

# CHAPTER I

## INTRODUCTION

This opening chapter establishes the structural and conceptual foundation of the study. It outlines the research background, defines the core problem, and states the target objectives. Additionally, it highlights the practical value of the work and sets the boundaries of the investigation, providing a clear roadmap for the entire project.

### 1.1 Background

Hypertension, commonly known as high blood pressure, remains a critical public health crisis globally and nationally. The sheer scale of the issue is staggering. According to the World Health Organization (WHO), the number of adults aged 30–79 living with hypertension doubled from 650 million in 1990 to 1.3 billion in 2019. This figure was projected to climb to 1.4 billion by 2024, accounting for roughly a third of the population within that age bracket [1][2]. By 2025, the global count was expected to reach 1.5 billion, hitting resource-constrained regions the hardest. Southeast Asia faces a similar burden; nearly one-third of its population suffers from the condition, leading to an estimated 1.5 million deaths every year [3].

Alarmingly, the widespread nature of hypertension contrasts sharply with low rates of awareness and medical management. Globally, the WHO reports that only 54% of individuals with hypertension are properly diagnosed, just 42% receive medical treatment, and a mere 21% manage to keep their blood pressure under control [1]. Indonesia mirrors this troubling trend. The 2018 National Basic Health Research (Riskesdas) highlighted a sharp rise in adult prevalence to 34.1%, up from 25.8% in 2013. More recently, the 2023 Indonesian Health Survey (SKI) recorded an overall prevalence of 30.8%. However, this national average masks stark regional disparities, ranging from 22.2% in Papua to a staggering 44.3% in South Kalimantan [3][4]. Crucially, the disease is no longer confined to older generations. It is increasingly creeping into younger populations, affecting 10.7% of young adults aged 18–24 and 17.4% of those aged 25–34 [5]. This shift toward the productive workforce underscores the urgent need for early risk detection.

Managing hypertension across Indonesia involves navigating a complex web of geographic and sociodemographic hurdles. The country's vast, fragmented geography and diverse educational levels mean that health literacy and medical access remain unevenly distributed. Because hypertension is famously a "silent killer" that develops without obvious warning signs, individuals rarely seek early screening or timely treatment. This lack of awareness, combined with limited healthcare infrastructure in remote areas, leaves vulnerable groups particularly the elderly exposed to severe risks. Unchecked high blood pressure damages more than just the cardiovascular system; it frequently triggers debilitating complications like heart disease, chronic kidney disease, and cognitive decline. While the government has launched various public health initiatives, early detection remains a significant roadblock, leaving countless cases undiagnosed. To bridge this gap, modern healthcare requires an adaptive, data-driven approach capable of pinpointing high-risk individuals before complications manifest [3][4].

Previous studies have utilized machine learning for hypertension prediction and screening, demonstrating strong performance. Estiko et al. conducted a large-scale study involving 268,210 participants in Indonesia and compared Random Forest, Gradient Boosting, and XGBoost for hypertension screening. XGBoost emerged as the best model with a sensitivity of 0.97 and an AUC of 0.75 based on 11 non-laboratory risk factors. This study demonstrates that non-laboratory risk factors hold strong potential for use in hypertension screening among the Indonesian population [6]. A study by Murtiningsih and colleagues evaluated LightGBM, XGBoost, and CatBoost using balancing techniques and hyperparameter tuning. As a result, CatBoost demonstrated the highest performance when combined with SMOTE and GridSearchCV, achieving an accuracy rate of 99.2% [7]. Harmoni and colleagues optimized the CatBoost algorithm using Optuna based on TPE for the classification of anti-tuberculosis drugs. The study results showed that hyperparameter optimization using Optuna was able to increase CatBoost's accuracy from 90% to 96%, thereby proving Optuna's effectiveness in improving model performance on medical data with both numerical and categorical features [8].

Despite these promising advances, key gaps remain in predictive modeling for hypertension. While CatBoost has proven its worth across several studies, it has not yet been combined with Optuna optimization to predict hypertension risk specifically through the lens of lifestyle choices. Furthermore, existing research frequently relies on traditional tuning methods like Grid Search. These approaches are computationally heavy and often struggle to navigate complex parameter spaces efficiently. To overcome this, this study introduces Optuna as a highly adaptive framework to pinpoint the most effective hyperparameter configurations. Merging CatBoost with Optuna offers a faster, more reliable path toward generating stable and highly accurate risk assessments.

Consequently, this study develops and refines a lifestyle-focused hypertension prediction model by integrating the CatBoost algorithm with the Optuna optimization framework. The core of this research involves using Optuna to fine-tune the algorithm's hyperparameters, seeking the optimal balance for predictive performance. To gauge success, the model will be evaluated across key metrics: accuracy, precision, recall, and F1-score, allowing for a clear before-and-after comparison. To guarantee stability and prevent overfitting, the study implements 5-fold cross-validation across three distinct data-splitting ratios: 90:10, 80:20, and 70:30. Ultimately, this work aims to offer fresh insights into how advanced machine learning can be practically applied to preventative healthcare.

## **1.2 Research Problem**

Driven by the clinical and technical challenges outlined in the background, this study addresses two central questions:

1. How effectively can a CatBoost model forecast hypertension risks when trained on diverse lifestyle and behavioral patterns?
2. In what ways, and to what extent, does automated hyperparameter tuning via Optuna alter the predictive performance of this CatBoost model?

## **1.3 Research Objectives**

To systematically answer these questions, the empirical goals of this research are structured as follows:

1. To critically evaluate the diagnostic accuracy of the CatBoost framework in mapping specific lifestyle inputs to hypertension risk levels.

2. To quantify the precise impact of Optuna's optimization process on the algorithm's overall operational metrics.

#### **1.4 Benefits**

By bridging the gap between machine learning and preventative healthcare, this research offers distinct academic and practical contributions.

1. **Academic Value:** The study provides a clear, empirical benchmark demonstrating how gradient boosting algorithms handle complex, non-linear lifestyle data. It offers deep insights into the mechanics of automated hyperparameter selection.
2. **Practical Utility:** For data scientists and healthcare technologists, this work illustrates exactly how targeted algorithmic adjustments can improve early risk-detection tools, paving the way for more reliable clinical decision-support systems.

#### **1.5 Scope and Limitations**

To maintain methodological rigor and analytical clarity, the boundaries of this project are defined by the following framework:

1. **Predictor Features:** The input data is strictly limited to ten specific variables covering demographic, physiological, and behavioral attributes: Age, Salt\_Intake, Stress\_Score, BP\_History, Sleep\_Duration, BMI, Medication, Family\_History, Exercise\_Level, and Smoking\_Status.
2. **Target Outcome:** The model evaluates a single binary target, Has\_Hypertension, classified exclusively as either Yes or No.
3. **Modeling and Fine-Tuning:** The core classifier relies entirely on the CatBoost architecture. Optimization is restricted to the automated tuning of four core hyperparameters via Optuna: iterations, learning\_rate, depth, and l2\_leaf\_reg.
4. **Testing and Metrics:** Model stability is guarded using a 5-fold cross-validation strategy. Success is measured comprehensively through a standard classification matrix: accuracy, precision, recall, and F1-score.
5. **Generalization Bound:** The resulting model is highly contextualized by the baseline distribution of the chosen dataset. As a result, its predictive validity applies specifically to patient populations with statistical characteristics matching the training data.