

CHAPTER I

INTRODUCTION

This chapter presents the background of the research problem, the formulation of the research questions, the research objectives and benefits, as well as the scope and limitations of this study.

1.1 Background

Determining the etiology in patients with acute fever in tropical regions often presents a complex diagnostic challenge due to identical and overlapping initial clinical symptoms. [1]. Fever itself is a symptom, not a diagnosis. The term “etiology” in this context refers to the causative pathogen behind a particular episode of fever. Acute fever is defined as an increase in body temperature lasting less than 14 days without signs of infection in a specific organ upon initial clinical evaluation, a condition known in the tropical medicine literature as Acute Undifferentiated Febrile Illness (AUFI). Thus, the etiology of acute fever refers to the effort to identify the causative pathogen behind an episode of acute fever that appears clinically similar, even though it is actually caused by different pathogens. The diversity of pathogens causing AUFI, ranging from viruses and bacteria to Rickettsia, makes determining the etiology a diagnostic challenge, particularly in South and Southeast Asia [2]. As of the third quarter of 2024, the Indonesian Ministry of Health has recorded over 203,900 dengue cases and more than 1,200 deaths nationwide, with East Java being one of the regions contributing the highest number of cases [3]. Regarding bacterial etiology, Indonesia is recorded as the country with the highest burden of mortality and disability-adjusted life years (DALYs) in Southeast Asia, according to the Global Burden of Disease study [4]. Meanwhile, regarding Rickettsial etiology, the absence of up-to-date national data actually underscores a particular urgency, as the disease burden in Indonesia is reported to remain "almost wholly unknown" [5]. For non-Dengue viral etiologies, this challenge of overlapping symptoms is clinically evident. A cohort study

reported that among all laboratory-confirmed cases of Influenza virus infection, not a single one was accurately diagnosed during hospital treatment [6].

Serological tests and laboratory parameters sometimes face challenges due to overlapping clinical symptoms [7]. A study of 1,486 patients with acute fever who were hospitalized at eight referral hospitals in Indonesia (AFIRE) also confirmed this complexity: while dengue was the most common cause (46.6%), more than half of the remaining cases were caused by non-dengue etiologies, and Rickettsia cases in the study went undiagnosed through standard testing during the hospitalization period [8]. Approximately 35% of all cases require more in-depth data analysis to reach an optimal diagnosis during hospitalization, opening up opportunities for the implementation of machine learning to identify hidden nonlinear patterns as a supplementary analysis tool for medical professionals [9].

However, the development of machine learning models for multiclass classification often faces a fundamental problem: class imbalance (imbalanced data) [10]. Because most endemic diseases generally have a much larger number of samples than other diseases, algorithms trained under these conditions are prone to predictive bias that increases the false negative rate for the minority class [11]. Therefore, optimizing classification performance is particularly important in cases of high symptomatic overlap, to ensure that the model can perform feature mapping objectively [12]. On the other hand, the complexity of classification in the acute fever domain is not limited to the issue of data imbalance, where multiclass classification approaches yield more moderate performance compared to binary classification approaches [13]. Studies in Indonesia have indeed shown high accuracy in predicting dengue, but they are still limited to binary classification and have not yet taken other etiologies into account simultaneously [14]. Therefore, research on strategies for handling imbalanced data and model optimization for multiclass classification is needed to bridge this performance gap.

To address this class imbalance, the Synthetic Minority Over-sampling Technique for Nominal and Continuous (SMOTE-NC) was applied, which preserves the integrity of categorical data through a majority-vote method, unlike standard SMOTE, which can corrupt medical data in text form [15]. The balanced data was then modeled using the CatBoost algorithm, which was chosen for its

ability to map probabilities among pathogens based on similarities in clinical features through the Oblivious Decision Tree and Ordered Boosting architectures, while minimizing the risk of overfitting in heterogeneous medical datasets [16]. Since CatBoost's performance depends on hyperparameter configurations such as tree depth, learning rate, and regularization coefficient, Optuna is used with a Bayesian optimization strategy to find parameter combinations adaptively and more efficiently than conventional grid or random search methods. This combination of SMOTE-NC, CatBoost, and Optuna is expected to optimize detection sensitivity for minority classes [17].

To bridge the gap between metric simulation and practical implementation, the CatBoost model which has been optimized using Optuna will be integrated into a server environment using the FastAPI framework [18]. The choice of this server-side inference architecture is intended to improve computational efficiency without overburdening devices and to facilitate centralized data updates [19]. The output consists of classification probabilities, which are then displayed through a mobile app interface built using the Flutter framework as a prototype for practical implementation [20].

1.2 Problem Formulation

Based on the background described above, the research questions in this study are:

1. How effective is the application of SMOTE-NC in addressing the class imbalance problem in the multiclass classification of acute fever etiologies?
2. How do the accuracy and reliability of the classification metrics (Recall, Precision, F1-Score) of the CatBoost algorithm which has been hyperparameter-optimized using Optuna compare to those of the baseline model in identifying the etiology of acute fever?
3. How can the classification model be implemented on a client-server architecture using FastAPI and Flutter as a prototype for a mobile-based decision support system for the etiology of acute fever?

1.3 Research Objectives

The objectives of this study are as follows:

1. To analyze the application of SMOTE-NC in addressing the class imbalance problem in the multiclass classification of acute fever etiologies.
2. To analyze the accuracy and reliability of the classification metrics (Recall, Precision, F1-Score) of the CatBoost algorithm which has been hyperparameter-optimized using Optuna compared to a baseline model in identifying the etiology of acute fever.
3. To develop a prototype of a mobile-based decision support system that integrates the classification model via server-side services to provide real-time probability information on etiologies.

1.4 Research Benefits

This research is expected to provide the following benefits:

1. For the Field of Computer Science: To provide a reference on the integration of the Synthetic Minority Over-sampling Technique for Nominal and Continuous (SMOTE-NC) method, the Categorical Boosting (CatBoost) algorithm, and hyperparameter optimization using Optuna in addressing class imbalance (imbalanced data) in medical tabular datasets with heterogeneous features (numeric and categorical).
2. For Healthcare Professionals: The resulting decision support system is explicitly positioned as a tool for initial screening (triage) and second opinions for healthcare professionals. This system is not intended to replace a final clinical diagnosis. Clinical validation by healthcare professionals remains mandatory before a definitive diagnosis is established.

1.5 Research Limitations

To ensure this study remains focused and well-defined, the following research limitations have been established:

1. The mobile application being developed requires an internet connection because the machine learning algorithm computations are executed on a server.
2. This study uses secondary data sourced from the Acute Febrile Illness Requiring Hospitalization (AFIRE) study, published via the Figshare repository by Nurmawati et al. (2024), which includes medical records from

8 referral hospitals in Indonesia for the period July 2013–June 2016; therefore, this study does not involve the direct collection of primary data.