data: A Comparison Study. PLOS ONE 8:e75619. https://doi.org/10.1371/journal.pone.0075619

Maksum I, Suhaili S, Amalia R, Kamara D, Rachman S, Rachman R (2018) PCR Multipleks untuk Identifikasi Mycobacterium tuberculosis Resisten terhadap Isoniazid dan Rifampisin pada Galur Lokal Balai Laboratorium Kesehatan Provinsi Jawa Barat. Jurnal Kimia VALENSI 4:107–118. https://doi.org/10.15408/jkv.v4i2.7226

Miotto, P., Tessema, B., Tagliani, E., Chindelevitch, L., Starks, A. M., Emerson, C., Hanna, D., Kim, P. S., Liwski, R., Zignol, M., Gilpin, C., Niemann, S., Denkinger, C. M., Fleming, J., Warren, R. M., Crook, D., Posey, J., Gagneux, S., Hoffner, S., . . . Rodwell, T. C. (2017). A standardised method for interpreting the association between mutations and phenotypic drug resistance inMycobacterium tuberculosis. *European Respiratory Journal*, *50*(6), 1701354. https://doi.org/10.1183/13993003.01354-2017

Nadjib M, Dewi RK, Setiawan E, Miko TY, Putri S, Hadisoemarto PF, Sari ER, Pujiyanto, Martina R, Syamsi LN (2022) Cost and affordability of scaling up tuberculosis diagnosis using Xpert MTB/RIF testing in West Java, Indonesia. PLoS One 17:e0264912. https://doi.org/10.1371/journal.pone.0264912

Nalugwa, T., Shete, P. B., Nantale, M., Farr, K., Ojok, C., Ochom, E., Mugabe, F., Joloba, M., Dowdy, D. W., Moore, D. a. J., Davis, J. L., Cattamanchi, A., & Katamba, A. (2020). Challenges with scale-up of GeneXpert MTB/RIF® in Uganda: a health systems perspective. *BMC Health Services Research*, *20*(1). https://doi.org/10.1186/s12913-020-4997-x

Nguyen TNA, Anton-Le Berre V, Bañuls A-L, Nguyen TVA (2019) Molecular Diagnosis of Drug-Resistant Tuberculosis; A Literature Review. Frontiers in Microbiology 10

Pai, M., & Schito, M. (2015). Tuberculosis Diagnostics in 2015: Landscape, Priorities, Needs, and Prospects. *The Journal of Infectious Diseases*, *211*(suppl_2), S21–S28. https://doi.org/10.1093/infdis/jiu803

Parwati CG, Farid MN, Nasution HS, Basri C, Lolong D, Gebhard A, Tiemersma EW, Pambudi I, Surya A, Houben RMGJ (2020) Estimation of subnational tuberculosis burden: generation and application of a new tool in Indonesia. Int J Tuberc Lung Dis 24:250–257. https://doi.org/10.5588/ijtld.19.0139

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Müller, A., Nothman, J., Louppe, G., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., & Duchesnay, É. (2012, January 2). *SciKit-Learn: Machine Learning in Python*. arXiv.org. https://arxiv.org/abs/1201.0490

Pelletreau S (2022) Desk Review: Pediatric Tuberculosis with a Focus on Indonesia | UNICEF Indonesia

Phelan JE, O'Sullivan DM, Machado D, Ramos J, Oppong YEA, Campino S, O'Grady J, McNerney R, Hibberd ML, Viveiros M, Huggett JF, Clark TG (2019) Integrating informatics tools and portable sequencing technology for rapid detection of resistance to anti-tuberculous drugs. Genome Medicine 11:41. https://doi.org/10.1186/s13073-019-0650-x

Schisterman EF, Faraggi D, Reiser B, Hu J (2008) Youden Index and the optimal threshold for markers with mass at zero. Stat Med 27:297–315. https://doi.org/10.1002/sim.2993

Shoaib, M., Shah, B., Sayed, N., Ali, F., Ullah, R., & Hussain, I. (2023). Deep learning for plant bioinformatics: an explainable gradient-based approach for disease detection. *Frontiers in Plant Science*, *14*. https://doi.org/10.3389/fpls.2023.1283235

Simonovska L, Trajcevska M, Mitreski V, Simonovska I (2015) The causes of

death among patients with tuberculosis. European Respiratory Journal 46

The CRyPTIC Consortium and the 100,000 Genomes Project (2018) Prediction of Susceptibility to First-Line Tuberculosis Drugs by DNA Sequencing. New England Journal of Medicine 379:1403–1415. https://doi.org/10.1056/NEJMoa1800474

Vasiliu, A., Saktiawati, A. M. I., Duarte, R., Lange, C., & Cirillo, D. M. (2022). Implementing molecular tuberculosis diagnostic methods in limited-resource and high-burden countries. Breathe, 18(4), 220226. https://doi.org/10.1183/20734735.0226-2022

Verma VK, Verma S (2022) Machine learning applications in healthcare sector: An overview. Materials Today: Proceedings 57:2144–2147. https://doi.org/10.1016/j.matpr.2021.12.101

Wilimitis, D., & Walsh, C. G. (2023). Practical Considerations and Applied Examples of Cross-Validation for model development and Evaluation in Health Care: tutorial. *JMIR AI*, *2*, e49023. https://doi.org/10.2196/49023

World Health Organization (2021) Global Tuberculosis Report 2021. World Health Organization

World Health Organization (2018). *The use of next-generation sequencing technologies for the detection of mutations associated with drug resistance in Mycobacterium tuberculosis complex: technical guide*. https://www.who.int/publications/i/item/WHO-CDS-TB-2018.19