

**MINISTRY OF EDUCATION AND TRAINING
LAC HONG UNIVERSITY**



DUONG THI KIM CHI

**ENHANCING THE EFFICIENCY OF MACHINE
LEARNING MODELS FOR BIOMEDICAL DATA**

DOCTORAL THESIS IN COMPUTER SCIENCE

Major: Computer science

Major code: **9480101**

Dong Nai, 2023

The thesis was completed at: Lac Hong University

Instructor: Associate Professor, PhD. Tran Van Lang

Reviewer 1:

Reviewer 2:

Reviewer 3:

The thesis will be defended before the University-level Thesis Examining Council meeting

.....
.....

At hour....., day.....month.....year

The thesis can be found at the library:

- Lac Hong University Library
- National Library

TABLE OF CONTENTS

CHAPTER 1. INTRODUCTION.....	5
1.1 OVERVIEW OF RESEARCH TOPIC	5
1.1.1 Introduction	5
1.1.2 Purpose and research topic	5
1.1.3 Challenges of the research.....	6
1.2 OBJECTIVES, SCOPE AND RESEARCH METHODS	6
1.2.1 Objectives.....	7
1.2.2 The research scope:	7
1.3 THE OBJECTIVES OF THE THESIS.....	7
1.3.1 Designing effective machine learning models for molecular biology data in drug development using recombinant techniques	8
1.3.2 Effective Machine Learning Models for Molecular Biology Data in Species Identification Task.....	9
1.3.3 Effective machine learning models in biomedical applications of disease diagnosis based on clinical data.	10
1.4 CONTRIBUTIONS OF THE THESIS	10
1.5 THESIS STRUCTURE	11
CHAPTER 2. KNOWLEDGE BASE AND RELATED WORKS	12
2.1 CONCEPTS IN BIOMEDICINE.....	12
2.2 RELATED STUDIES THAT UTILIZE MACHINE LEARNING ALGORITHMS IN THE PROPOSED PROBLEMS	12
2.3 RELATED STUDIES USING MACHINE LEARNING ALGORITHMS IN THE PROPOSED PROBLEMS	12
2.3.1 Dimensional reduction.....	12
2.3.2 Unsupervised Learning Methods.....	12
2.3.3 Supervised Learning Methods	13
2.3.4 Ensemble Learning Method.....	13
2.3.5 The Algorithms to Address the Biomedical Data Challenges	13
2.4 RELATED STUDIES	13
2.5 BIOMEDICAL DATA USED IN THE DISSERTATION TASKS	14
2.6 EVALUATION OF MACHINE LEARNING MODELS	14
CHAPTER 3. HIGHLY EXPRESSED GENES DETECTION MODEL	16
3.1 THE PROBLEM OF FINDING HIGHLY EXPRESSED GENES (HEG)	16

3.1.1 Experimental Results	16
3.2 THE PROBLEM OF FINDING A SUITABLE EXPRESSION SYSTEM FOR THE TARGET GENE	17
CHAPTER 4. SPECIES IDENTIFICATION MODEL	18
4.1 INTRODUCTION	18
4.2 BUILD A DATA SET FOR THE TRAINING PROCESS	18
4.3 PROPOSED ALGORITHM TO BUILD A SPECIES IDENTIFICATION MODEL BASED ON ENSEMBLE LEARNING	18
4.4 MÔ HÌNH HỌC MÁY ĐỊNH DANH LOÀI NĂM MỖI	19
4.5 EXPERIMENTAL RESULTS	19
CHAPTER 5.....DISEASE DIAGNOSIS MODEL BASED ON CLINICAL DATA.....	22
5.1 DISEASE PREDICTION MODEL BASED ON CLINICAL DATA	22
5.1.2 Experimental results	22
5.2 CLASSIFICATION MODEL FOR CoViD-19 AND INFLUENZA	23
CHAPTER 6. CONCLUSION AND FUTURE DEVELOPMENT	27

Chapter 1. INTRODUCTION

1.1 Overview of research topic

1.1.1 Introduction

Biomedical computation (or bioinformatics) is an interdisciplinary research field between medicine and computer science. It combines methods of data analysis, machine learning, statistics, and information theory to address issues in the field of biomedical science such as disease detection and diagnosis, drug design, and molecular biology research. Important studies in the field of bioinformatics include disease outbreak prediction, drug development, vaccine production, and disease diagnosis and treatment. Machine learning methods can drive fundamental research and applications in bioinformatics [2]. These models can help determine disease risk, early detection of pathology, and prediction of treatment outcomes.

To build effective and highly applicable machine learning models, most machine learning algorithms require homogeneous structured data, digitized attributes, and no missing values. However, this is challenging to achieve with biomedical data because: i) the majority of molecular biological data has complex string sequence structures and a large volume of data due to the intricate biological mechanisms; ii) hospital data, such as medical records and medical test results for various diseases, are fragmented, duplicated, incomplete, and imbalanced in real-world biomedical datasets. These challenges serve as motivation for conducting this important research in the thesis.

1.1.2 Purpose and research topic

The last decade has seen great advances in Next-Generation Sequencing technologies, there has been a rise in the number of genomes sequenced and publication of biomedical data continue to grow through large-scale life science projects each year. That has been a strong motivation for researchers to utilize machine learning in the field of biomedicine. However, this research direction faces significant challenges due to the unique characteristics of biomedical data, including its complexity, high

dimensionality, and presence of errors. To overcome these challenges, effective preprocessing of raw biomedical data is crucial to enable machine learning algorithms to achieve optimal performance in tasks such as data description and classification predictions [1, 2].

The main objective of the thesis is to propose new approaches to the problem of "Efficient application of machine learning models to biomedical data through the selection of effective preprocessing solutions for two types of data: molecular biology data and clinical examination data in clustering and classification algorithms for biomedical applications such as predicting the origin of species, supporting vaccine/drug development, and aiding in disease diagnosis"

1.1.3 Challenges of the research

Applying machine learning models to biomedical data is highly necessary as it brings numerous opportunities for better disease treatment, optimized treatment costs, and reduces the burden on the healthcare system, especially during disease outbreaks. However, to make accurate predictions, most machine learning algorithms require homogeneous structured data, where attributes are digitized and devoid of missing values. However, this is challenging to achieve with biomedical data due to the following reasons:

- ❖ **The molecular biology data:** Gene sequence data has a high dimensionality (thousands of dimensions), complex biological mechanisms, and imbalanced data distribution. These are significant challenges in applying machine learning to biomedical problems, particularly in drug production.
- ❖ **The clinical data:** Patient's clinical data from hospitals, such as medical records and medical test results for various diseases, are fragmented, duplicated, incomplete, and imbalanced, especially for rare new diseases.

Therefore, preprocessing biomedical data to meet the requirements of machine learning algorithms is essential but challenging. Overcoming these challenges is crucial for the successful application of machine learning in the biomedical field.:

1.2 Objectives, scope and research methods

1.2.1 Objectives

The thesis aims to accomplish two primary objectives, which involve addressing two significant challenges associated with the implementation of automated machine learning models in two distinct groups of biomedical data. The specific tasks are as follows:

❖ For molecular biology data:

Designing efficient machine learning models for molecular biology data to address gene expression identification tasks, particularly in the field of drug development.

Constructing effective machine learning models for species identification tasks in molecular biology.

❖ For clinical and near-clinical data:

Developing automated preprocessing mechanisms for clinical diagnosis data in the field of biomedicine. This involves handling erroneous data, encoding data, and reducing the dimensionality of the attribute set.

Utilizing the preprocessed and standardized clinical data as input for training machine learning algorithms in various clinical biomedical applications, specifically for disease diagnosis purposes.

1.2.2 The research scope:

i) Designing methods for representation, encoding, and feature selection for gene data in the task of identifying target genes. This includes selecting the appropriate host cell environment for the target gene in drug design using recombinant techniques.

ii) Proposing a species identification algorithm based on Boosting techniques.

iii) Designing an automated method for encoding and dimensionality reduction of clinical data for the diagnosis diseases

1.3 The Objectives of the Thesis

The main objectives of the machine learning models for biomedical applications include the ability to explain the model's predictions and enhance the accuracy of the predictions. Considering the features of this type of data, the thesis approaches the objectives from the following

1.3.1 *Designing effective machine learning models for molecular biology data in drug development using recombinant techniques*

Molecular biology data, specifically gene sequence data, poses significant challenges for machine learning applications in drug development using recombinant techniques. These challenges include high-dimensional data (thousands of dimensions), complex biological mechanisms, and imbalanced data distribution. Addressing these challenges is crucial for machine learning applications in the pharmaceutical field. For example, in the process of drug production using recombinant technology, identifying gene sets with high protein expression or selecting a suitable host cell environment for target genes can improve the quality of recombinant proteins. Specifically, *finding an appropriate host cell environment for the target gene involves determining the level of codon response between the host cell environment and the desired recombinant protein product.* The main challenge of this task is to identify specially-genes that have the potential for optimal protein expression in genome. These specially-genes typically account for only 5% in genome containing thousands of genes. Additionally, finding a suitable host cell environment for the target gene is a critical step. In this context, the thesis proposes two effective solutions on this gene dataset: i) The first solution involves constructing a model for *"Predicting High Protein-Expressing Genes for Gene Design in Recombinant Techniques."* This model aims to identify specially-genes that have a high potential for protein expression, which can be used in gene design for recombinant applications; ii) The second solution is to build a *"Gene Correlation Prediction Model with the Host System in Recombinant Techniques."* This model focuses on predicting the correlation between

genes and the host cell system used in recombinant techniques. The process of solving the first task is described in Figure 1.1.

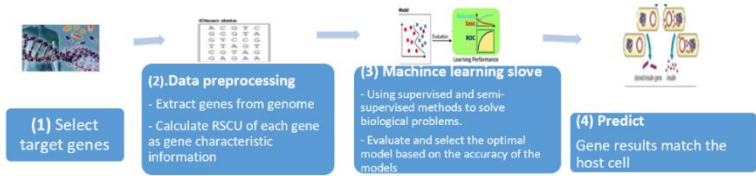
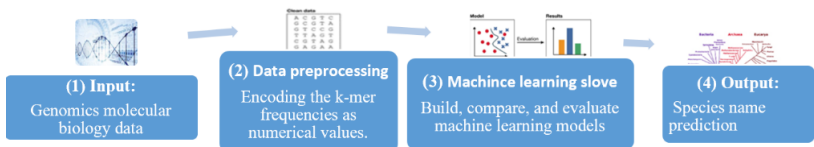


Figure 1.1 General process of the first task problem

1.3.2 Effective Machine Learning Models for Molecular Biology Data in Species Identification Task

In drug development applications that utilize gene sequence data (genomics), there are several tasks, including species identification, disease mechanism analysis, and detection of anomalies in gene sequences. Species identification involves determining the genetic characteristics of a species to gain a better understanding of its origin. This information can provide insights into the survival and development mechanisms of the considered species. Traditionally, this task relied on biological observations and external phenotypic features to determine species names. Alternatively, sequence comparison algorithms such as Neighbor Joining (NJ), distance-based methods, and clustering methods were used [10, 11]. These algorithms are effective in comparing individual sequences but lack efficiency when applied to multiple gene pools, especially when the gene pools contain genes of varying lengths and differences. By employing machine learning techniques, it is possible to simultaneously identify thousands of genes belonging to a particular species and provide efficient genetic information about these genes in terms of time and explanatory power. This general approach is illustrated in Figure 1.2.

Figure 1.2. General process of species identification problem.



1.3.3 Effective machine learning models in biomedical applications of disease diagnosis based on clinical data.

Biomedical data includes paraclinical and clinical data. This is biomedical data collected from screening test results at medical facilities. This data has height characteristics, The data often contains errors, missing data, and serious imbalance for the rare disease class. To solve the challenges of this task, the thesis proposes a general method as shown in Figure 1.3.

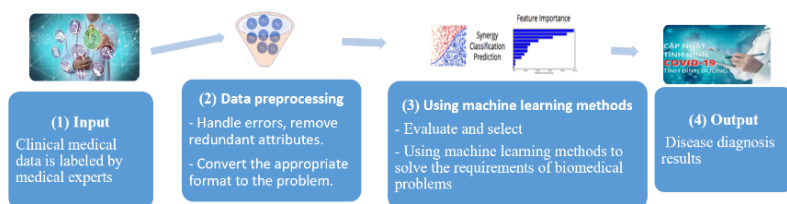


Figure 1.3. General process of disease diagnosis problem from clinical data.

1.4 Contributions of the Thesis

To accomplish the objectives of the stated problems, theoretical research and practical experiments have been conducted in the dissertation to propose solutions to the research tasks. The main contributions of the dissertation include:

Firstly, the dissertation addresses the issue of dimensionality reduction for machine learning models in molecular biology applications, specifically for the problem of selecting suitable genes in recombinant protein production. By using the HSCU index calculation method to represent the characteristic information of each gene, the dissertation compresses the input dataset for two machine learning tasks: selecting highly expressed genes in the expression system and selecting appropriate host cells for the target gene in molecular biology applications for recombinant protein production [CT1, CT2, CT3].

Secondly, the dissertation tackles the issue of high-dimensional information representation for machine learning models in biological species identification applications. By utilizing the k-mer method to represent the characteristic information of each gene, machine learning

algorithms, ensemble techniques, and hierarchical clustering are employed to create a species identification model. The dissertation constructs a model to predict the names of 17 termite mushroom species. The model achieves an accuracy of 91% AUCROC of 99% [CT4, CT7].

Thirdly, the dissertation utilizes a combination of Boosting algorithms to build an automatic feature reduction model for clinical data in the disease diagnosis task. Specifically, the dissertation constructs a model to predict patients with CoViD-19 based on this clinical dataset. The dissertation employs a combination of LightGBM and XGBoost techniques to automatically reduce features from the clinical dataset regarding CoViD-19 and Influenza. Additionally, a disease prediction model is built with an accuracy of 99.85% [CT5, CT6].

1.5 Thesis Structure

Chapter 1: Presents the necessity of the thesis, determines the objectives, objects and scope of the research, the thesis determines the tasks and approach of the thesis. The thesis presents the contributions of the thesis and summarizes the structure of the thesis.

Chapter 2: Presents basic issues related to biomedical data. The thesis presents an overview of each requirement of the three proposed problems, approaches and solutions for each problem.

Chapter 3: Building and selecting a set of highly expressed genes in the expression system using the PAM and CLARA clustering algorithms. Select host cells that match the target gene using Random Forest and XGBoost algorithms.

Chapter 4: Machine learning model for species identification, proposed model for species identification, issues of species identification as well as data processing process and model for predicting species names.

Chapter 5: Disease prediction model from clinical data. Approach to disease diagnosis from CoVid-19 disease data, using machine learning to distinguish CoVid-19 disease and seasonal flu.

Chapter 6: Presents conclusions summarizing the contributions of the thesis as well as proposing future research directions

Chapter 2. **KNOWLEDGE BASE AND RELATED WORKS**

2.1 Concepts in biomedicine

Outline the fundamental concepts of biomedicine used in the dissertation problems, such as genomics, DNA recombination, synonymous codons, gene expression systems, and species identification.

2.2 Related Studies that Utilize Machine Learning Algorithms in the Proposed Problems

2.2.1 Dimensionality Reduction Issues:

The data characteristics in biological problems often involve high-dimensional data. Therefore, it is necessary to reduce the size of the dataset during the preprocessing step before conducting further analysis. There are two main approaches to achieve this: dimensionality reduction techniques and feature selection. In the proposed dissertation, the focus is on feature selection: (i) The techniques used for sequence data in the dissertation involve gene encoding using synonymous codon ratios for each gene and k-mer technique for feature selection. (ii) For clinical data, multiple methods are employed to automatically remove attributes containing missing data, correlated data, and low information gain data.

2.3 Related studies using machine learning algorithms in the proposed problems

2.3.1 *Dimensional reduction*

The data characteristics in biological problems are often high-dimensional, so it is necessary to reduce the dimensionality of the dataset during the preprocessing step before conducting subsequent analyses. There are two main approaches to achieve this: dimensionality reduction techniques and feature selection. In the proposed dissertation, the focus is on feature selection: (i) The techniques used for sequence data in the dissertation involve gene encoding using synonymous codon ratios for each gene and the k-mer technique for feature selection. (ii) For clinical data, a combination of methods is used to automatically remove attributes containing missing data, correlated data, and low information gain data

2.3.2 *Unsupervised Learning Methods*

Clustering is a characteristic technique of this method, which relies on similarity measurements and the definition of these measurements. Clustering techniques are highly effective in extracting information from datasets. The dissertation applies clustering techniques to predict gene sets for high expression and utilizes hierarchical clustering to classify species

2.3.3 Supervised Learning Methods

Classification is the representative method of supervised learning. Supervised learning is defined as the estimation of the functional relationship between *the attributes* x_i and the target variable y_i , $y_i \approx f(x_i)$, where y_i can be either discrete (classification) or continuous (regression) depending on the nature of the variable. In the proposed dissertation, this technique is used in all three tasks: the first task, the second task, and the third task

2.3.4 Ensemble Learning Method

Ensemble learning is a machine learning approach that trains multiple models simultaneously to solve the same problem, and then combines their outputs to improve accuracy. Ensemble learning methods, particularly Boosting techniques, have been widely used in various fields, including biology. Boosting can be used to classify patient data or predict the response of a specific patient to a certain medication.

2.3.5 The Algorithms to Address the Biomedical Data Challenges

The machine learning methods mentioned above, namely PAM, CLARA, Random Forest, Xgboost, and Catboost, are all utilized in the thesis to tackle the outlined issues in biomedical data.

2.4 Related studies

There have been many successes in drug discovery supported by machine learning methods. For example, Deep Genomics has utilized an AI workspace platform to generate new genetic targets and corresponding oligonucleotide drug candidates, such as DG12P1, for managing a rare genetic disorder called Wilson's disease [25] [26].

Machine learning is also employed in the identification of species and control of gene sequence variations, providing strong support for vaccine and drug development, such as identifying the species of the CoViD-19 virus and its variants [28].

Furthermore, machine learning aids in obtaining proactive preventive and therapeutic agents to control potential future disease outbreaks. In a study conducted by Norah Alballa, 93 published dissertations from January 2020 to January 2021 were surveyed across reputable journals such as PubMed, Scopus, IEEE Xplore, and Google Scholar. The study focused on the application of machine learning for diagnosing CoViD-19 and predicting hospitalization risks. The author summarized the algorithm groups applied, as described in Figure 2.6

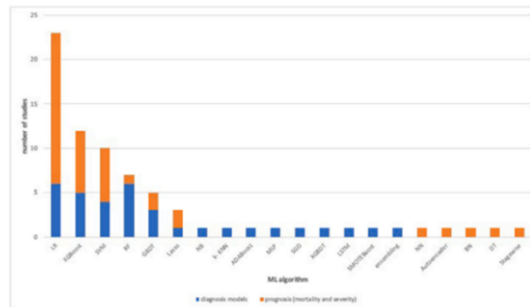


Figure 2.6: Studies have been Utilizing ML Algorithms for CoViD-19 Diagnosis Based on Clinical Data [29].

2.5 Biomedical Data Used in the Dissertation Tasks

This is also the first task of the dissertation, involving genomic data with over 10,000 sequences belonging to three groups of host organisms. In the second task, sequence data is still used for species identification, with over 2,000 ITS sequences belonging to 29 termite fungal species. The final task involves clinical data collected with a total of over 7,000 samples from patients with CoViD-19 and H1N1 influenza.

2.6 Evaluation of Machine Learning Models

To evaluate the machine learning models in the thesis, the classification results of the proposed model are compared with other models. Specifically, the dissertation utilized the k-fold cross-validation (CV) method, as well as metrics such as Accuracy, Sensitivity (also known as Recall), Specificity, False Positive Rate (FPR) or Fall-out, F1 score, and Receiver Operating Characteristics (ROC) curve

Chapter 3. **HIGHLY EXPRESSED GENES DETECTION MODEL**

3.1 **The problem of finding Highly Expressed Genes (HEG)**

Highly Expressed Genes (HEG) are genes that have the ability to produce a large amount of protein with specific characteristics in a cell or specific cell system [5]. These genes are favored in the production of recombinant proteins because they can increase protein production yield and reduce production costs. Finding HEG within a gene system is a challenging task as they are often expressed at higher levels compared to other genes under the same environmental and experimental conditions. The thesis utilized a partitioning method to cluster genes within a gene system in order to select groups of genes that tend to use similar codons within each group. This was done to identify HEG that are closest to the center of each cluster. The thesis proposed the use of cluster analysis methods such as PAM [36] and CLARA [38] for gene clustering, and the general approach can be summarized as follows:

- (1) Encoding genes using the RSCU technique for the entire gene set in the gene system.
- (2) Applying clustering algorithms PAM and CLARA to the data obtained in step (1) to form clusters and find new cluster centroids.
- (3) Evaluating a clustered group. Selecting the gene group closest to the cluster centroid, which is highly likely to be a HEG.

3.1.1 ***Experimental Results***

1. **Execution time of PAM-CLARA on the Bsubtilis dataset**

The execution time criterion is highly important when selecting algorithms for computational tasks involving large datasets. The execution time of PAM and CLARA on the B. subtilis gene set was measured and compared.

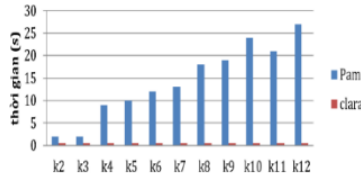


Figure 3.1. Execution time of PAM and CLARA operations on k-clusters

2. **Silhouette index:** used to evaluate clustering quality using clustering quality as shown in Figure 3.2 and Figure 3.3.

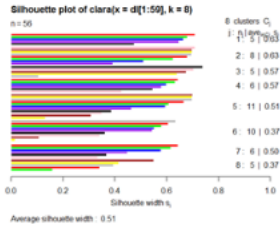


Figure 3.2. Silhouette index in CLARA

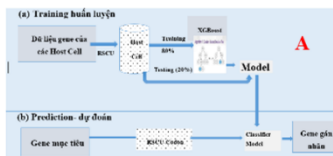
```
Objective function: 83.84559
Clustering vector: 1m (1:3543) 1 2 3 4 5 6 7 8 9 10 11 9 9 32 .
Cluster sizes: 86 53 40 72 36 66 75 33 45 20 17 53 51 94 75 27 42 76
21 15 25 33 44 48 20 72 70 76 56 46 38 110 14 46 50 74 400 89 26 39 72 27 59 60
99 70 29 34 48 44 11 30 26 43 56 36 13 40 40 10 50 21 26 20 80 46 57 50 5 35 35
32 76 30 5
Best sample:
[1] 17 45 51 79 82 84 89 138 152 191 219 242 259 1777
[15] 285 285 287 300 306 329 338 388 397 450 481 473 529 340
[29] 563 606 633 638 647 649 676 685 732 744 751 769 781 786
[43] 797 803 837 866 883 897 924 955 1004 1020 1038 1033 1065 1072
[57] 1085 1089 1139 1153 1159 1172 1174 1195 1196 1205 1206 1230 1235 1244
[71] 1244 1279 1282 1301 1306 1320 1321 1322 1328 1355 1379 1417 1451 1469
[85] 1482 1485 1490 1502 1518 1550 1561 1621 1640 1651 1669 1675 1681 1687
[99] 1739 1774 1786 1790 1820 1829 1832 1838 1841 1937 1944 1959 1975 1983
[113] 2003 2007 2098 2137 2144 2147 2148 2150 2170 2233 2260 2300 2326 2347
[127] 2352 2360 2366 2428 2438 2454 2455 2462 2510 2532 2533 2567 2582 2599
[141] 2625 2626 2645 2669 2692 2710 2717 2762 2765 2784 2785 2836 2838 2898
[155] 2906 2916 2944 2951 2996 2997 3006 3020 3023 3088 3102 3109 3163 3164
[169] 3211 3214 3234 3285 3347 3357 3376 3378 3385 3400 3425 3458 3461 3490
[183] 3492 3497 3521 3524 3528 3536 3540 3543
```

Figure 3.3. List of names of 5% of selected

HEG genes

3.2 The Problem of Finding a Suitable Expression System for the Target Gene

To improve the productivity of recombinant protein expression in recombinant protein production, it is crucial to find a suitable Host Cell for the target gene. This was also a project classification for which the thesis proposes a machine learning method to address. The general procedure for predicting gene correlation with host cells is depicted in Figure 3.4.A. The host cells experimental dataset used includes E. coli, the bacterial strain *Bacillus subtilis*, and *Latococcus lactis*. Experimental results of the process are presented in Figure 3.4B



Host cell	B	
	Random Forest	Xgboost
Host Ecoli (Class 1)	0,97	0,99
Host 2 B.Subtilis (Class 2)	0,94	0,99
Host 3 Latococcus (Class 3)	0,88	0,97

Figure 3.4A The overall process of the proposal and Figure 3.4B results of the process.

Chapter 4. SPECIES IDENTIFICATION MODEL

4.1 Introduction

Understanding the origin of a biological species is a highly important task in both the field of biology and medicine. It facilitates subsequent activities such as species conservation, discovery of new species, and in the field of medical science, it supports diagnosis and treatment. A DNA segment commonly used for classification is referred to as a DNA barcode, which corresponds to the ITS (Internal Transcribed Spacer) sequence. These gene groups typically utilize rRNA genes, including 18S, 5S, and 16S genes, to evaluate the evolutionary relationships between organisms. To address this issue, this study proposes a novel method that utilizes ensemble learning techniques based on ITS sequence data to classify species

4.2 Build a data set for the training process

The ITS sequence data used for the proposed classification model is the ITS data of termite mushrooms. The majority of this data in gene banks is incomplete, and the naming conventions are not standardized, making information retrieval difficult. The thesis synthesized ITS sequence data from two gene banks BOLD and NCBI. By excluding termite mushroom species with fewer than 7 sequences, the thesis obtained 1704 sequences from 17 termite mushroom species. The CountVectorizer encoding method was applied to construct the dataset for the termite mushroom species classification model. The data processing in the thesis utilized k-mer techniques and applied the CountVectorizer encoding method to build the dataset for the termite mushroom species classification model. The proposed ITS information encoding technique is illustrated in Figure 4.1.

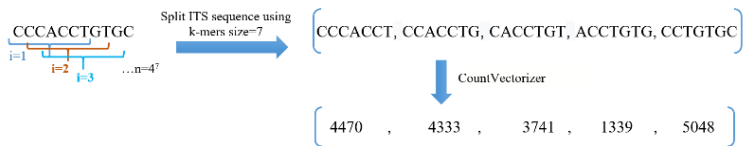


Figure 4.1 Illustration of the ITS encryption process with k-mer of length $k=7$.

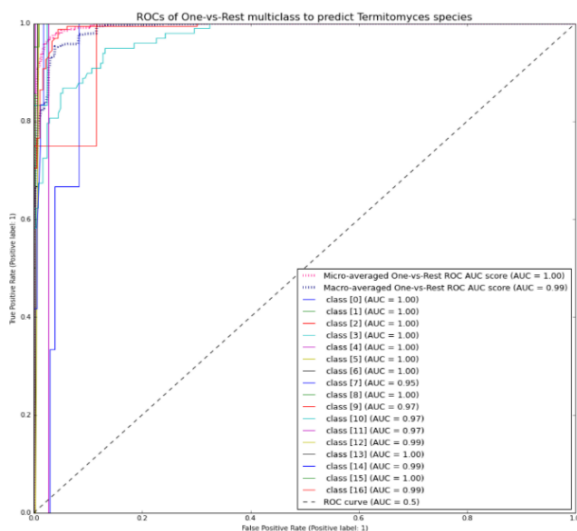
4.3 Proposed algorithm to build a species identification model based on ensemble learning

The thesis conducted experiments on the same data set with four different algorithms to compare the performance of characteristic measures of the machine learning model. Detailed results are presented in Table 4.1

Table 4.1. Performance comparison of machine learning algorithms on the same dataset

Method	AUROC	Accuracy	Precision	Recall	F1 score
Naive Bayes	0.93	0.60	0.84	0.60	0.62
RandomForest	0.98	0.88	0.88	0.88	0.88
XGBboost	0.99	0.91	0.90	0.91	0.90
Catboost	0.99	0.87	0.87	0.87	0.87

In addition, the AUCROC of each class is calculated according to the macro-average ROC OvR method for the multiclass model used [64]. The thesis uses the last class (grade 16) as the positive class, the remaining classes are negative. And the visual results of each layer are presented in the figure 4.1



Hình 4.1 The ROC curve using the OvR macro-average for each class in the XGBoost method by size k-mer=7

❖ Performance comparison with fungal classifiers

The proposed results of the thesis have higher performance than other classification models when applying k-mer size of size 7 with the selected classification algorithm XGBoost. The comparison results are presented in table 4.2

Table 4.2: Performance comparison with different fungal classifiers also using ITS

Ref	Method	Accuracy
(Schloss PD 2009)	K-mer (k=5), kNN, PGMA	0.86
(Delgado 2016),	K-mer (k=5), Naïve Bayes	0.87
(Deshpande 2016)	K-mer (k=8) Bayesin	0.87
(Prabina Kumar 2019)	K-mer (k=4), Random Forest.	0.89
Thesis proposal	K-mer (k=7), XGBoost	0.91

❖ **Comparing the prediction results of the proposed model with the prediction results of BLAST:**

The ITS sequences of termite mushrooms collected in Binh Duong province, Vietnam, have been published on NCBI. These sequences were predicted using the proposed classification model from the thesis, and the results were consistent with the species identification results stored on NCBI. Additionally, the sequences belonging to the termite mushroom genus MF163445-BD3, MF163446-BD6, MF163149-BD4 were obtained, but their species were not clear. The classification system of the thesis successfully identified two sequences, MF163445-BD3 and MF163446-BD6, as belonging to the species *Termitomyces striatus*, which matched the reference specimen collected during sequencing.

Chapter 5. DISEASE DIAGNOSIS MODEL BASED ON CLINICAL DATA.

5.1 Disease prediction model based on clinical data

5.1.1 Introduction to the disease prediction problem and proposed model

Data from blood test results is a type of paraclinical data that is often used in initial disease diagnosis. This is also the type of data that is widely stored in medical records. Disease diagnosis with automatic support using machine learning based on blood test results can detect blood diseases, cardiovascular diseases, and liver diseases. The thesis uses blood test results of patients with Covid-19 provided by Israelita Albert Einstein hospital in Brazil as training data [32].

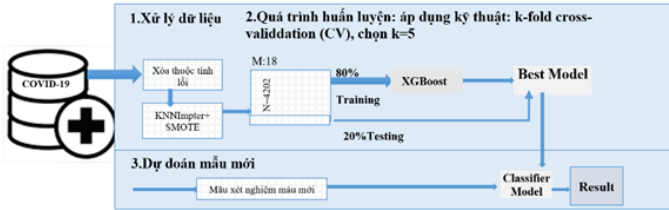


Figure 5.2. Overview model proposed for diagnosing Covid-19 disease

To get the best data for the classification model, the thesis applied the method of removing attributes with a null error rate of more than 99.9% and removing attributes with high relevance in the set. data. To drop all attributes of input data that have an error rate greater than 99.9 %. The thesis has proposed to remove attributes with high error rates using Algorithm 5.1

Algorithm 5.1: Algorithm to remove attributes with high missing

Input: set $D_{ogr} = X_{ogr} = \{(x_1, y_1), (x_2, y_2), \dots, (x_n, y_n)\}, x \in R^m, y \in \{0, +1\}$ m -dimensional, $Y \in [0, 1], thresh$

Out put: D_{lean}

Begin

- 1: missing_total $\leftarrow$ Calculate null values in columns
- 2: missing_percent $\leftarrow$ missing_total / len(D_{ogr})
- 3: data_missing $\leftarrow D_{ogr}.index[missing_percent \geq thresh]$
- 4: $Clean \leftarrow Dog.drop[data_missing]$

End

5.1.2 Experimental results

The thesis uses many different algorithms to build a classifier based on Datasmall such as: XGBoost [22], LGBMClassifier [47], CatBoost [48]. The AUROC evaluation results of each algorithm are described in Figure 4. Realizing that XGBoost gives the best AUROC value, it should be used to build a classifier .

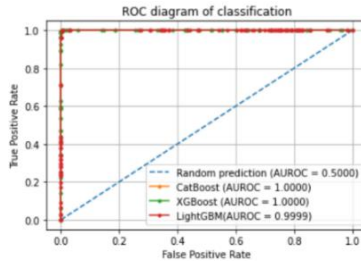


Figure 5.2. Comparison of ROC performance of gradient boosting algorithms

The results of the proposed model are also compared with the results of predicting the same function as researched by author Maryam AlJame [25].

Table 5.1 Performance comparison table of the proposed model

Classifier	Algorithm	AUC	Precision	Recall	F1 score
Proposed model	XGBoost	0.998	1.0	1,0	1,0
Maryam AlJame [25]	XGBoost	0.994	0.98	0.99	-

Realizing that the proposed model uses input data processing methods such as removing attributes with errors of more than 99%, deleting empty data using the KNNimputer technique, and reducing imbalance between two classes using the SMOTE technique. This combination helps the proposed classification model have better prediction results than research with the same goal and data set.

5.2 Classification Model for CoViD-19 and Influenza

5.2.1 Problem Introduction and Solution Model

The data collected from hospitals or medical records always contain numerous missing or erroneous attributes. If methods that involve removing errors are used, it may affect the prediction results by disregarding data containing valuable attributes used in diagnosis and treatment. The thesis

proposes a model for automatically processing erroneous data and distinguishing between two diseases with similar symptoms: CoViD-19 and H1N1 influenza.

The proposed model can directly learn from raw data without the need for intervention to remove empty data. The proposed method consists of two stages: In the first stage, it evaluates and processes the data to reduce non-predictive attributes in the dataset using the Light Gradient Boosting algorithm (LightGBM). In the second stage, it constructs a classification model for each disease group based on the XGBoost method. The proposed model is illustrated in Figure 5.3.

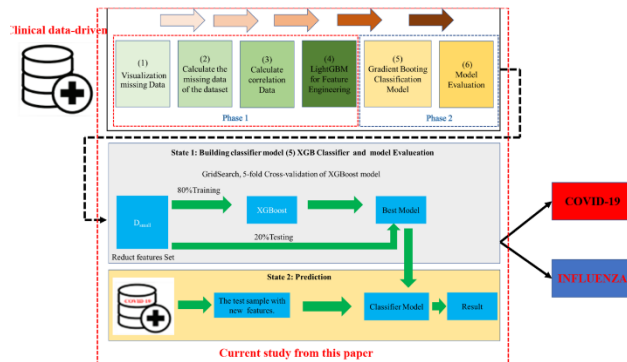


Figure 5.2. Overview model distinguishes Covid -19 and Influenza

The thesis utilized LightGBM to select important attributes in the clinical dataset from the CoViD-19 and influenza platforms. Subsequently, the tree model used the Information Gain (IG) method to calculate values for the important attributes. The thesis employed 5.2 Algorithm to extract the

important attributes, and Algorithm 5.3 provided detailed steps for constructing the disease classification model.

Thuật toán 5.2. Trích chọn các thuộc tính

Đầu vào: D_{learn} ; iter = number of iterations, hyperparameters = {objective = 'binary', boosting_type = 'goss', n_estimators = 1000, class_weight = 'imbalanced'}

Đầu ra: D_{small} : The dataset

Begin

- 1: Initialize: FeatureImportances = {}
- 2: Model $\leftarrow$ LGBM classifier (hyperparameters)
- 3: **Repeat** iter times
 - Divide D_{learn} dataset into D_{Train} and D_{Test}
 - Train model based on D_{Train} and early-ending hyperparameters
 - FeatureImportances = model.FeatureImportances

End repeat

- 4: Tính trung bình FeatureImportances
- 5: Chọn ZeroFeatureImportances //Chọn các thuộc tính có giá trị là không
- 6: Return $D_{small} \leftarrow D_{learn} - \text{ZeroFeatureImportances}$

End

Thuật toán 5.3: Xây dựng bộ phân loại bệnh dùng kỹ thuật XGboost

Đầu vào: $D_{small} = \{(x_i, y_i) \in R^m \times \{0, 1\}, \forall i = \overline{1, m}\}$; hyperparameter is $\Theta = \{\text{'n_estimators'}$, $\text{'colsample_bytree'}$, 'learning_rate' , 'eval_metric' : 'auc'}

Đầu ra: Best_Model

Begin

- 1: Initialize: FeatureImportances = {}
- 2: Model $\leftarrow$ **XGBoostClassifier** (Θ)
- 3: KFold $\leftarrow$ StratifiedKFold (n_splits=5, shuffle=True, random_state=2020)
- 4: For i=1 to KFold
 - Divide the D_{small} dataset into D_{Train} and D_{Test}
 - Train the model based on D_{Train} and early-ending hyperparameters
- 5: Tính các điểm roc_auc_score, accuracy_score, precision_score, recall_score, and f1_score
- 6: Chọn the best_model ở bước 4
- 7: Visualize the mean value ở bước 4
- 8: Return the Best_Model ở bước 4

End

5.2.1 Performance of the Proposed Model

❖ Comparing the accuracy of the prediction data

From the input dataset consisting of 50 attributes, the thesis extracted 24 important attributes for disease diagnosis. Among them, the GroundGlassOpacity attribute plays a particularly important role in diagnosing CoViD-19, as it appears in the majority of hospital admission

cases with a rate of 89%. The inclusion of this feature in the proposed model for predicting CoViD-19 and influenza is highly significant in uncovering the underlying data patterns.

❖ Performance of the Proposed Model

To assess the accuracy of the thesis's proposed model, it compared its performance with Li's model. The specific results are described in Table 5.6.

Bảng 5.6: . So sánh hiệu suất của mô hình đề xuất với mô hình của Li

Bộ phân loại		Accuracy (%)	AUC score (%)	Recall score (%)	Precision score (%)	F1 score (%)
Mô hình đề xuất của luận án	XGBoost	99.9	99	99	100	99
	Gradient-Boosting	99.6	99	98	100	99
	LGBM	99.7	99	99	100	99
	Random Forest	99.8	99	99	100	99
Li [77]	XGBoost	99	97.7	92	92	92
	RIDGE regression	-	96.6	-	-	-
	Random Forest	-	95.3	-	-	-
	LASSO regression	-	96.3	-	-	-

Chapter 6. **CONCLUSION AND FUTURE DEVELOPMENT**

6.1 Conclusion:

Machine learning models can effectively handle prediction, classification, and visualization of biological and medical data. The thesis proposed efficient solutions to support computational tasks in urgent biological issues such as drug production in recombination, species identification, and medical diagnosis. In the proposed approaches, the thesis focused on accomplishing three main tasks, addressing two major challenges in deploying machine learning models for two types of biomedical data: molecular biology data and clinical data. The thesis achieved the following research objectives:

Designing an efficient machine learning model for molecular biology data in gene expression determination, applied in the field of drug development.

Constructing an effective machine learning model for species identification by extracting gene information using the K-mer method. This method converts raw gene sequence data into numerical information and utilizes it in the machine learning model, facilitating accurate species classification and convenient data visualization.

Developing an efficient machine learning model for disease diagnosis in the biomedical field by handling missing data and selecting important information from clinical diagnostic data. This was achieved by combining two LightGBM and XGBoost. This proposal automated the selection of important attributes before inputting them into the disease diagnosis model, ensuring the avoidance of information loss while maintaining the best accuracy for the prediction model.

6.2 Future Development:

The application of bioinformatics holds vast potential for studying computational solutions that support biological applications in human healthcare. Continuing to improve the accuracy of prediction models in Problem 1 and Problem 2 is a key concern. Deep learning solutions can be explored to automatically extract gene information in Problem 2, replacing

the K-mer technique. While the K-mer method may be suitable for fungal species identification, it may not be suitable for species with longer gene sequences, making automatic feature extraction necessary. For Problem 3, additional features can be incorporated into the prediction model to automatically stratify the disease severity, making the deployment more practical and relevant in real-world scenarios.

.

CÁC CÔNG TRÌNH ĐÃ CÔNG BỐ

[CT1]. **Dương Thị Kim Chi**, Trần Văn Lăng, Lê Mậu Long, *Xác định các tham số quan trọng cho việc thiết kế gene dùng trong tái tổ hợp*. Kỷ yếu Hội nghị Quốc gia lần IX về Nghiên cứu cơ bản và Ứng dụng Công nghệ thông tin, Cần Thơ, 04-05/8/2016, ISBN: 978-604-913-472-2, NXB. KHTN&CN, DOI: 10.15625/vap.2016.000103, tr. 846-853.

[CT2]. **Dương Thị Kim Chi**, Trần Văn Lăng, Huỳnh Xuân Hiệp, *Dự đoán gene biểu hiện cao cho thiết kế gene dùng trong tái tổ hợp*. Kỷ yếu Hội nghị Quốc gia lần IX về Nghiên cứu cơ bản và Ứng dụng Công nghệ thông tin, Cần Thơ, 04-05/8/2016, ISBN: 978-604-913-472-2, NXB. KHTN&CN, DOI: 10.15625/vap.2016.00017, tr. 134-142.

[CT3] **Dương Thị Kim Chi**, Trần Văn Lăng, *Mô hình dự đoán gene tương quan với hệ thống vật chủ dùng trong Tái tổ hợp*. Kỷ yếu Hội nghị Quốc gia lần X về Nghiên cứu cơ bản và ứng dụng Công nghệ thông tin, Đà Nẵng, 17-18/8/2017, ISBN:978-604-913-614-6, Nxb. KHTN&CN, DOI:10.15625/vap.2017.00049, tr. 408 – 416.

[CT4]. **Dương Thị Kim Chi**, Nguyễn Thị Ngọc Nhi, Nguyễn Thế Bảo, Lê Mậu Long, Phạm Công Xuyên, *Ứng dụng học máy cho định danh loài nấm mới*”, Kỷ yếu Hội nghị Quốc gia lần XI về Nghiên cứu cơ bản và ứng dụng Công nghệ thông tin, Hà Nội, 09-10/8/2018, ISBN:978-604-913-749-5, Nxb. KHTN&CN, DOI:10.15625/vap.2018.00069 tr. 529 – 536.

[CT5]. **Dương Thị Kim Chi**, Trần Văn Lăng, Trần Bá Minh Sơn, *Mô hình học tăng cường độ dốc dự đoán bệnh CoViD-19 từ dữ liệu lâm sàng*, Kỷ yếu Hội thảo quốc gia STAIS-2022 “Ứng dụng Công nghệ thông minh trong công nghiệp 4.0 thành phố thông minh và phát triển bền vững”, Bình Dương, 14/7/2022, ISBN: 978-604-357-047-2, tr. 245-253

[CT6]. **Dương Thị Kim Chi**, Tran Van Lang, Thanh Q. Nguyen; *Clinical data-driven approach to identifying CoViD-19 and influenza from a gradient-boosting model*, *Cogenet Engineering*; Volume 10, Issue 1, 2023, DOI: 10.1080/23311916.2023.2188683 (ESCI, Scopus, Scimago Q2)

[CT7]. **Thi Kim Chi Duong**, Van Lang Tran, The Bao Nguyen, Thi Thuy Nguyen, Ngoc Trung Kien Ho Thanh Q. Nguyen, *Ensemble learning-based approach for automatic classification of termite mushrooms*, Frontiers Genetics, Sec. Computational Genomics, 2023, DOI: 10.3389/fgene.2023.1208695 (SCI-E, Scopus, Scimago Q2)