A DEEP REINFORCEMENT LEARNING HYBRID ALGORITHM FOR THE COMPUTATIONAL DISCOVERY AND CHARACTERIZATION OF SMALL PROTEINS UTILIZING MYCOBACTERIUM TUBERCULOSIS AS A MODEL

BABALOLA ABDULHAFEEZ OLUWABUNMI

UNIVERSITI SAINS MALAYSIA

2025

A DEEP REINFORCEMENT LEARNING HYBRID ALGORITHM FOR THE COMPUTATIONAL DISCOVERY AND CHARACTERIZATION OF SMALL PROTEINS UTILIZING MYCOBACTERIUM TUBERCULOSIS AS A MODEL

by

BABALOLA ABDULHAFEEZ OLUWABUNMI

Thesis submitted in fulfilment of the requirements

for the degree of

Master of Science

August 2025

ACKNOWLEDGEMENT

In the name of Allah, the most beneficent, the most merciful, I express my gratitude for His blessings upon me. I would like to sincerely thank my main supervisor, Dr. Shuhaila Mat Sharani, for her invaluable feedback and unwavering encouragement, which have played a crucial role in shaping the direction of my research. I am also grateful to Dr. Hazrina Yusof Hamdani, Professor Mohd Firdaus Mohd Raih, and Professor Norazmi Mohd Nor for their meticulous attention to detail and expertise, which have significantly contributed to the quality of my work.

I am deeply thankful to my parents for their unwavering love, encouragement, and sacrifices. Your prayers and support have been a constant source of comfort and motivation for me. Without the collective support and encouragement of these exceptional individuals, this thesis would not have been possible. May Allah grant everyone a prestigious position in Jannah. Ameen.

TABLE OF CONTENTS

LIST OF TABLES

Page

LIST OF FIGURES

Page

LIST OF SYMBOLS

$<$	Less than
$>$	Greater than
$\leq$	Less than or equal to
$\geq$	Greater than or equal to
%	Percentage

LIST OF ABBREVIATIONS

3’ UTR	3’ Un-Translated Region
5’ UTR	5’ Un-Translated Region
AUC	Area Under Curve
smORFs	Small Open Reading Frames
CDS	Coding Sequence
circRNA	circular Ribonucleic Acid
dStop	downstream stop site
dStart	downstream start site
GO	Gene Ontology
intStart	Internal start site
intStop	Internal stop site
log10(FDR)	logarithm to the base 10 of the False
FPKM	Fragments Per Kilobase per Million
TPR	True Positive Rate
FPR	False Positive Rate

UStart	Upstream start site

LIST OF APPENDICES

ALGORITMA HIBRID PEMBELAJARAN TETULANG MENDALAM UNTUK PENEMUAN PENGIRAAN DAN PENCIRIAN PROTEIN KECIL MENGGUNAKAN MYCOBACTERIUM TUBERCULOSIS SEBAGAI MODEL

ABSTRAK

Ramalan dan pencirian tepat bagi rangka bacaan terbuka kecil (smORF) adalah penting untuk memahami peranan fungsinya dalam pengawalan gen dan proses selular. Kajian ini membentangkan pembangunan dan penilaian satu algoritma pembelajaran mesin hibrid baharu yang menggabungkan kekuatan model Random Forest dan Gradient Boosting bagi meningkatkan ketepatan ramalan smORF. Prestasi algoritma hibrid ini dinilai secara menyeluruh dan dibandingkan dengan model individu menggunakan metrik penilaian komprehensif termasuk ketepatan, sensitiviti, spesifisiti dan kawasan di bawah lengkung ROC (AUC). Keputusan menunjukkan bahawa model hibrid mencapai prestasi yang lebih tinggi dengan ketepatan 0.998, sensitiviti 0.998, dan spesifisiti 1.00, sekali gus mengatasi prestasi model Random Forest dan Gradient Boosting secara individu. Selain itu, data transkriptom daripada *Mycobacterium tuberculosis* digunakan untuk mengesahkan ramalan tersebut, menonjolkan kaitan biologi dan potensi aplikasi pendekatan yang dicadangkan dalam biologi pengiraan. Kajian ini menekankan kepentingan gabungan teknik pembelajaran mesin untuk meningkatkan ketepatan ramalan dan menyediakan kerangka kukuh bagi kemajuan penemuan smORF. Walaupun tumpuan diberikan kepada perbandingan antara model individu dan hibrid, kajian ini turut mengenal pasti peluang untuk penanda aras lanjutan terhadap alatan luaran bagi mengesahkan lagi sumbangannya. Penemuan ini menyumbang kepada bidang penyelidikan biologi dan pengiraan, serta

menawarkan pandangan mendalam tentang metodologi ramalan smORF dan aplikasinya.

A DEEP REINFORCEMENT LEARNING HYBRID ALGORITHM FOR THE COMPUTATIONAL DISCOVERY AND CHARACTERIZATION OF SMALL PROTEINS UTILIZING MYCOBACTERIUM TUBERCULOSIS AS A MODEL

ABSTRACT

The accurate prediction and characterization of small open reading frames (smORFs) are critical for understanding their functional roles in gene regulation and cellular processes. This study presents the development and evaluation of a novel hybrid machine learning algorithm that integrates the strengths of Random Forest and Gradient Boosting models to enhance the prediction of smORFs. The performance of the hybrid algorithm was rigorously assessed and compared to the standalone models using comprehensive evaluation metrics, including accuracy, sensitivity, specificity, and area under the receiver operating characteristic curve (AUC). Results demonstrated that the hybrid model achieved superior performance, with an accuracy of 0.998, a sensitivity of 0.998, and a specificity of 1.00, significantly outperforming both the Random Forest and Gradient Boosting models individually. Additionally, transcriptomic data from *Mycobacterium tuberculosis* were utilized to validate the predictions, highlighting the biological relevance and potential applications of the proposed approach in computational biology. This study underscores the importance of combining machine learning techniques to improve prediction accuracy and provides a robust framework for advancing smORF discovery. While the focus was on comparing standalone and hybrid models, the study identifies opportunities for future benchmarking against external tools to further validate its contributions. The findings

contribute to both computational and biological research, offering insights into smORF prediction methodologies and their applications.

CHAPTER 1

INTRODUCTION

1.1 Background of the Study

In recent years, there has been a growing interest in studying small proteins. These small proteins have been discovered and studied across various domains of life including prokaryote, archaea and eukaryote. However, many of them still poorly understood functions, lack defined secondary structures, and show limited similarities across different species (Weidenbach et al., 2022). Small open reading frames (smORFs) are gaining increasing attention due to their significant role in encoding small peptides. In the past, these peptides were considered non-functional or junk DNA by early gene prediction methods (Guerra-Almeida et al., 2021). However, they are now known to be involved in a wide range of physiological functions, including muscle formation, cell proliferation, immune activation, and more (Kute et al., 2022). Advancements in genomics and molecular biology have led to the discovery of numerous small open reading frames (smORFs) in various transcripts. Initially, many of these smORFs were considered non-functional; however, a significant number have since been identified as playing important roles in physiological functions and human diseases (Kute et al., 2022). The association between smORFs and human diseases emphasizes the importance of understanding their roles and functions. This knowledge can provide valuable insights into disease mechanisms and potential therapeutic targets (Kute et al., 2022).

Small open reading frames (smORFs) can encode proteins that act as key factors and play important roles in organisms. smORFs also constitute the potential pool for promoting *de novo* gene birth, leading to evolutionary progress and the development of various species. As a result, discovering these entities through theoretical and practical approaches has become a remarkable endeavor (Yu et al., 2023). smORFs have been identified as key player in a various of biological processes, including muscle formation and contraction, cell growth, and immune stimulation (Ji et al., 2020). A significant portion of these smORFs functions is still unknown. Hence, the development of computational techniques to determine the function of smORFs has become increasingly important (Ji et al., 2020). smORFs evolve continuously in species, serving as templates for protein production and potentially as building blocks for evolutionary adaptations (Sandmann et al., 2023). Comprehensive analyses have also categorized smORFs into different functional categories, ranging from inert DNA sequences to those that encode biologically active peptides (Couso & Patraquim, 2017).

smORFs have been found to play significant roles in various biological processes (Ladoukakis et al., 2011; Orr et al., 2020). These smORFs can be located in non-coding RNAs such as circular RNAs, mitochondrial RNAs, and long noncoding RNAs (lncRNAs), as well as in coding transcripts (5′UTR, CDS, and 3′UTR) (Orr et al., 2020). smORFs are present in all domains of life and are characterized by having start and stop codons within a span of 100 codons or fewer (Guerra-Almeida et al., 2021; Ladoukakis et al., 2011). These smORFs are highly abundant, surpassing the number of annotated protein-coding ORFs. While functional proteins containing fewer than 100 amino acids are known, the coding potential of smORFs has been largely overlooked in the past (Couso & Patraquim, 2017; Yu et al., 2023).

Therefore, despite their prevalence in genes with less than 100 codons (Cheng et al., 2011), smORFs have often been disregarded in gene prediction and annotation. However, recent studies have shed light on the significance of smORFs in various biological processes, including muscle formation, cell growth, and their potential roles in adaptation (Ji et al., 2020; Sandmann et al., 2023). Functional smORFs often remain unannotated due to the lack of experimental confirmation (Couso & Patraquim, 2017). there are frequently overlooked simply because they haven't been experimentally verified (Couso & Patraquim, 2017). Overcoming this challenge is difficult and rarely achieved by chance (Couso & Patraquim, 2017).

Computational annotation relies on identifying similar sequences, which may suggest the significance or function of the smORFs, or even match with known proteins, offering valuable insights into their roles (Couso & Patraquim, 2017; Hood et al., 2009; Kochetov, 2008; Samandi et al., 2017; Yu et al., 2023). The discovery of small peptides has drawn increasing attention, leading to expanded research in this field. Current studies on smORFs predominantly rely on computational prediction and biological experiments. Common experiments methos include ribosome profiling (Erhard et al., 2018; Fritsch et al., 2012), immunoblot assays (Hemm et al., 2008) and mass spectrometry (Kersten et al., 2011; Oyama et al., 2007). Despite the advances, biological investigation of small peptides or smORFs are hampered by their small size, short length, and low relative mass. These limitations often result in lengthy, costly experiments that may be ineffective and inaccurate.

However, rapid developments of algorithms have greatly benefited several fields, including lncRNA-disease association (Yu et al., 2020), cell-penetrating peptide identification (Wei et al., 2017), lncRNA identification (Meng et al., 2021), and

miRNA-lncRNA interaction (Kang et al., 2020). Machine learning algorithm also holds great potential as valuable tools for validating biological experiments, reducing costs and time, and advancing research.

Tuberculosis has recently been declared a global health emergency due to the rise in cases of Multidrug-resistant Tuberculosis (MDR-TB) worldwide. In a study by (Ejalonibu et al., 2021), researchers developed new and more effective antibiotics against resistant *M. tuberculosis* (Ejalonibu et al., 2021). These antibiotics are designed to inhibit essential bacterial proteins, offering a promising strategy for combating the global tuberculosis (TB) epidemic (Ejalonibu et al., 2021). The main objective of drug design and discovery is to identify compounds that can specifically target a protein's active site thereby disrupting its enzymatic activity. Once a compound with these properties is identified, it undergoes rigorous testing, including clinical trials, to evaluate its efficacy against the pathogen in the host (Ejalonibu et al., 2021).

In recent years, computational and analytical methods have emerged as valuable tools in drug development. These techniques offer a significant improvement over traditional methods, which can be time-consuming and laborious. Specifically, computational techniques have been enhanced with the use of advanced software that aids in the development and optimization of active compounds (Ejalonibu et al., 2021). These compounds have the potential to be used in future chemotherapeutic development to combat the global problem of tuberculosis resistance (Ejalonibu et al., 2021).

Machine learning (ML) refer to a set of algorithms that learn hidden patterns from datasets, enabling tasks such as classification and clustering (Zhu & Gribskov, 2019). The development of an effective ML-based method for a particular problem

depends on a good dataset and a good choice of a specific ML algorithm (Barbierato & Gatti, 2024). Many bioinformatics tools have been developed using machine learning (ML), such as the ability to predict ORF coding potential (Kang et al., 2017; Wang et al., 2013). Algorithm techniques also play a crucial role in genomics prediction, with ML methods widely applied in various genomic tasks, including protein coding potential identification (Yu et al., 2023), classification of disease-related genes (Le Thi et al., 2008), protein binding site detection (Pan & Yan, 2017), and disease diagnosis (Manogaran et al., 2018).

Despite the increasing use of machine learning and deep learning methods in genomics, accurate prediction remain a challenge (Oubounyt et al., 2019). Moreover, most current approaches rely on single shallow machine learning model, such as Support Vector Machines (SVM) and Logistic Regression. These classifiers, however, come with inherent limitations, leaving a room for further development and improvement in this field.

In this study, we present a hybrid algorithm of Random Forest and Gradient Boosting algorithms specifically for identifying small proteins or small open reading frames less than 100 codons. These smORFs, through crucial in many biological processes are often overlooked due to their size, lack of conservation and difficult in annotating smORFs in prokaryotic genomes. Enhancing the accuracy, sensitivity, specificity, and robustness of smORFs prediction is essential. An effective algorithm is needed for the computational discovery and characterization of smORFs with *Mycobacterium tuberculosis* as a model organism.

1.2 Problem Statement

Small proteins often exhibit diverse structures and functions that are not well characterized. Their short sequences pose a challenge for traditional algorithms, which rely on larger datasets to effectively identify patterns (Fuchs & Engelmann, 2023). As a result, these algorithms tend to have high false-negative rates, where actual small proteins are overlooked (Jeffery, 2023).

Previous studies have compiled and annotated numerous smORFs, but the functions of the majority of these smORFs remain unknown (Ji et al., 2020). Despite this, some smORFs that have been studied play crucial roles in various biological functions (Bartholomäus et al., 2021; Ji et al., 2020; Yu et al., 2023; Zhu & Gribskov, 2019). Identifying smORFs experimentally has proven challenging (Yu et al., 2023). While several *in silico* techniques have been developed to distinguish between long non-coding RNAs (lncRNAs) and coding RNAs (mRNAs), they are less effective with RNAs containing smORFs (Zhang et al., 2021). Furthermore, traditional methods are often computationally intensive and not scalable for large datasets and complex genes features (Yu et al., 2023).

This study aims to address these issues by introducing a more effective algorithm for the computational discovery and characterization of smORFs, using *Mycobacterium tuberculosis* (*M. tuberculosis*) as a model. *M. tuberculosis* causes tuberculosis, which is the ninth leading causes of death globally and the leading cause of mortality due from a single infectious agent, with the highest infection and death rates occurring in developing and low-income countries (Bagcchi, 2023).

1.3 Study Rationale

Recent advancements in genome sequencing and transcriptomics have generated vast amounts of data (Satam et al., 2023), yet the identification and functional characterization of smORFs remain underexplored. These smORFs, typically less than 100 amino acids in length, are challenging to detect with traditional bioinformatics tools due to their small size and the inherent noise in genomic data (Leong et al., 2022). Although previous studies have highlight the presence and potential significance of smORFs in various biological processes, but comprehensive datasets and robust prediction models are still lacking (Ji et al., 2020).

This study aims to address these gaps by focusing on the computational discovery and characterization of smORFs. Our goal is to contribute to the development and innovation of smORFs prediction in bacterial species. This research highlights the urgency of filing current knowledge gaps and focuses specifically on bacterial species and smORFs characterization.

1.4 Research Questions

1. How can an algorithm be developed to improve the efficiency of discovering and characterizing of small open reading frames (smORFs) or small proteins?
2. How can a computational framework be developed to identify small open reading frames (smORFs)?
3. How to validate transcriptome analysis in the discovery of novel smORFs?

1.5 Objectives of the Study

1.5.1 General Objective

The main objective of this study is to develop an algorithm for the computational discovery and characterization of small open reading frames (smORFs) or small proteins using *Mycobacterium tuberculosis* as a model.

1.5.2 Specific Objectives

1. To develop a hybrid algorithm for the prediction and identification of small open reading frames (smORFs).
2. To evaluate and compare the performance of the standalone Random Forest and Gradient Boosting models against the developed hybrid algorithm in predicting smORFs.
3. To validate the predicted smORFs from *Mycobacterium tuberculosis* by mapping them to its transcriptome data.
4. To identify and characterize novel smORFs in *M. tuberculosis*.

1.6 Conceptual Framework

The conceptual framework provides an approach to explore the diverse and functional significance of smORFs within bacterial genomes. By combining computational methods and transcriptome analysis to identify and characterize small proteins (Figure 1.1).

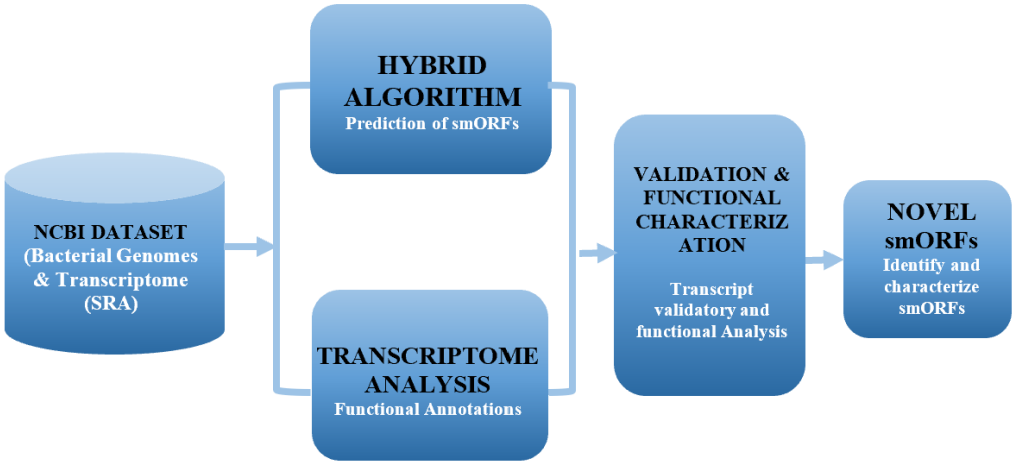

Figure 1.1 Diagram showing the conceptual framework

CHAPTER 2

LITERATURE REVIEW

2.1 Algorithm

Machine learning (ML) algorithms have evolved tremendously over the past two decades, transitioning from a laboratory curiosity to essential tools in widespread commercial applications. Within artificial intelligence (AI), machine learning has become the preferred approach for developing practical software for tasks such as computer vision, speech recognition, natural language processing and robot control (Jordan & Mitchell, 2015; Soori et al., 2023). ML encompasses a broad range of algorithms and modelling tools for various data processing tasks and has impacted most scientific disciplines in recent years (Carleo, 2019).

2.1.1 Hidden Markov Models

Hidden Markov Models (HMM) are used to account for unequal and unpredictable evolutionary rates across different positions in molecular sequences (Baldi et al., 1994). Evolutionary rates at distinct positions are constrained to a limited range of possible rates. The likelihood of phylogeny is calculated as a sum of terms, where each term is the probability of observing the data given a specific assignment of rates to sites, multiplied by the prior probability of that rate combination (Felsenstein & Churchill, 1996).

An HMM consists of a finite set of states linked by transitions. Each state is characterized by two sets of probabilities: a transition probabilities, which determine the likelihood of moving from one state to another, and output probability distribution or density functions (Yang et al., 1994). These distributions provide the probability of

emitting each output symbol from a finite alphabet or a continuous random vector of the current state (Yang et al., 1994).

2.1.2 Hybrid Algorithm

Hybrid algorithms are optimization algorithms that combine two or more different techniques to achieve better performance (Ting et al., 2015). This approach leverages the strengths of each technique while minimizing their weaknesses. Hybrid algorithms are important in enhancing algorithm search capability. The goal of hybridization is to integrate the benefits of each algorithm to develop a hybrid algorithm while minimizing any significant downsides. In general, hybridization results in increased computational speed or accuracy (Ting et al., 2015). In contrast, ensemble algorithms combine the predictions of multiple models often of the same or similar types (e.g., multiple decision trees or classifiers) to achieve better generalization performance. These models operate independently, and their outputs are aggregated using techniques such as majority voting, weighted averaging, or stacking. Popular ensemble strategies include bagging, boosting, and random forests (Brown, 2017).

2.1.3 Random Forest Algorithm

Random Forest (RF) is an algorithm proposed by Breiman (2001), RF builds multiple decision trees by randomly selecting subsets of data and features. This approach helps improve the model's robustness against noise, making it suitable for handling complex datasets (Figure 2.1). Numerous empirical studies have confirmed the high prediction accuracy of the random forest algorithm, as well as its ability to tolerate abnormal values and noise (Gao et al., 2019). RF has been utilized in various fields in recent years. For instance, Malek et al. (2018) successfully predicted paediatric

fracture healing time by combining Random Forest with self-organizing maps. Yu et al. (2023) introduced RF to predict protein-coding potential. Similarly, Wang et al. (2018) applied RF to condition monitoring and fault diagnosis in manufacturing and proposed a panoramic crack detection method based on structured RF. In the study of Tabatabaee Malazi & Davari (2018), they achieved a high level of accuracy in identifying complex activities of elders at home by using RF and emerging pattern algorithms, as measured by the F-measure index. In addition, in the study of de Santana et al. (2018), the authors quantified the quality of soil parameters through multivariable regression of RF, enabling a fast and automatic analysis process. Finally, Anitha & Raja (2017), proposed a new computer-aided method for detecting brain tumors using the RF classifier.

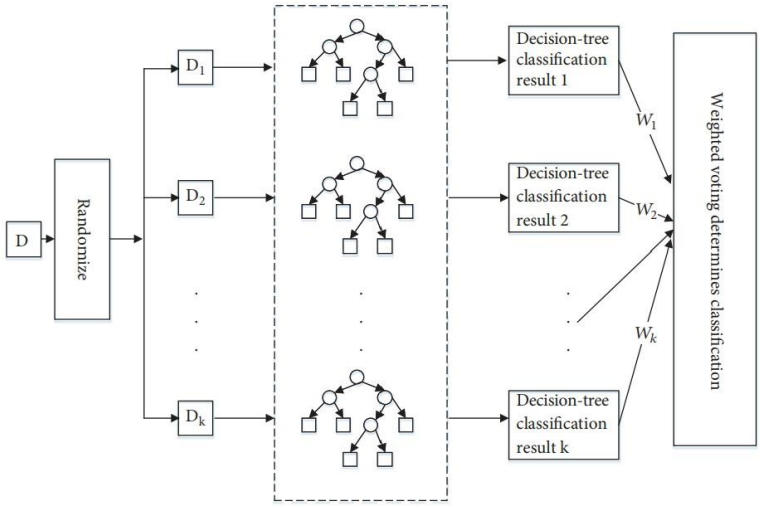

Figure 2.1 Structure of the Weighted Random Forest (Gao et al., 2019)

2.1.4 Gradient Boosting Algorithm

The gradient boosting machine (GBM) is an ensemble learning method that constructs a predictive model by sequentially adding fitted weak learners (Friedman, 2002). Gradient Boosting Tree (GBT) uses a boosting method to improve a decision

tree (DT) (Friedman, 2001). Rather than creating a new optimized model, the key idea is to combine weak models to form a strong consensus model. The feature space is initially partitioned into sub-regions in a DT (Hastie et al., 2001) to represent the dependent variable for each region (Breiman et al., 2017). The objective is to learn a functional mapping $y = F(x; \beta)$ from data $\{xi, yi\}$ n i=1, where β represents the set of parameters of F, with the objective of minimizing a cost function $\sum n$ i=1 $\Phi(yi, F(xi; \beta))$ (Chen et al., 2013). Boosting assumes that $F(x)$ follows an "additive" expansion form $F(x) = \sum M$ m=0 $\rho m f(x; \tau m)$, where f is referred to as the weak or base learner and has a weight ρ and a parameter set τ. Therefore, the whole parameter set β is composed of $\{\rho m, \tau m\}$ M m=1 (Chen et al., 2013). These parameters are learned in a greedy "stage-wise" process, which involves: (1) setting an initial estimator $f0(x)$; (2) for each iteration $m \in \{1, 2, \ldots, M\}$, solving $(\rho m, \tau m) = \arg\min_{\rho,\tau} \sum n$ i=1 $\Phi(yi, Fm-1(xi) + \rho f(xi; \tau))$ (Y. Chen et al., 2013).

GBM approximates step two steps. First, it fits $f(x; \tau m)$ by

$$\tau m = \arg min_T \sum_{1=1}^{n} (gim - f(xi; \tau))^2 \quad (1)$$

Where;

$$gim = -\left[\frac{\partial \phi(y_1, F(x_i))}{\partial F(x_i)}\right]_{F(x)=F_{m-1}(x)} \quad (2)$$

Secondly, it learns ρ by

$$\rho m = arg_{\rho}^{min} \sum_{i=1}^{n} \phi\left(y_i, F_{m_{-1}}(x_i) + \rho f + (x_1; \tau m)\right) (3)$$

Then, it modifies Fm $(x) = Fm-1$ $(x) + \rho m f(x; \tau m)$. However, in practical, shrinkage is frequently added to prevent overfitting; as a result, the update becomes

$Fm(x) = Fm-1(x) + \nu\rho mf(x; \tau m)$, where $0 < \nu \leq 1$. Tree factors, such as tree size (or depth) and the minimal number of samples at terminal nodes, influence the complexity of $f(x)$ if the regression tree is the weak learner. In addition to utilizing appropriate shrinkage and tree parameters, subsampling, that is fitting each base learner on a random subset of the training data could enhance GBM performance (Figure 2.2), this technique is known as stochastic gradient boosting.

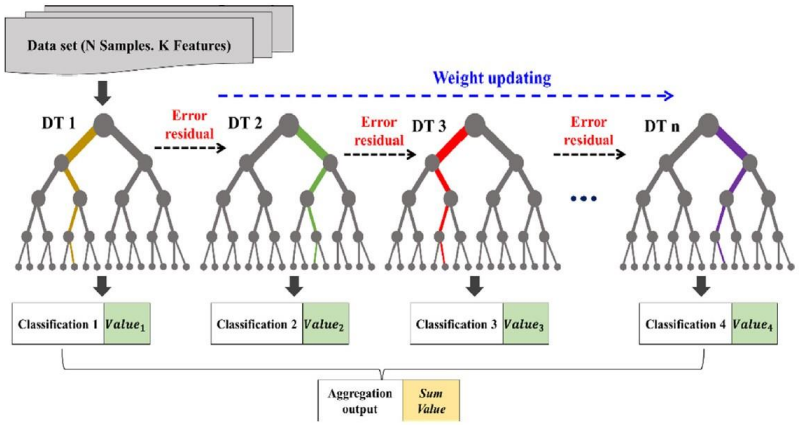

Figure 2.2 Structure of Gradient Boosting Tree (Chen et al., 2022)

2.2 smORFs prediction approach

Various algorithms have been employed to predict smORFs, including Random Forest (Yu et al., 2023), class-imbalance learning (Zhao et al., 2023), profile hidden Markov models and deep learning (Durrant & Bhatt, 2021), Convolutional Neural Networks (Al-Ajlan & El Allali, 2019), and DeepCPP, a deep neural network for coding potential prediction (Zhang et al., 2021). In addition, logistic regression was utilized in the MiPepid study (Zhu & Gribskov, 2019), while ORFpred (Srinivas et al., 2016) employed a machine learning method to predict the likelihood of ORF translation initiation and elongation. Other approaches include the smORFer algorithm

(Bartholomäus et al., 2021) and RNA expression analysis using reannotated microarray probes (Ji et al., 2020). Below, we will provide a detailed discussion on the distinct characteristics of algorithms employed by various smORF prediction tools. The purpose of this comprehensive analysis is to enhance understanding of the different approaches used in smORF prediction, while considering the tools. Specifically, we will examine both the advantages and limitations of each algorithm.

2.2.1 Prokaryotic Predictions

2.2.1(a) ProsmORF-pred: A machine learning based method for the identification of smallORFs in prokaryotic genomes

ProsmORF-pred used Random Forest approach for the prediction of small open reading frames (smORFs) encoding proteins with less than 100 amino acids (Khanduja et al., 2023). ProsmORF-pred, a machine learning-based technique created after a complete examination of known prokaryotic smORFs, was introduced in the research of (Khanduja et al., 2023). Based on sequence and genomic neighbourhood similarity searches, this technique shows potential in predicting smORFs and assisting in their functional annotation (Khanduja et al., 2023). The ProsmORF-pred methodology to identify smORFs entails training two separate ML models within ProsmORF-pred. The model for detecting protein-like sequences is trained using annotated smORFs from Escherichia coli, whereas the model for identifying initiation site recognition is trained using longer ORFs (>100aa) from the same genome. The work describes a detailed benchmarking of ProsmORF-pred, proving its performance on annotated smORFs from 32 bacterial genomes in comparison to previous techniques. The technique achieved sensitivity of 0.96 in predicting smORFs.

The discussion revolves around the significance of smORFs in cellular processes and the challenges associated with their computational identification. The authors elucidate the limitations of existing approaches, emphasizing the necessity for more accurate prediction tools, especially. They highlight the potential impact of ProsmORF-pred in complementing high-throughput experimental approaches and the importance of machine learning in advancing smORF prediction.

2.2.1(b) Automated Prediction and Annotation of Small Open Reading Frames in Microbial Genomes

smORFs and microproteins have been identified to play a significant role in microbes (Durrant & Bhatt, 2021). However, there are still numerous unknown smORFs in human-associated microbes. A recent bioinformatic analysis aimed to improve the prediction of small protein families by utilizing evolutionary conservation signals. To facilitate the annotation of specific smORFs, they developed a tool called SmORFinder (Durrant & Bhatt, 2021). This tool combined profile hidden Markov models of each smORF family with deep learning models that possess superior generalization capabilities for smORF families not included in the training set. Consequently, the predictions made by SmORFinder incorporate Ribo-seq translation signals. An analysis of feature importance revealed that the deep learning models can identify Shine-Dalgarno sequences, prioritize specific positions in each codon, and group synonymous codons present in the codon table. Furthermore, an examination of the core genome of 26 bacterial species identified several core smORFs with unknown functions. They have pre-computed smORF annotations for thousands of RefSeq isolate genomes and Human Microbiome Project metagenomes, which are accessible through

a public web portal. In this study, they developed two deep learning models using the hyperband algorithm to fine-tune hyperparameters. These models, collectively referred to as DeepSmORFNets (DSN), include the first model (DSN1), which was optimized to achieve the lowest validation loss on a validation set of observed smORF families ("Validation - Observed"), and the second model (DSN2), which was optimized to achieve the highest F1 score on a validation set of unobserved smORF families. Based on these findings, it can be concluded that the deep learning models generally exhibit better performance in generalizing to unobserved smORF families, while the pHMMs still demonstrated superior precision at a significance cutoff of an E-value < 1 x 10^{-6}. This suggests that the models may complement each other when used together to identify smORFs. The authors utilize a combination of profile hidden Markov models (pHMMs) and deep learning models to predict smORFs and their encoded microproteins (Durrant & Bhatt, 2021).

This approach, named SmORFinder, improves the detection of smORFs that are often missed by traditional methods. The deep learning models within SmORFinder identify biologically meaningful features of smORFs sequences, including Shine-Dalgarno sequences, codon synonyms, and codon wobble positions. Through rigorous evaluations and comparisons with existing tools, the authors demonstrate the effectiveness of their approach in identifying smORFs with improved precision and recall. Applying their approach to 26 bacterial species, the authors identified several core smORFs of unknown function (Durrant & Bhatt, 2021).

This discovery highlights the potential functional significance of these smORFs in microbial genomes. Moreover, the research provides a comprehensive analysis of thousands of RefSeq isolate genomes and Human Microbiome Project metagenomes, offering valuable annotations through a public web portal. This resource facilitates

further research and exploration of smORFs in various microbial genomes. The results of this research indicate that deep learning algorithms outperform in terms of generalizing to unobserved smORF groups in general, while the accuracy of pHMMs remains superior with a significance threshold of an E-value < 1 x 10^{-6}. The algorithms can potentially supplement each other when employed jointly for smORFs identification. The study introduces a novel approach to addressing the challenges of smORFs prediction and annotation in microbial genomes. Their deep learning models not only enhance detection accuracy but also uncover biologically relevant features of smORF sequences (Durrant & Bhatt, 2021). By identifying core smORFs of unknown function, this study provides insights into potential novel regulators of microbial processes. The SmORFinder annotation tool and the accompanying web portal offer valuable resources for the scientific community to explore smORFs' roles in microbial genomics. This research opens avenues for further investigations into the functional significance of these often-overlooked small proteins, additionally the study only focused on smORFs in microbial genomes and did not consider smORFs in other organisms (Durrant & Bhatt, 2021).

2.2.1(c) smORFer: a modular algorithm to detect small ORFs in prokaryotes

Small proteins are increasingly recognized as important in physiological processes (Bartholomäus et al., 2021). However, the functional identification and genome annotation of these proteins remain challenging. Ribosome profiling, a method that sequences ribosome-protected fragments, can detect active open-reading frames (ORFs) and annotate coding sequences (CDSs) (Bartholomäus et al., 2021). While multiple identifiers had been successful in eukaryotic smORFs annotation, they faced difficulties in prokaryotic genomes due to unique features such as polycistronic

messages and non-canonical initiation (Bartholomäus et al., 2021). To address this issue, a new algorithm called smORFer was developed (Bartholomäus et al., 2021). This algorithm aims to detect putative smORFs in prokaryotic organisms by using an integrated approach that considers the structural features of the genetic sequence and in-frame translation. This was achieved by converting these parameters into a measurable score using Fourier transform (Bartholomäus et al., 2021). The algorithm can be executed in a modular way, allowing different modules to be selected for smORFs search depending on the available data for a particular organism.

In the study conducted by Bartholomäus et al. (2021), presented a novel approach to identifying smORFs in prokaryotic genomes. SmORFs, which encode small proteins with fewer than 50 amino acids, have recently gained recognition for their central roles in various physiological processes (Bartholomäus et al., 2021). However, their systematic annotation and functional identification remain challenging both experimentally and computationally. The paper introduces a new algorithm, smORFer, which addressed these challenges by leveraging ribosome profiling data and considering unique features of prokaryotic genomes. The smORFer algorithm utilized ribosome profiling, also known as Ribo-Seq, which involves deep sequencing of ribosome-protected fragments to identify actively translated open-reading frames (ORFs). Unlike previous approaches that rely on the 3-nt periodicity in Ribo-Seq data sets for eukaryotic smORFs annotation, smORFer took into account the distinct characteristics of prokaryotic genomes. This includes factors such as overlapping ORFs, polycistronic messages, non-canonical initiation and leaderless translation, which can complicate smORFs prediction in prokaryotes (Bartholomäus et al., 2021).

One of the unique features of smORFer is its integrated approach, combining structural features of genetic sequences with in-frame translation (Bartholomäus et al., 2021). The

algorithm employed Fourier transform to convert these parameters into a measurable score, enabling accurate selection of putative smORFs. The modular nature of smORFer allows researchers to tailor the algorithm to specific organisms by selecting different modules based on available data. This flexibility enhances the algorithm's versatility and adaptability to various prokaryotic species.

The author compared RibORF with smORFer in detecting lengthy ORFs in the *E. coli* genome, specifically those longer than 1000 nt. Using the genomic sequence, RibORF and smORFer predicted many possible ORFs. The number was higher than the identified ORFs in *E. coli*, however, because they considered numerous start codons that shared the same stop codon. Both techniques detected 99.6% of the known annotated ORFs when counting ORFs based on unique stop codons. RibORF identified 235 translated ORFs (1.2% of all known ORFs >1000 nt) by including extra criteria, whereas smORFer recognized 740 (45% of all known ORFs >1000 nt) (Bartholomäus et al., 2021). This discrepancy might be attributable in part to the use of TIS-Ribo-Seq data, highlighting the relevance of include such datasets for precise mapping of initiation locations. It is worth noting that RibORF, which does not require TIS-Ribo-Seq, runs slower than the algorithm used in the study. The study contributes to the development of smORFer, a novel algorithm designed to accurately detect smORFs in prokaryotic genomes (Bartholomäus et al., 2021).

By accounting for the unique genomic architecture of prokaryotes, smORFer overcame challenges associated with traditional smORFs prediction methods. This algorithm holds promise for advancing our understanding of small protein function and expanding our knowledge of their roles in microbial physiology. In addition, the research paper introduces smORFer as an innovative algorithm for the detection of smORFs in prokaryotic genomes. By incorporating structural features of genetic

sequences and in-frame translation, smORFer demonstrates high accuracy in identifying putative smORFs. This modular approach provides researchers with a versatile tool to address the challenges of smORF annotation in prokaryotic organisms. The limitation shows that the use of a single dataset for peptide identification and the need for further validation of smORFer's predictions, including the need for more comprehensive datasets and the development of methods to study the function of small proteins (Bartholomäus et al., 2021).

2.2.1(d) Pervasive translation in *Mycobacterium tuberculosis*.

The study by Smith et al. (2022) investigates pervasive translation in *M. tuberculosis*. The authors employed advanced ribosome profiling techniques to map translation activity across the *M. tuberculosis* genome, addressing limitations of conventional gene prediction algorithms that often miss small and unconventional open reading frames (ORFs).

The researchers utilized two ribosome profiling approaches: Ribo-seq and Ribo-RET. Ribo-RET treats bacterial cultures with retapamulin to trap initiating ribosomes at start codons, enriching translation initiation sites. Ribo-seq captures elongating ribosomes across mRNAs. RNA fragments of about 31 nucleotides were size-selected, dephosphorylated, and prepared into sequencing libraries, which were sequenced to identify regions of active translation. Bioinformatic analyses included mapping ribosome footprints to the genome, detecting novel ORFs, particularly short ORFs, and analyzing their features. Evolutionary analyses examined codon usage patterns and signatures of purifying selection, providing insights into the potential functional importance of identified ORFs.

The study revealed widespread translation of numerous previously unannotated ORFs, most being short and encoding peptides of 50 amino acids or less. While many showed characteristics consistent with non-functional proteins, such as lack of conservation and signatures of neutral evolution, a subset demonstrated signs of evolutionary constraint. Approximately 90 ORFs exhibited evidence of purifying selection, suggesting functional relevance. The total number of these ORFs exceeds the annotated genes in the *M. tuberculosis* genome, highlighting pervasive translation. The findings imply that *M. tuberculosis* continuously translates a broad spectrum of genomic regions, generating short peptides that could serve as raw material for gene evolution or other roles in bacterial physiology.

The authors interpret their findings as evidence that pervasive translation in *M. tuberculosis* produces a large repertoire of short peptides, most likely non-functional. However, the subset under purifying selection suggests that some of these short ORFs may evolve in genes with specific functions. This dynamic indicates that pervasive translation might serve as a substrate for genetic innovation, allowing bacteria to adapt rapidly to environmental pressures. These insights into the bacterial translational landscape emphasize the complexity of microbial genomes and challenge traditional notions of gene annotation.

The study further provided a comprehensive analysis demonstrating that *M. tuberculosis* exhibits extensive translation of novel genomic regions, mainly short ORFs. Their integration of ribosome profiling with evolutionary evidence highlights the potential for new gene emergence and enhances our understanding of bacterial gene regulation and evolution. The findings underscores the importance of experimental approaches in revealing hidden layers of the bacterial transcriptome and proteome,

offering a foundation for future studies into the functional roles of these newly identified peptides (Smith et al., 2022).

2.2.1(e) Leaderless Transcripts and Small Proteins Are Common Features of Mycobacterial Translational Landscape

The study by Shell et al. (2015) provides a comprehensive investigation into the translational landscape of mycobacteria, specifically *Mycobacterium smegmatis* and *M. tuberculosis*. The authors employ an integrative approach combining RNA-seq, transcription start site (TSS) mapping, ribosome profiling, and N-terminal mass spectrometry to elucidate gene structures and translation mechanisms. They reveal that approximately 25% of transcripts are leaderless, initiated directly at an AUG or GUG start codon without upstream ribosome-binding sites, yet are translated efficiently, challenging the canonical view based on *E. coli* paradigms. The study uncovers numerous unannotated small proteins and alternative start codons, indicating significant underannotation and proteome diversity. Functional assays demonstrate that leaderless translation in mycobacteria solely depends on the presence of a start codon, whereas leadered translation requires Shine-Dalgarno interactions (Shell et al., 2015). However, these findings are limited to specific growth conditions, and the mechanisms of leaderless translation across different environmental states or stress conditions remain unclear. Moreover, while the study provides evidence for the prevalence and efficiency of leaderless translation, the precise molecular mechanisms behind this process are not fully characterized. These limitations suggest that additional studies are necessary to generalize the findings and fully understand how diverse conditions influence translation modes and regulation in mycobacteria. Despite these constraints, the work significantly advances our understanding of mycobacterial gene regulation and offers

important insights for genome annotation, pathogenicity, and therapeutic targeting of bacterial translation machinery (Shell et al., 2015).

Therefore, several approaches have been utilized for predicting smORFs using different algorithms, including Random Forest (Yu et al., 2023), class-imbalance learning (Zhao et al., 2023), profile hidden Markov models and deep learning (Durrant & Bhatt, 2021), Convolutional Neural Networks (Al-Ajlan & El Allali, 2019), and DeepCPP, a deep neural network for coding potential prediction (Zhang et al., 2021). Furthermore, in the study of MiPepid (Zhu & Gribskov, 2019), utilized logistic regression and ORFpred (Srinivas et al., 2016), used machine learning method to predict the likelihood of ORF translation initiation and elongation. Moreover, there are other approaches like the smORFer algorithm (Bartholomäus et al., 2021), and RNA expression analysis using reannotated microarray probes (Ji et al., 2020).

2.2.1(f) Integrated sequence and omic features reveal novel small proteome of *M. tuberculosis*

Sinha et al. (2024) developed a bioinformatics pipeline integrating sequence features with high-throughput omics data, RNA-Seq, Ribo-Seq, and proteomics to predict and validate small proteins in *M. tuberculosis*. Public datasets from various conditions, including exponential growth, nutrient starvation, and hypoxia, were used to capture diverse gene expression profiles. Features such as codon usage bias, GC content, and sequence conservation were used as inputs to train a Random Forest classifier to distinguish coding from non-coding regions. Incorporating Ribo-Seq data improved the detection of actively translated smORFs.