Chapter 1

Introduction

1.1. Background

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 (WHO, 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 WHO (2021), approximately 10 million people were infected by TB with 800 thousand of them coming from Indonesia in 2020. 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). The condition was worsened due to the COVID-19 pandemic as the numbers of reported TB incidence between 2019 and 2020 significantly decreased by about 31%, which hinders the attempt to stop the epidemic that utilizes tracking TB incidence extensively to stop the transmission (WHO, 2021).

Diagnosis of TB has been done using various methods such as nucleic acid amplification tests (NAAT), culture-based testing, AFB (acid-fast bacilli) microscopy, and also TB skin test. However, diagnosing multidrug resistant tuberculosis (MDR-TB), which is TB that is able to evade multiple first-line drugs, is a whole different task that cannot be achieved with previously mentioned methods. 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; however, drug-susceptibility testing needs around 3-4 weeks long to

culture di TB bacteria (Maksum et al., 2018). To deal with this issue, utilization of sequencing technologies mixed with machine learning models to predict MDR-TB from whole genome sequences (WGS) can be done for diagnosis of MDR-TB or area profiling of MDR-TB to create the best drug regimen for each area.

According to The CRyPTIC Consortium and the 100,000 Genomes Project (2018), whole genome sequencing is more scalable, faster, and also could be cheaper than drug-susceptibility testing. Utilizing a database with a list of mutations that confers resistance as reference to create machine learning models to predict MDR-TB could be a better alternative for diagnosing MDR-TB or resistance profiling. Moreover, the machine learning models could be developed to detect multiple drug resistances, which would be advantageous as GeneXpert could only diagnose Rifampicin resistance on Mtb. Machine learning itself is a branch of artificial intelligence (AI) that can be used to predict outcomes using statistical methods and algorithms that could gradually improve their accuracy (IBM Cloud Education, 2020). There are different types of machine learning models to predict different outcomes depending on what is the desired output. For the diagnosis of drug resistance, logistic regression model is more favorable as it outputs categorical variables. Logistic regression uses the sigmoid function to predict the output, which in this case is either resistant to the drug or not.

Another way to predict MDR-TB would be utilizing deep learning models to look at patterns that signify the resistance from the whole genome sequences and determine if they are resistant or not. Deep learning is also known as artificial neural networks, a subfield of machine learning whose methods are inspired by the function and structure of the brain (IBM Cloud Education, 2020). For prediction of MDR-TB the convolutional neural network (CNN) would be suitable because they are able to do classification. Moreover, CNN is able to come up with their own features from patterns they detected, unlike traditional ML that needs to be fed with the appropriate features. CNN is usually used for image classification, processing, and segmentation (Kaushik & Kumar, 2019). However, CNN can also be used

for genome sequences using an algorithm called one dimensional convolutional neural network that is able to process data with only one dimension. Machine learning and deep learning has been applied in different sectors of the healthcare industry such as research, clinical trials, personalized treatment, medical imaging diagnostics, ML-based behavioral modification, smart health records, and etc. (Verma & Verma, 2021).

So, the development of convolutional neural network and logistic regression to predict the drug resistance from the first-line antituberculosis drugs of TB samples will be performed in this project. The accuracy, sensitivity, and specificity of both of the algorithms will be compared to assess which algorithm is suitable for the MDR-TB prediction for the four first-line drugs. Convolutional neural network was hypothesized to outperform logistic regression 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 hope to give promising results in advancing the diagnosis of MDR-TB in Indonesia.

1.2. Objectives

The main objective of this research is to create predictive models using logistic regression and convolutional neural network that have high accuracy, sensitivity, and specificity to predict resistance against the first-line antituberculosis drugs and compare between both models to find the best model to predict the drug resistance. It would be done by conducting these following objectives:

1. Implementation of logistic regression to create a predictive model to predict resistance against first-line antituberculosis drugs using non-phylogenetic SNPs as features.
2. Implementation of convolutional neural network to create a predictive model to predict resistance against first-line antituberculosis drugs using gene sequences as the input.
3. Assessment of accuracy, sensitivity, and specificity using confusion matrix for each first-line antituberculosis drugs.

4. Comparison of accuracy, sensitivity, specificity, and computational complexity of detecting drug resistance between both models to find the best model for drug resistance prediction of the first-line antituberculosis drugs.

1.3. Research Scope

The scope of this study for both logistic regression and convolutional neural network predictive models will involve:

1. Raw whole genome sequence of *Mycobacterium tuberculosis* retrieval from databases
2. NGS secondary analysis to form VCF files
3. Extraction of gene sequence for convolutional neural network input sequence
4. Preprocessing and model fitting for logistic regression (LR) and convolutional neural network (CNN)
5. Cross validation and confusion matrix to assess accuracy, sensitivity, and specificity of both models for each drug
6. Assess the computational complexity by measuring CPU utilization, training time, and prediction time
7. Comparison between both models' performance based on the accuracy, sensitivity, specificity, and computational complexity as parameters for benchmarking