CHAPTER I
INTRODUCTION

1.1. Research Background

Stress has become a global mental health phenomenon and affects various aspects of individual life, from work productivity, quality of social relationships, to physical health conditions[1] . Unmanaged stress levels can develop into anxiety disorders, depression, and even chronic health problems such as cardiovascular disease [2] . In the context of modern, fast-paced society, the ability to detect and predict stress levels early is an important need for *preventive intervention* .

Sleep conditions have a close reciprocal relationship with stress levels. Insufficient or poor-quality sleep can be both a cause and a consequence of stress experienced by individuals[3] . Scientific research has demonstrated a strong correlation between psychological stress levels and variables like sleep duration, sleep quality (efficiency), sleep disorders (insomnia), and irregular sleep habits. However, there exists a nonlinear and intricate association between stress levels and sleep-related variables, often influenced by interactions among multiple elements. [4] .

Traditional stress assessments often rely on psychometric questionnaires such as the Perceived Stress Scale (PSS) or the DASS-42, which require self-completion by respondents and are prone to subjective bias [5] . In addition, these manual approaches are impractical for large-scale implementation and do not allow for *real-time analysis.* Beside that, the increasingly easy collection of sleep data through *wearable devices* or mobile applications offers the potential to develop more objective and efficient prediction systems [6] .

The complexity of the non-linear relationship between sleep state variables and stress levels demands sophisticated analytical approaches. Traditional statistical methods are often inadequate to capture this complexity. In this case, Because boosting-based *machine learning algorithms* can handle complicated feature interactions and non-linear correlations, they become a viable solution in this situation[7] .

Two popular ensemble algorithms are *XGBoost* (Extreme *Gradient boosting*) and *Random forest* (RF). *XGBoost* is widely known for its ability to handle complex and high-dimensional data efficiently, and uses L1 and L2 regularization techniques

to prevent *overfitting* [8] . In research by [9] . *XGBoost* , optimized with Cuckoo Search Algorithm (CSA), successfully achieved 92% accuracy and 91% F1 score, demonstrating superior performance in detecting contamination in coconut milk products compared to *Random forest* . Meanwhile, *Random forest* (RF), which relies on *bagging techniques* , has been proven to be stable and easy to interpret, so it is often used in medical applications and spectral data analysis [8] . When compared to boosting models like *XGBoost*, RF frequently loses accuracy but is better at creating more comprehensible and reliable models. According to research by [8], both models exhibit extremely high accuracy above 96%, however *XGBoost* marginally surpasses RF in terms of positive case detection in cervical cancer risk prediction. *Random forest's* advantages in stability and interpretability make it a popular choice in applications that require more transparent model explanations. On the other hand, *XGBoost* excels at addressing issues with imbalanced data and iteratively correcting prediction errors, making it more suitable for tasks that require progressive accuracy improvements [9] .

However, there aren't many research that explicitly compare the *Random Forest* (RF) and *XGBoost algorithms'* performance when it comes to predicting stress levels based on sleep conditions. Prior research has mostly concentrated on alternative techniques like *Support Vector Machine* (SVM) and *Logistic Regression.* To find out which algorithm is better for this particular problem domain in terms of accuracy, precision, *recall*, F1-score, and computing efficiency, a thorough comparison between RF and XGBoost is actually required.

This work seeks to fill the existing gap by thoroughly evaluating the predictive performance of Random Forest and XGBoost models based on sleep state characteristics. By comparing these two state-of-the-art algorithms, which possess different characteristics—where *Random Forest* excels in stability and interpretability, while *XGBoost* has proven to be more flexible and efficient in handling large and complex datasets,this study is expected to provide practical contributions toward the development of precise and efficient systems for early stress detection. Furthermore, this study also aims to provide empirical guidance for researchers and practitioners in selecting the most optimal algorithm for sleep-based mental health prediction problems.

1.2. Research Problem

This background allows for the formulation of a number of research questions that will be addressed in the undergraduate thesis, such as:

1. How to build *Random forest* and *XGBoost models* to forecast stress levels based on sleep conditions?
2. How effectively do the Random Forest and XGBoost models classify stress levels based on sleep conditions in terms of accuracy, precision, recall, F1-score, and computational time?

1.3.Research Purposes

The objectives of the thesis entitled "Predicting Stress Levels Based on Sleep Conditions Using *the Random Forest* and *XGBoost Methods* " are as follows:

Building and implementing a stress level prediction model based on sleep conditions using two algorithms, namely *Random forest* (*bagging*) and *XGBoost* (*boosting*).

1.4. Benefits of Research

Some of the benefits that can be taken from the thesis entitled " Predicting Stress Levels Based on Sleep Conditions Using *the Random Forest* and *XGBoost Methods* " include:

1. Offers a comparative evaluation of how *the Random forest* (*Bagging- based*) and *XGBoost* (*Boosting- based*) algorithms can be applied to process complex sleep condition variables and non-linear relationships with stress levels, resulting in accurate prediction models and showing the differences in characteristics of the two ensemble approaches.
2. Serve as a reference in selecting effective mental health prediction algorithms based on *bias-variance tradeoff considerations* . The research results can be used as an empirical consideration in choosing between *Random Forest* (with the advantages of stability and resistance to *overfitting*) and *XGBoost* (with the advantages of maximum accuracy and strict regularization) based on the effectiveness of evaluation metrics within the scope of sleep-based stress prediction.
3. The resulting predictive model becomes an application that helps individuals or health professionals identify stress risks earlier.

1.5. Research Limitations

The limitations of this study entitled "Stress Level Prediction Based on Sleep Conditions Using Random Forest and XGBoost Methods" are as follows:

1. *The dataset* employed in this research is publicly available. obtained from the Kaggle platform consisting of physiological parameters of sleep and stress levels, rather than primary psychological survey data or more comprehensive clinical data.
2. The scope of modeling is limited to 8 physiological sleep predictor variables (*Snoring rate* , *Respiration rate* , *Body temperature* , *Limb movement* , *Blood oxygen* , REM, *Sleeping hours* , *Heart rate*) with 1 target variable (Stress Level) in categorical form, where the comparison is only carried out comparing *the Random forest* and *XGBoost algorithms* and their performance measurement is based on commonly used classification metrics (accuracy, precision, *recall* , F1-score).
3. Implementation is limited to website application using the *streamlit library* , not in the form of an application other than a website.