INTRODUCTION

The assessment of air quality in a particular area relies heavily on the Air Quality Index (AQI) [1], [2], which is determined by measuring different air pollutants and evaluating their potential health effects [3]. Jakarta, the capital of Indonesia, holds the unenviable record of having the highest global air pollution levels. According to IQAir, a Swiss air filtration company that monitors worldwide air quality, Jakarta ranks fifth in AQI scores [4].

As per the 2019 World Health Organization (WHO) report, around 4.2 million premature deaths occur annually worldwide due to air pollution [5]. Additional research [6] and [7] delineate the primary origins and adverse consequences of air pollution, including impacts on respiratory health and other diseases and its role in contributing to global warming. Commonly identified sources of air pollution involve household combustion tools, motor vehicles, industrial facilities, and occurrences like forest fires [8].

Several studies have been conducted to predict Jakarta's Air Quality Index (AQI). In 2022, Permai et al. [9] conducted research using Support Vector Machine (SVM) and Multilayer Perceptron (MLP). The results showed impressive accuracy levels of 98% and 92%, respectively. The following year, Muljana et al. [10] conducted a study using Random Forest Classifier (RFC), achieving an outstanding accuracy rate of 95%. In the same year, Rafif et al. [11] conducted research using Decision Tree and SVM, obtaining accuracy rates of 87.86% and 90.56%, respectively.

Although the achieved accuracy in the mentioned studies ([9], [10], [11]) is relatively high, these studies still have some shortcomings. Firstly, these studies did not perform normalization on their datasets. Normalization is a crucial aspect of data processing, aiming to create uniformity and ensure variables have a consistent scale [12], [13]. This precautionary measure is vital to avoid potential issues such as unwanted dominance or bias in experimental results [14], [15].

Secondly, studies [10] and [11] did not include hyperparameter tuning, which is a crucial key to optimizing model performance [16], [17]. On the other hand, the study [9] only performed hyperparameter tuning on the MLP model, and the specific method used was not detailed. Additionally, the [9] research did not incorporate k-fold cross-validation in evaluating the performance of their models. Using k-fold cross-validation enhances result reliability by assessing the models' performance across multiple subsets of the dataset [18].

In addition to the previously mentioned limitations, there are crucial aspects related to the data used in the studies [9], [10], [11]. All three studies utilized datasets from the same source, namely the official Jakarta Open Data website. Notably, these studies did not incorporate the most recent complete dataset available, specifically the 2021 dataset accessible on Jakarta Open Data. This latest dataset introduces a new parameter, PM2.5, which was absent in the datasets used in studies [9], [10], [11].

Lastly, an important point is that those studies did not explore utilising deep learning algorithms for predicting the Air Quality Index (AQI) in Jakarta. While they achieved commendable accuracy using Support Vector Machine (SVM), Multilayer Perceptron (MLP), Random Forest Classifier, and Decision Tree, the absence of deep learning algorithms in their experiments represents a notable gap. Integrating deep learning methodologies could potentially lead to substantial improvements in AQI prediction results, given the ability of deep learning models to capture intricate patterns and dependencies within complex datasets [19]. Considering deep learning approaches could offer a more sophisticated and nuanced understanding of air quality dynamics in Jakarta.

To address the identified shortcomings in previous studies, we propose an air quality detection system in Jakarta based on the Long Short-Term Memory (LSTM) deep learning algorithm in this study. It leverages the LSTM architecture and enhances its performance through Bayesian Optimization (BO). The proposed system, APD-BayTM, is designed to predict Jakarta's Air Quality Index (AQI) accurately.

The main contribution of this study is:

- Implementing of the Long Short-Term Memory (LSTM) model within the APD-BayTM.
- Parameter tuning was incorporated through Bayesian Optimization to enhance their effectiveness in predicting Jakarta's Air Quality Index (AQI). This approach is designed to improve model performance, leading to more accurate predictions of AQI.
- Utilizes the most recently available dataset from the Jakarta Open Data website, encompassing essential parameters such as PM2.5. This ensures the dataset's relevance to the current air conditions in Jakarta.
- Includes the normalization of the dataset to uphold consistency and ensure that variables possess a consistent scale. Normalization is essential to prevent unwanted dominance or bias in the experimental outcomes.

By combining the optimized LSTM model with the latest dataset and normalization, this study aims to improve accuracy in Jakarta's air pollution detection significantly.

The APD-BayTM system's anticipated success is poised to positively impact the comprehension and management of air quality challenges in metropolitan cities. This study is expected to contribute significantly to the existing air quality prediction literature by prioritising deep learning methods and parameter optimisation. Furthermore, the success of this system is likely to inspire the development of more sophisticated and accurate prediction methods in the future, further advancing the field of air quality assessment and management.