CHAPTER 1

INDTRODUCTION

Tuberculosis (TB) remains a serious global health concern, particularly when co-occurring with Human Immunodeficiency Virus (HIV) infection. Diagnosing TB in HIV-positive individuals is especially challenging due to the limitations of current diagnostic techniques. Meanwhile, advancements in gene expression analysis and machine learning have opened up new possibilities for improving early and accurate TB detection. This study seeks to develop a more effective prediction model by combining ensemble learning with optimization methods, focusing on TB diagnosis among HIV-infected patients. The detailed background and motivation of this research are presented in the following section.

1.1 Background

Tuberculosis (TB) is a disease caused by Mycobacterium tuberculosis (MTB), one of the leading infectious diseases globally, is spread through droplet transmission, which occurs when a person with pulmonary TB coughs up bugs. In general, a relatively small proportion of people infected with MTB develop active TB, but this proportion is much higher in people with impaired immunity [1]. TB is a significant global issue, with approximately 30% of the world's population infected. It is estimated that there are nine million new TB cases each year worldwide [2], [3]. In the early 1990s, a drug-resistant TB strain caused an outbreak in New York, killing 80% of infected patients [4]. By 2020, Human Immunodeficiency Virus (HIV) coinfection with TB had become the primary cause of death, with an estimated 20 million fatalities among 37.7 million individuals infected with both [5], [6].

Due to the high risk of TB in individuals with HIV, early and accurate detection is critical [5]. Two main approaches have been employed as solutions: RNA-based techniques and microscopic and culture-based methods. RNA-based techniques, particularly messenger RNA (mRNA) microarray studies, analyze the whole transcriptome to identify host gene expression signatures associated with disease progression, offering a molecular-level insight into infection status [5], [7]. Meanwhile, microscopic examination and culture-based analysis detect Mycobacterium tuberculosis directly from sputum samples, allowing visualization and growth of the bacteria to confirm infection [8], [9]. However, both approaches have significant limitations. RNA-based methods lack gene signatures specifically capable of identifying patients co-infected with HIV and TB, limiting their diagnostic precision in this critical group [8]. Microscopic and culture-based methods suffer from low sensitivity, especially in HIV co-infected patients, children, and extrapulmonary TB cases; they also cannot detect drug resistance effectively and require long processing times [5], [9], [10].

To address these limitations, machine learning techniques applied to gene expression data offer a powerful and innovative approach for improving TB detection in HIV-infected patients [5], [8].

Early detection of tuberculosis (TB) in HIV patients is very important because TB infection can worsen the patient's condition and may even be fatal. Data-driven approaches, particularly those using gene expression data, have shown potential in improving the accuracy of TB diagnosis in HIV patients. Several studies [11], [12], [13] have investigated TB prediction in HIV patients by applying machine learning methods to gene expression data. Although the results are promising, the number of studies specifically focusing on TB prediction in HIV patients remains limited. One of the main challenges in this research is the high dimensionality of gene expression data, which can increase model complexity and reduce predictive performance. Therefore, effective feature selection methods are needed to address this high dimensionality issue in order to improve the efficiency and accuracy of the prediction system [14].

Recent studies have been proposed to address the issue of high dimensionality in gene expression data through various dimensionality reduction and feature selection techniques [14], [15], [16]. These approaches generally aim to eliminate irrelevant or redundant features while preserving the most informative ones, to improve model performance and reduce computational complexity. However, the findings of these studies [14], [15], [16], indicate that existing methods are still limited in capturing the complex and nonlinear patterns inherent in gene expression profiles. Many traditional techniques still rely on fixed selection criteria or make simplifying assumptions, which may limit their ability to fully represent complex biological patterns present in the data.

In addition to these methodological limitations, several general challenges also persist. First, the number of genes selected for analysis often varies greatly between studies—from a few to hundreds—which makes it difficult to achieve consistent and generalizable results [14]. Second, many models tend to perform well only on specific datasets or populations, reducing their reliability when applied to new or unseen data [15]. Third, some advanced techniques, although powerful, often require large computational resources or are difficult to interpret, making them less practical for real-world clinical use, especially when working with limited sample sizes, as is often the case in TB-HIV co-infection studies [16].

Consequently, high dimensionality remains a persistent challenge, especially in tasks such as TB detection among HIV-infected individuals. This underscores the need for more flexible and adaptive approaches that can better manage high-dimensional data and enhance classification accuracy. To address the limitation above, alternative approaches such as ensemble learning methods have been explored as promising solutions [17]. These methods are capable of improving classification performance by combining multiple models, which helps in capturing complex patterns and enhancing robustness, particularly in high-dimensional datasets like gene expression data [17], [18].

One of the most popular ensemble techniques is Bagging (Bootstrap Aggregating), which trains multiple models on randomly resampled subsets of the data and combines their predictions through voting or averaging [19]. While effective in reducing variance, conventional Bagging assigns equal weights to all models without considering their individual performance. To address this issue, several approaches have proposed weighting strategies in tuberculosis prediction based gene expression data [20], [21], [22]. However, many Bagging implementations still rely on uniform weighting, assuming that all base learners contribute equally to the final decision. This can reduce overall performance, especially when dealing with complex and high-dimensional data such as gene expression data.

To overcome this limitation, more adaptive solutions are needed. Various metaheuristic algorithms have been widely explored for classification and feature selection due to their capability to search large, complex solution spaces inspired by natural behaviors. Algorithms such as Genetic Algorithms (GA), Particle Swarm Optimization (PSO), and Ant Colony Optimization (ACO) have been successfully applied in these domains, particularly in optimizing ensemble models. One increasingly prominent and promising approach in this area is the Firefly Algorithm (FA), which has attracted growing interest due to its demonstrated effectiveness in solving complex optimization problems, particularly for its ability to efficiently explore nonlinear and multimodal search spaces, maintain solution diversity, and avoid local optima [23–26]. Its unique brightness-attraction mechanism allows it to explore diverse solutions and avoid local optima more effectively than many alternatives. FA offers flexible and powerful mechanisms for navigating high-dimensional optimization problems, making it especially effective in tasks such as ensemble model weighting based on gene expression data. By dynamically adjusting weights according to individual model performance, FA-enhanced ensembles can significantly improve accuracy and robustness, making them well-suited for complex classification problems such as tuberculosis (TB) prediction.

This study aims to predict tuberculosis in HIV patients using a Firefly Algorithm-based weighted ensemble, which optimizes the weights of base models to improve predictive performance and avoid local optima. By utilizing FA for weight optimization, the ensemble model can achieve higher accuracy and robustness, effectively handling the complexity and variability of gene expression data in predicting TB in HIV patients. This method is expected to demonstrate superior performance by dynamically adjusting weights to optimize model accuracy. This approach enhances feature selection and ensures that the solution converges to a global optimum, rather than becoming trapped in local optima, thereby providing a more reliable prediction model for clinical applications. The Firefly Algorithm has been successfully applied in various optimization problems, including feature selection and fault detection, where it aids in discriminative feature selection in classification and regression models [21]. Its ability to handle multimodal and nonlinear optimization prob-

lems makes it suitable for complex data scenarios [27]. By integrating the Firefly Algorithm into the weighted ensemble framework, this study seeks to enhance the predictive modeling of tuberculosis in HIV patients, leveraging the algorithm's strengths in optimization to address the challenges posed by complex gene expression data.

1.2 Problem Statement

Tuberculosis (TB) remains a significant public health concern, particularly among individuals co-infected with Human Immunodeficiency Virus (HIV), where compromised immune function complicates early and accurate diagnosis. Conventional diagnostic methods, including RNA-based and culture-based techniques, have shown limitations such as low sensitivity, long processing times, and reduced reliability in detecting TB among immunocompromised patients [5], [9], [10]. In response, machine learning techniques applied to gene expression data have emerged as a promising approach for enhancing diagnostic precision. However, the high dimensionality and complexity of gene expression data present challenges in feature selection and model generalization [5], [8].

While machine learning classifiers such as Support Vector Machine (SVM), Random Forest, and deep learning models have been utilized to predict TB in HIV-positive individuals, their performance remains inconsistent. This is often due to overfitting and inadequate feature selection strategies. Ensemble learning methods have been proposed to improve prediction accuracy and robustness by combining multiple classifiers. However, many ensemble models apply equal or fixed weights to all base learners, which assumes that each model contributes equally to the final prediction. This assumption can result in suboptimal performance, especially when some models perform better than others.

To improve ensemble performance, weight optimization strategies have been introduced using metaheuristic algorithms. These algorithms adaptively assign weights to models based on their performance, allowing more accurate predictions. Among various metaheuristics, the Firefly Algorithm (FA) is particularly promising due to its effectiveness in high-dimensional optimization problems and its ability to avoid local optima by exploring diverse solutions. This study is therefore guided by the following research questions:

1. How does a baseline ensemble model without weight optimization perform in predicting TB in HIV-infected individuals?

2. What is the impact of different Firefly Algorithm (FA) parameter settings on the performance of the ensemble model?

3. How does the Firefly Algorithm (FA), as an optimization strategy, compare to baseline ensemble models in improving predictive performance for TB detection in HIV-positive patients?

By answering these questions, the study aims to evaluate the effectiveness of a Firefly Algorithm-based weighted ensemble model in improving accuracy, robustness, and generalization in TB classification using gene expression data.

1.3 Objective and Hypotheses

1.3.1 Objectives

The primary objective of this study is to develop and evaluate a Firefly Algorithm-based weighted ensemble model for predicting tuberculosis (TB) in HIV-infected patients using high-dimensional gene expression data. This approach aims to enhance diagnostic accuracy and model robustness by addressing the limitations of conventional diagnostic methods, overcoming the high dimensionality of gene expression data, and improving ensemble learning performance through weight optimization. To achieve this goal, the study is designed with the following specific objectives:

1. To evaluate the predictive performance of a baseline ensemble model without weight optimization in classifying TB among HIV-infected individuals.
2. To assess the effectiveness of various Firefly Algorithm (FA) parameter configurations in the optimization process, and their impact on improving classification performance within the ensemble model.
3. To investigate the impact of applying the Firefly Algorithm (FA) for weight optimization within the ensemble framework, and to compare its performance against non-optimized and classically optimized ensemble models in terms of accuracy and robustness.

1.3.2 Hypotheses

Based on objectives and the reviewed literature, this study hypothesizes that a baseline ensemble model without weight optimization is expected to produce only moderate predictive performance in classifying TB among HIV-infected individuals. By incorporating classical metaheuristic algorithms for weight optimization, the classification accuracy and robustness of the ensemble models can be improved. Furthermore, the Firefly Algorithm (FA), when used for weight optimization in ensemble learning, is expected to significantly outperform both baseline and classically optimized ensemble models, particularly in predicting TB in HIV-positive patients using gene expression data. In addition, it is hypothesized that the performance of the FA-based ensemble model is strongly influenced by its parameter settings; appropriate combinations of FA parameters are expected to enhance the optimization process, whereas suboptimal settings may hinder model convergence or reduce predictive accuracy. Overall, the Firefly Algorithm-based weighted ensemble model

is hypothesized to deliver superior performance in terms of accuracy and robustness, especially in handling high-dimensional gene expression data and addressing the complexity of TB detection in HIV co-infection.

1.4 Justification for Research

Based on the aforementioned issues, this study proposes the use of a weighted ensemble learning model to improve the accuracy of tuberculosis prediction classification in HIV-infected patients based on gene expression data. Ensemble learning is chosen for its ability to handle high-dimensional data and improve generalization by combining multiple base learners. However, conventional ensemble methods, such as bagging, apply uniform weights to all models, assuming equal contributions. This may lead to suboptimal performance when some models outperform others. To overcome this, weight optimization is introduced using the Firefly Algorithm (FA), a metaheuristic optimization method inspired by the natural behavior of fireflies. FA has demonstrated strong performance in high-dimensional search spaces, enabling the model to dynamically adjust weights and avoid local optima. The architecture and implementation of the proposed FA-based weighted ensemble model are described in detail in Chapter 3, while the experimental results and performance evaluation are presented and discussed in Chapter 4. Thus, integrating FA into the ensemble framework is expected to produce a more accurate and robust predictive model for TB detection in HIV-positive patients.

1.5 Scope and Delimitation

This study focuses on developing and evaluating a Firefly Algorithm-based weighted ensemble model to predict tuberculosis (TB) in HIV patients using high-dimensional gene expression data. The research leverages machine learning techniques, specifically a metaheuristic optimization strategy, to optimize model weights within the ensemble framework, aiming to improve diagnostic accuracy and robustness in classification. The research utilizes publicly available microarray datasets representing gene expression profiles of individuals with HIV and TB co-infection. The study compares the performance of the proposed FA-based ensemble model against both non-optimized ensembles and those optimized using classical metaheuristic algorithms, focusing on its effectiveness in handling high-dimensional biomedical data.

This research is limited to in silico (computational) experiments and does not include clinical or laboratory-based validation. Although it addresses high-dimensionality in gene expression data, the study does not propose new feature selection methods; instead, it relies on preprocessed datasets where dimensionality reduction has already been applied or integrated. Furthermore, while the Firefly Algorithm is central to the study, other metaheuristic algorithms are not directly implemented or compared. The focus is on

evaluating the impact of FA on ensemble weighting rather than on comparing different ensemble architectures or optimization frameworks.

In addition, outcomes are evaluated using computational performance metrics, including accuracy, sensitivity, and robustness, rather than through real-world implementation in clinical environments. By delineating this scope, the study aims to provide methodological insights into improving tuberculosis prediction in HIV-positive individuals through enhanced ensemble learning strategies, contributing to the development of more reliable computational tools in biomedical informatics.

1.6 Significance of the Study

This study aims to support the development of more effective diagnostic methods for tuberculosis (TB) in patients living with HIV by introducing a weighted ensemble model optimized using the Firefly Algorithm. By focusing on gene expression data, the proposed approach is designed to improve both the accuracy and reliability of TB prediction, especially in high-dimensional and complex biomedical datasets. This model addresses some of the common weaknesses found in traditional TB diagnostic techniques, such as low sensitivity and long processing time, which are particularly problematic for immunocompromised patients.

From a research perspective, the study contributes to the growing field of machine learning in healthcare by demonstrating how metaheuristic optimization, specifically the Firefly Algorithm, can be used to enhance the performance of ensemble models. While many previous studies have applied machine learning to gene expression data, few have explored how weight optimization can further improve predictive results, particularly in the context of TB-HIV co-infection. By comparing baseline ensembles, classically optimized ensembles, and the Firefly-optimized model, this study provides a clearer understanding of how optimization strategies affect prediction outcomes.

Although this research is limited to computational analysis and does not involve real-world clinical validation, the results may still offer practical value. The findings could be used as a foundation for developing more accurate and scalable diagnostic tools, helping medical practitioners detect TB more effectively in high-risk populations such as people living with HIV. Ultimately, the study contributes to both academic knowledge and future improvements in digital health technology.

1.7 Structure of the Thesis

This thesis is structured to systematically address the research problem, respond to the formulated research questions, and achieve the defined objectives regarding the prediction of tuberculosis (TB) in HIV-positive individuals using a Firefly Algorithm-based weighted ensemble model.

The study begins in **Chapter 1**, which introduces the background of the investigation by highlighting the limitations of traditional TB diagnostic methods, especially among immunosuppressed patients. Then, the emergence of machine learning (ML) applied to gene expression data is presented as a modern solution. The *problem statement* (Section 1.2) outlines the core challenges, including the high dimensionality of gene expression data and the limited performance of existing ensemble models. From these, a set of research questions and objectives (Sections 1.2 and 1.3) were formulated to guide the study, particularly focusing on the contribution of weight optimization using the Firefly Algorithm (FA).

Chapter 2 presents a comprehensive *review of related literature*, which discusses relevant studies on ML and ensemble learning in biomedical diagnostics, metaheuristic algorithms, and high-dimensional gene expression data. This chapter establishes the theoretical foundation and highlights gaps that this study seeks to address.

Chapter 3 details the *research methodology*, describing the dataset, preprocessing techniques, design of the baseline ensemble model, implementation of the FA-based optimization, and evaluation procedures. This chapter explains how the methodology aligns with the research objectives and how each model variant is evaluated.

Chapter 4 presents the *results and discussion*, structured around the research questions:

1. The performance of the baseline ensemble model without optimization is examined to answer **research question 1/objective 1**.

2. The effect of classical metaheuristic-based weight optimization is analyzed for **research question 2/objective 2**.

3. The proposed Firefly Algorithm-based ensemble is evaluated against the previous models to address **research question 3/objective 3**.

In **Chapter 5**, the thesis concludes with a synthesis of key findings, a reflection on the study's limitations, and recommendations for future research directions. The overall structure ensures coherence between the identified research problem, the applied methodology, and the resulting conclusions, forming a clear and logical flow from start to finish.