

Development of Support System for Acupuncture and Moxibustion Treatment by Using Machine Learning



Hang Yang
The Graduate School of East Asian Studies
Yamaguchi University

A thesis submitted for the degree of
Doctor of Philosophy

March 2026

Abstract

Title of Thesis: Development of Support System for Acupuncture and
Moxibustion Treatment by Using Machine Learning

Name of Degree Candidate: Hang Yang
Asian Education Course
The Graduate School of East Asian Studies
Yamaguchi University, Japan

Degree and Year: Doctor of Philosophy, 2026

Dissertation directed by: Dr. Qi-Wei Ge
Professor
The Graduate School of East Asian Studies
Yamaguchi University, Japan

Acupuncture and moxibustion treatment (AMT), a core component of traditional Chinese medicine (TCM), has been widely used to restore balance and alleviate pain by stimulating specific acupoints. However, acupoint prescription remains highly dependent on the clinical experience of individual practitioners, leading to considerable variability and a lack of standardized protocols. Although large amounts of clinical knowledge have been accumulated over thousands of years, systematic methods for objective analysis and knowledge discovery are still limited. Artificial intelligence (AI), particularly Machine Learning, has shown great potential in AMT by uncovering data-driven patterns that support clinical decision-making, in particular identifying acupoint selection from large-scale data and developing personalized treatment prescriptions. Nevertheless, its application in AMT still faces critical challenges, including data quality, model reliability, and clinical trustworthiness.

This thesis presents an intelligent AMT support system designed to overcome the limitations of traditional acupoint selection. By integrating a hierarchical classification framework with interpretable machine learning, the system provides personalized acupoint prescriptions and enhances the transparency of clinical decision-making. With physician oversight, the system integrates AI recommendations with human expertise to ensure alignment with TCM principles and established clinical standards.

Specifically, to address the challenges of data quality, model reliability, and clinical trustworthiness, this study conducted the following work:

We built a dataset of 5,000 cases covering 159 diseases and introduced a tailored data augmentation method, expanding it to 10,138 samples. This establishes a foundation for future data expansion.

Based on our analysis of AMT data from a machine learning perspective, we adopted the Hierarchical Attention-based Recurrent Neural Network (HARNN) model for hierarchical classification tasks in this domain and validated its performance and applicability through extensive computational experiments.

To further enhance the reliability and interpretability of machine learning results, we introduced the Local Interpretable Model-Agnostic Explanations (LIME) method into the field of AMT.

Building upon this, we design an AI-assisted physician decision-making framework, integrating AI recommendations with physician expertise to ensure that AI-driven AMT aligns with TCM principles and clinical standards. Finally, we implement the system using Streamlit, providing an interactive interface for AI-assisted clinical decision-making.

The computational experiments demonstrated that on the original dataset of 5,000 samples, HARNN achieved an IoU (Intersection over Union) of 0.883 in predicting acupoint prescriptions. After data augmentation, the IoU reached 0.954 in 5-fold cross-validation, and 0.932 on a test set of 1,000 original samples. The application of LIME enables intuitive visualization of the model's prediction rationale, thereby enhancing the model's reliability.

This study develops a hierarchical and interpretable machine learning model that predicts acupoint prescriptions based on symptoms, integrat-

ing HARNN for hierarchical classification and LIME for interpretability. This provides an effective technical approach and methodology for the intellectualization of AMT.

学位論文内容要旨

学位論文題目：機械学習による鍼灸治療に向けた支援
システムの開発

指導教員：葛 崎偉 教授

申請者名：楊 航 (ヨウ コウ)

令和 2 年度入学

山口大学大学院東アジア研究科 (博士後期課程)
アジア教育開発コース

鍼灸治療 (Acupuncture and moxibustion treatment: AMT) は、東洋医学 (Traditional Chinese Medicine: TCM) の中核的な要素であり、特定の経穴を刺激することによって身体のバランスを整え、疼痛を緩和する目的で広く用いられてきた。しかし、経穴の処方は依然として個々の臨床医の経験に大きく依存しており、その結果、施術内容にばらつきが生じ、標準化されたプロトコルの欠如が問題となっている。数千年にわたって膨大な臨床知識が蓄積されてきたものの、客観的な分析や知識発見を行う体系的な方法は依然として限られている。

近年、人工知能 (AI)、特に機械学習 (Machine Learning) は、大規模データから臨床判断を支援するデータ駆動型パターンを抽出し、経穴選択の最適化や個別化治療処方の構築において大きな可能性を示している。しかし、AMT への応用においては、データ品質、モデルの信頼性、臨床的妥当性など、依然として重要な課題が残されている。

本論文では、従来の経穴選択における限界を克服することを目的として、知的な鍼灸治療支援システムを提案する。本システムは、階層的な分類フレームワークと説明可能な機械学習を統合することにより、個別化された経穴処方を提供し、臨床判断の透明性を高めるものである。さらに、医師の監督のもと、AI による推奨と専門家の知見を

統合することで、東洋医学の原則および既存の臨床基準との整合性を確保している。

特に、本研究ではデータ品質、モデルの信頼性、臨床的妥当性といった課題に対応するため、以下の研究を行った。

まず、159 種類の疾患を対象とする 5,000 件の症例データセットを構築し、独自のデータ拡張手法を導入することで 10,138 件に拡張した。これにより、今後のデータ拡充に向けた基盤を確立した。次に、機械学習の観点から AMT データを分析し、この分野における階層的分類タスクに適した階層型アテンション付きリカレントニューラルネットワーク（Hierarchical Attention-based Recurrent Neural Network: HARNN）モデルを採用し、広範な計算実験を通じてその性能と有用性を検証した。さらに、機械学習モデルの解釈性を高めるために、LIME（Local Interpretable Model-Agnostic Explanations）を導入した。これらの成果に基づき、AI による提案と医師の専門的判断を統合する意思決定支援フレームワークを設計し、AI による鍼灸治療が東洋医学の理論と臨床基準に適合するようにした。最後に、Streamlit を用いてシステムを実装し、AI 支援型臨床意思決定のための対話型ユーザーインターフェースを提供した。

計算実験によると、5,000 件の元データセットにおいて、HARNN は経穴処方予測において IoU（Intersection over Union）= 0.883 を達成した。データ拡張後、5 分割交差検証で IoU=0.954、1,000 件のテストセットでは IoU=0.932 を示した。また、LIME を適用することで、モデルの予測根拠を直感的に可視化し、モデルの信頼性を高めることができた。

本研究は、症状に基づいて経穴処方を予測する階層的かつ可解釈な機械学習モデルを開発し、階層分類に HARNN を、可解釈性向上に LIME を組み合わせたものである。これにより、鍼灸治療の知能化に向けた有効な技術的アプローチと方法論を提供する。

**Development of Support System for Acupuncture and
Moxibustion Treatment by Using Machine Learning**

by

HANG YANG

Dissertation submitted to the Graduate School of East
Asian Studies of Yamaguchi University,
in partial fulfillment of the requirements for the degree of
Doctor of Philosophy

Advisory Committee:

Professor Qi-Wei GE, Chairman/Advisor
Professor Mitsuru NAKATA
Professor Chisato KITAZAWA

©Yamaguchi University. All rights reserved.

Acknowledgements

I would like to express my deepest gratitude to my supervisor, Professor Qi-Wei GE, for his invaluable guidance, continuous encouragement, and insightful advice throughout my doctoral study. His dedication to research and rigorous attitude have greatly inspired me and shaped my academic journey.

I am very grateful to the other members of the examination committee for this dissertation: Professor Mitsuru NAKATA, Professor Chisato KITAZAWA, and Professor Shingo YAMAGUCHI. Their careful review and valuable suggestions have greatly contributed to improving the quality of the final work.

I am also sincerely grateful to the members of my laboratory and colleagues for their kind support and helpful discussions during my research.

Finally, I would like to express my heartfelt appreciation to my family for their endless love, patience, and understanding. Their constant support has been the greatest source of strength that enabled me to complete this dissertation.

Contents

Abstract	iii
Acknowledgements	xi
1 Introduction	1
1.1 Research Background	1
1.2 Research Motivation	4
1.3 Thesis Organization	6
2 Literature Review	7
2.1 Acupuncture and Moxibustion Treatment	8
2.1.1 Basic Principles of AMT	8
2.1.2 Acupoint Prescription	9
2.2 Machine Learning for AMT	10
2.2.1 Traditional machine learning	11
2.2.2 Ensemble Learning	11
2.2.3 Deep Learning	14
2.2.4 Interpretability in Machine Learning for AMT	15
3 Application of Machine Learning Methods to AMT	16
3.1 Materials and Methods	17
3.1.1 Database construction	17
3.1.2 Algorithm selection	19
3.1.3 Data preprocessing	22
3.1.4 Evaluation Metrics	24
3.2 Experiments	26
3.2.1 Comparative experiments	26
3.2.2 Further Experimental Analysis of Seq2seq with attention . . .	27
3.2.3 Case Study	29

3.2.4	Discussion	29
3.3	Conclusion	31
4	Hierarchical and Interpretable Machine Learning Model for Acupoints De-termination	32
4.1	Materials and Methods	33
4.1.1	Data analysis and database construction	33
4.1.1.1	Data sources and analysis	33
4.1.1.2	Database construction	35
4.1.1.3	Data expansion	36
4.1.2	Hierarchical and Interpretable Machine Learning Model . . .	41
4.1.2.1	Overview of research framework	41
4.1.2.2	Hierarchical classification module	41
4.1.2.3	Result interpretation module	43
4.1.2.4	Evaluation metrics	44
4.2	Experiments	45
4.2.1	Model evaluation and performance comparison	45
4.2.1.1	Hierarchical model results on original dataset	45
4.2.1.2	Performance improvement through data augmentation	46
4.2.1.3	Model comparison on expanded dataset	46
4.2.2	Case Study	50
4.2.2.1	Case 1	50
4.2.2.2	Case 2	53
4.2.3	Discussion	57
4.2.3.1	Model Performance and Interpretability	57
4.2.3.2	Limitations and Challenges	58
4.3	Conclusion	60
5	System Design and Implementation	61
5.1	Overview of the Intelligent Support System	62
5.2	System Implementation	63
5.3	Case Study	66
5.3.1	Experimental Setup	66
5.3.2	Case 1	66
5.3.3	Case 2	74
5.4	Conclusion	81

CONTENTS	xv
----------	----

6 Conclusions	83
----------------------	-----------

Bibliography	85
---------------------	-----------

List of Figures

2.1	The decision-making process of traditional AMT.	10
2.2	Structure of Bagging.	12
2.3	Structure of Boosting.	13
2.4	Structure of Stacking.	13
3.1	Take “chronic fatigue syndrome” as an example to illustrate the construction process of the database. Here the text of this figure is a translation of the Chinese text on page 845 of reference [2].	18
3.2	Coding method of the traditional machine learning algorithms and FNN.	23
3.3	Feature encoding of TextCNN.	23
3.4	Exemplary Acupoint Prescription by Seq2seq with Attention.	29
4.1	Take “multiple sclerosis” as an example to illustrate the construction process of the database. Here the text of this figure is a translation of the Chinese text on page 863 of reference [2].	34
4.2	Hierarchical label structure for disease cases in acupoint prescriptions. The numbers in the figure represent the statistical information of our database.	35
4.3	Diseases, symptoms, acupoints, and subcategories with assigned IDs.	38
4.4	Example of data augmentation: from original to augmented data. . .	38
4.5	t-SNE visualization of feature distributions for original and augmented data.	40

4.6	Framework of the hierarchical and interpretable machine learning model. D : Textual data of a disease case. k : The dimension of the embedding vector, defining the length of each symptom's representation. γ : Hierarchical category structure corresponding to the disease cases. V : The complete semantic representation of a disease case's symptoms. S : The vector matrix of the hierarchical category structure γ . HAM: the hierarchical attention-based memory unit. $!^{h-1}$: the information of previous level h . P^i : The prediction result of the i -th layer. L^i : The interpretability results of LIME in the i -th layer.	41
4.7	Hierarchical attention-based memory unit [15].	42
4.8	Trend analysis of performance metrics for three hierarchical levels with data expansion.	46
4.9	LIME visualization of a disease case in disease name classification. . .	50
4.10	LIME visualization of prediction probabilities for disease name, subcategory, and acupoint.	51
4.11	LIME visualization of the top 15 acupoints ranked by prediction probabilities to explain the model's decision process.	52
4.12	LIME visualizations for Level 1 predictions in Case 1 across three repeated runs.	53
4.13	Hierarchical Attention on Symptoms for Case 1 (Fold 1 Model). . . .	54
4.14	Hierarchical Attention on Symptoms for Case 1 (Fold 2 Model). . . .	55
5.1	Diagnosis and Treatment Workflow with the Intelligent AMT Support System.	62
5.2	Main Interface of the AMT Intelligent Support System.	64
5.3	Treatment Record Interface of the AMT Intelligent Support System. . .	65
5.4	System Output for Case 1: Hierarchical Predictions Including Disease Name, Subcategory, and Recommended Acupoints. (Source: Created by the authors)	67
5.5	Top Prediction Probabilities for Disease Names - Case 1 (Level 1). (Source: Created by the authors)	68
5.6	LIME Explanation for Top-1 Disease Name Prediction - Case 1 (Level 1). (Source: Created by the authors)	69
5.7	Top Prediction Probabilities for Disease Subcategories - Case 1 (Level 2). (Source: Created by the authors)	69

5.8	LIME Explanation for Top-1 Disease Subcategory Prediction - Case 1 (Level 2). (Source: Created by the authors)	70
5.9	Top Prediction Probabilities for Acupoint Prescriptions - Case 1 (Level 3). (Source: Created by the authors)	70
5.10	LIME Visualization of the Top 15 Predicted Acupoints for Case 1 (Part 1).	71
5.11	LIME Visualization of the Top 15 Predicted Acupoints for Case 1 (Part 2).	72
5.12	LIME Visualization of the Top 15 Predicted Acupoints for Case 1 (Part 3).	73
5.13	System Output for Case 2: Hierarchical Predictions Including Disease Name, Subcategory, and Recommended Acupoints. (Source: Created by the authors)	74
5.14	Top Prediction Probabilities for Disease Names - Case 2 (Level 1). (Source: Created by the authors)	75
5.15	LIME Explanation for Top-1 Disease Name Prediction - Case 2 (Level 1). (Source: Created by the authors)	75
5.16	Top Prediction Probabilities for Disease Subcategories - Case 2 (Level 2). (Source: Created by the authors)	76
5.17	LIME Explanation for Top-1 Disease Subcategory Prediction - Case 2 (Level 2). (Source: Created by the authors)	76
5.18	Top Prediction Probabilities for Acupoint Prescriptions - Case 2 (Level 3). (Source: Created by the authors)	77
5.19	LIME Visualization of the Top 15 Predicted Acupoints for Case 2 (Part 1).	78
5.20	LIME Visualization of the Top 15 Predicted Acupoints for Case 2 (Part 2).	79
5.21	LIME Visualization of the Top 15 Predicted Acupoints for Case 2 (Part 3).	80

List of Tables

2.1	Meridians and acupoints in human body.	9
3.1	Database-1.	19
3.2	Database-2.	20
3.3	Four possible situations for the results of classification	25
3.4	The evaluation results of an acupoint prescription with two acupoints using different evaluation criteria.	25
3.5	Experimental results of different machine learning algorithms.	26
3.6	Experimental results of the 5-fold cross validation of Seq2seq with attention.	27
3.7	Experimental results of the 5-fold cross validation of Seq2seq with attention on ascending and descending data.	27
3.8	Experimental results of the 5-fold cross validation of Seq2seq without attention.	28
3.9	Experimental results of the 5-fold cross validation of Seq2seq with attention using 6000 augmented data.	28
4.1	Rotated disease case table.	37
4.2	Numerical representation of database elements.	39
4.3	Performance metrics for three hierarchical levels on 5,000 original data.	45
4.4	Summary of baseline models for extended dataset.	47
4.5	Experimental results of different machine learning algorithms on the expanded dataset.	47
4.6	Experimental results of the 5-fold cross validation of HARNN on the expanded dataset.	48
4.7	Paired t-test results comparing HARNN with other models.	48
4.8	Related Training and Test Samples for Case 2.	57

Chapter 1

Introduction

1.1 Research Background

Acupuncture and Moxibustion Treatment (AMT) originated in China and is a traditional medical practice with remarkable efficacy and minimal side effects. As early as the 6th century, AMT was introduced to Korea and Japan. Today, it is widely practiced in over 190 countries and regions worldwide and has gained recognition from the World Health Organization (WHO) [1]. It has been proved through clinical practice that it has certain effects on more than 300 kinds of diseases, such as internal medicine, surgery, gynecology and pediatrics, and has good effects on about 100 kinds of diseases, such as chronic fatigue syndrome, withdrawal symptoms, etc. It is particularly recognized for its efficacy in managing pain, neurological conditions, and chronic diseases, serving as a valuable alternative or complementary therapy in modern healthcare. AMT is a clinical discipline based on Traditional Chinese Medicine (TCM) theory, which mainly treats diseases using acupuncture and moxibustion techniques guided by the Channel Theory (Jing-Luo) [2]. Acupoints are special parts of human body surface that are infused with Qi from internal organs and meridians, and also the places where acupuncture, moxibustion and other stimuli are operated. Stimulating appropriate acupoints has the effect of dredging meridians, harmonizing Qi and blood, restoring the balance of yin and yang, and coordinating viscera, so as to achieve the purpose of disease prevention and treatment. Acupoint prescription is the basis of AMT. Clinically, acupoints selection is generally based on the acupuncturist's judgment of the patient's condition through four diagnosis, namely "inspection, auscultation and olfaction, interrogation, and palpation", and the specific acupoint prescription is guided by the theoretical principles of TCM and the clinical experience of acupuncturists. For thousands of years, acupuncturists have never stopped exploring acupoint prescriptions in acupuncture and moxibustion treatment. Up to now,

there are no standardized protocols for acupoint selection, as each acupoint in TCM theory has multiple functions, leading to complex relationships between acupoints and diseases. Despite its long-standing history and widespread application, the clinical practice of acupuncture remains largely experience-driven, resulting in inconsistencies in acupoint prescriptions. The absence of standardized treatment protocols presents challenges for reproducibility and scientific validation, thereby hindering the broader integration of acupuncture into evidence-based medical frameworks [3]. After thousands of years of development, a large number of clinical experiences of AMT have been recorded in the form of text. If these data can be used effectively, it can provide assistance for acupuncturists and even support for patients' self-help treatment in the near future.

Artificial Intelligence (AI for short hereafter) technology has flourished and has been widely used in many fields in recent years. As one of the core technologies of AI, machine learning has been applied to all walks of life including the medical field [4]. Machine learning is to recognize input data sets through computer, encode the data into computer models or algorithms, train appropriate mathematical models, and test and verify the trained models on new data. The ability of machine learning to analyze data related to medical treatment and results is expected to transform medicine into a data-driven, results oriented discipline, which will have a profound impact on the detection, diagnosis and treatment of diseases [5]. Deep learning is a new research direction of machine learning, that is, through data-driven way, using nonlinear activation function to introduce nonlinear factors, insinuating higher-level and more abstract vector representation, to build multivariate neural network, so that machines can learn and think like people through this neural network, so as to solve the complex relationship problems beyond the reach of human resources. Deep learning has good generalization ability and has played an important role in the medical field in recent years [6]. The advent of generative AI technologies, such as ChatGPT, has introduced innovative solutions for medical applications [7]. These technologies have not only revolutionized clinical diagnosis and pharmaceutical development but have also opened unprecedented avenues for AMT. Therefore, it is expected to analyze the patient's condition and decide the corresponding treatment schemes in the way of artificial intelligence by the traditional machine learning methods and deep learning methods through the learning of a number of cases. The application of AI in acupuncture holds the potential to enhance the precision of acupoint prescriptions, improve therapeutic outcomes, and promote the standardization of clinical practices.

Moreover, AI-assisted systems can support practitioners by generating recommendations informed by extensive clinical data, contributing to more consistent, accurate, and personalized treatment strategies.

The rapid advancement of AI has led to a series of significant breakthroughs in various aspects of AMT practice, including acupoint selection and prescription, identification of AMT techniques, and prediction of AMT efficacy [8]. In terms of technical approaches, existing studies predominantly rely on data mining and unsupervised machine learning techniques. For instance, AI-based data mining methods have been employed to uncover effective acupoint combinations, principles of acupoint compatibility, and optimal AMT plans. These advancements have significantly enhanced the reliability and stability of acupuncture's clinical efficacy [9, 10]. Unsupervised machine learning techniques, such as association rule mining and clustering algorithms, are among the most widely used approaches in this field, with the Apriori algorithm being particularly prominent [11, 12]. Regarding research content, most existing studies concentrate on exploring acupoint prescriptions tailored to specific diseases. For example, machine learning-based acupoint selection methods have been applied in AMT for various conditions, including postoperative syndromes of lumbar disc herniation, facial paralysis, neurogenic dysphagia, urticaria, and insomnia [8].

1.2 Research Motivation

AI has demonstrated potential in AMT, offering data-driven support for clinical decision-making. A key application is to generate personalized acupoint prescriptions based on patient symptoms, yet challenges remain in data processing, model selection, and clinical applicability. This study aims to address these challenges by developing a machine learning-based AMT support system that provides personalized acupoint prescriptions to assist physicians in clinical decision-making. Differing from existing research, this study research employs supervised machine learning techniques and integrates data from multiple diseases into a unified model. This approach enables the model to comprehensively learn the complex relationships between symptom combinations and acupoint prescriptions, thereby providing personalized acupoint prescriptions tailored to each patient’s condition.

While this approach demonstrates great potential, its development and implementation within this study present several key challenges: 1. Although a large amount of textual data on AMT exists, there is currently no standardized method for pre-processing this raw data for use in machine learning, nor are there publicly available datasets specifically prepared for such applications; 2. The task requires identifying and selecting appropriate acupoints from a set of several hundred candidates based on combinations of patient symptoms. This constitutes a multi-label classification problem, which remains a significant challenge in machine learning due to factors such as a large and complex output space [13] and severe label imbalance commonly observed in long-tailed data distributions [14]. 3. In the medical domain, prediction reliability is paramount. The lack of transparency in machine learning models, particularly complex deep learning models, hinders their adoption in healthcare. Ensuring both the reliability and interpretability of predictions remains a critical challenge for building trust and enabling safe clinical applications.

To address the above challenges, this study conducted the following work: 1. We built a dataset of 5,000 cases covering 159 diseases and introduced a tailored data augmentation method, expanding it to 10,138 samples. This establishes a foundation for future data expansion. 2. Based on our analysis of AMT data from a machine learning perspective, we adopted the Hierarchical Attention-based Recurrent Neural Network (HARNN) model [15] for hierarchical classification tasks in this domain and validated its performance and applicability through extensive computational experiments. 3. To further enhance the reliability and interpretability of machine learning results, we introduced the Local Interpretable Model-Agnostic Explanations (LIME)

[16] method into the field of AMT. 4. A web-based platform developed with Streamlit [17] enables clinicians to input patient symptoms and receive AI-generated diagnostic suggestions along with interpretability visualizations. To our knowledge, this study is the first to introduce hierarchical learning and interpretable machine learning methods into the field of AMT.

1.3 Thesis Organization

The rest of this thesis is organized as follows:

Chapter 2 presents a comprehensive literature review. It first introduces the basic principles of AMT and the characteristics of acupoint prescription. It then reviews the development of machine learning techniques in the context of AMT, including traditional machine learning, ensemble learning, deep learning, and issues of interpretability.

Chapter 3 explores the preliminary application of a spectrum of machine learning methods to acupoint prescription in AMT. The study investigates three primary categories of models: traditional machine learning, ensemble methods, and deep learning. The chapter then outlines the experimental setup, presents the results, and provides a critical discussion of the limitations of these approaches.

Chapter 4 proposes our novel hierarchical and interpretable machine learning model for acupoint determination. This chapter details the methodology, from dataset construction and model architecture to the evaluation metrics employed. The experimental results are then presented and comprehensively analyzed to validate the effectiveness and advantages of the proposed approach.

Chapter 5 describes the translation of our proposed model into a practical application. It presents the design and implementation of an interactive AMT system that leverages the hierarchical and interpretable model from Chapter 4 to support clinical decision-making.

Chapter 6 concludes the thesis by summarizing the key findings and contributions. It also discusses limitations of the current study and suggests directions for future research to improve the proposed methods and extend their applications.

Chapter 2

Literature Review

This chapter provides a comprehensive review of the literature relevant to this dissertation. The objective is to establish a theoretical and methodological foundation by examining two major research domains: AMT and machine learning approaches applied to AMT. First, Section 2.1 introduces the fundamental concepts of AMT, including its basic principles and the role of acupoint prescriptions in clinical practice. This part highlights the theoretical underpinnings of AMT and sets the stage for understanding how computational approaches can be integrated with traditional medical knowledge. Next, Section 2.2 surveys machine learning methods in the context of AMT. It begins with traditional algorithms, progresses to ensemble learning, and then covers advances in deep learning. Finally, special attention is given to interpretable machine learning, which aims to provide greater transparency and improve trustworthiness in the application of computational models to medical decision-making. By reviewing these two domains, this chapter aims to bridge traditional medical theories with modern computational techniques, thereby laying the groundwork for the methodological developments presented in subsequent chapters.

2.1 Acupuncture and Moxibustion Treatment

2.1.1 Basic Principles of AMT

The basic contents of AMT mainly include AMT theory, AMT technology and clinical application of AMT. AMT theory mainly includes meridian theory and the rules of acupoints. AMT technology mainly includes acupuncture, moxibustion and other needling methods. Clinical application of AMT is a comprehensive application of AMT theory and technology. Meridian theory is an important part of TCM, which covers the distribution, physiological functions, pathological changes of human meridian system and its relationship with internal organs, and runs through the diagnosis and treatment of AMT [2]. Meridians were gradually discovered by ancient medical experts through long-term medical practice and continuous observation and summary. Meridians are closely related to the tissues and organs around the body and plays an important role in physiological function and pathological changes. Meridians are the channels for human body to transport Qi and blood, which are all over the body. Here Qi, as a special concept in TCM, is the most fundamental substance in the construction of the human body and in the maintenance of its life activities. Acupoints are special parts of human body surface that are infused with Qi from internal organs and meridians, and also the places where acupuncture, moxibustion and other stimuli are operated. Stimulating appropriate acupoints has the effect of dredging meridians, harmonizing Qi and blood, restoring the balance of yin and yang, and coordinating viscera, so as to achieve the purpose of disease prevention and treatment. Meridians consist of 12 principal meridians and 8 extra meridians. Each principal meridian has acupoints. In 8 extra meridians, only conception and governor vessels have acupoints. There are 361 acupoints in 12 principal meridians and the meridians of conception governor vessels, and 48 acupoints that exist outside meridians in human body. There are totally 409 acupoints identified by WHO (World Health Organization) [18]-[20]. The names and WHO notations of each meridian, as well as the acupoints, are shown in Table 2.1. In addition, the parts that have neither specific name nor fixed position but are stimulated by tenderness or other reactions are collectively referred to as “Ashi points”. Ashi points are usually near or far away from lesion, and they usually appear with the occurrence of a disease and disappear with the recovery of a disease [2].

Table 2.1: Meridians and acupoints in human body.

Meridian name	WHO notation of meridian	WHO notation of acupoints	Number of acupoints
Lung Meridian	LU	LU1~LU11	11
Large Intestine Meridian	LI	LI1~LI20	20
Stomach Meridian	ST	ST1~ST45	45
Spleen Meridian	SP	SP1~SP21	21
Heart Meridian	HT	HT1~HT9	9
Small Intestine Meridian	SI	SI1~SI19	19
Bladder Meridian	BL	BL1~BL67	67
Kidney Meridian	KI	KI1~KI27	27
Pericardium Meridian	PC	PC1~PC9	9
Triple Energizer Meridian	TE	TE1~TE23	23
Gallbladder Meridian	GB	GB1~GB44	44
Liver Meridian	LR	LR1~LR14	14
Conception Vessel	CV	CV1~CV24	24
Governor Vessel	GV	GV1~GV28	28
Extras Points	EX	EX-HN1~EX-HN15	48
		EX-CA1	
		EX-B1~EX-B9	
		EX-UE1~EX-UE11	
		EX-LE1~EX-LE12	
Total			409

2.1.2 Acupoint Prescription

The decision-making process of traditional AMT is shown in Figure 2.1. Acupuncturists obtain the symptoms of patients through the four diagnostic methods, and analyze and summarize these symptoms according to meridian theory, so as to clarify the etiology, location, pathogenesis and urgency of the disease. On this basis, the appropriate AMT prescriptions are determined by comprehensively considering the meridian theory and the rules of acupoints. AMT prescription includes acupoint prescription and manipulation. Acupoint prescription is the first component of AMT prescription. Each acupoint in the body has relative specificity and may have the same or different therapeutic functions. Selecting acupoints with the same or similar functions can enhance the effectiveness of the treatment by strengthening the synergistic effect between the acupoints. Manipulation is the second component of AMT prescription, which includes treatment methods, specific operation and timing of treatment [2]. In this paper, we focus our discussion on ③ deciding acupoint prescription.

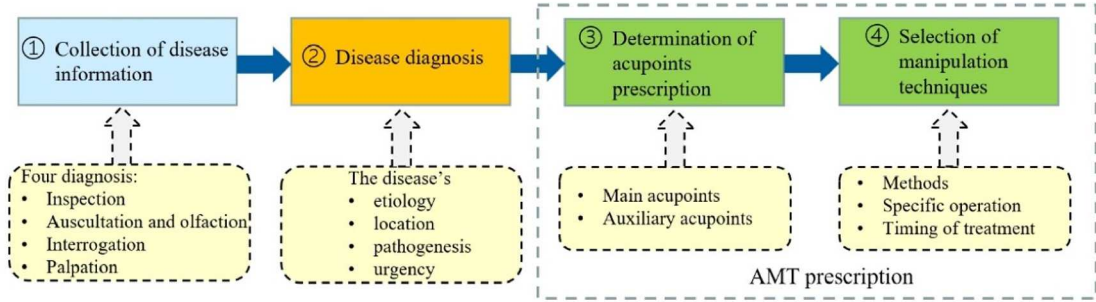


Figure 2.1: The decision-making process of traditional AMT.

2.2 Machine Learning for AMT

The process of selecting acupoints can be regarded as establishing the corresponding relationship between the combination of a series of symptoms and the combination of acupoints in the acupoint prescriptions, rather than the corresponding relationship between a single symptom and a single acupoint. In machine learning, this kind of problem can be classified as multi-label classification problem. The combination of symptom numbers is used as features of the model input, while the combination of acupoint numbers is used as labels of the model output.

Suppose $\chi = \mathbf{R}^d$ (or $\mathbf{Z}^d$) denotes the d -dimensional instance space, and $\mathbf{Y} = \mathbf{f}_{y_1}; \mathbf{y}_2; \dots; \mathbf{y}_q \mathbf{g}$ denotes the label space with q possible class labels. The task of multi-label learning is to learn a function $\mathbf{h} : \chi \rightarrow \mathbf{Y}$ from the multi-label training set $\mathbf{D} = \{(\mathbf{x}_i; \Upsilon_i) \mid 1 \leq i \leq m\}$. For each multi-label example $(\mathbf{x}_i; \Upsilon_i)$, $\mathbf{x}_i \in \chi$ is a d -dimensional feature vector $(x_{i1}; x_{i2}; \dots; x_{id})^T$ and $\Upsilon_i \in \mathbf{Y}$ is the set of labels associated with $\mathbf{x}_i$. For any unseen instance $\mathbf{x} \in \chi$, the multi-label classifier $\mathbf{h}(\cdot)$ predicts $\mathbf{h}(\mathbf{x}) \subseteq \mathbf{Y}$ as the set of proper labels for $\mathbf{x}$ [13].

In the acupuncture and moxibustion treatment support system, χ indicates the set of symptoms. Each symptom in the database is encoded as a vector of the same length. $\mathbf{Y}$ indicates the set of 410 acupoints, which comprises the 409 acupoints certified by the WHO and Ashi points. $\mathbf{D}$ indicates the established database. By learning the data in the database, the function $\mathbf{h}$ from symptom combinations to acupuncture treatment scheme is obtained. When inputting the symptom combination $\mathbf{x}$ not in the database, the trained function $\mathbf{h}$ can predict the corresponding acupuncture treatment scheme $\mathbf{h}(\mathbf{x})$.

2.2.1 Traditional machine learning

Machine learning is to study how computers imitate human learning behavior, acquire new knowledge or experience, reorganize existing knowledge structure and improve their own performance. Machine learning can learn the rules and patterns of data in massive data by computer, and obtain potential information. It is widely used to solve the problems of classification, regression, clustering and so on.

Maron et al.[21] proposed NB algorithm for classification according to probability principle based on bayesian theory. Cover et al.[22] proposed KNN classification algorithm based on distance measurement. Breiman proposed the early Decision Tree (DT) classification algorithm-cart algorithm, which uses the tree structure algorithm to divide the data into discrete classes[23]. Support vector machine (SVM) is a kind of generalized linear classifier that classifies data according to supervised learning[24]. Artificial Neural Network (ANN) classification adjusts the parameters of artificial neural network according to the given training samples to make the network output close to the known sample class label[25]. Based on the traditional methods of machine learning, there are three methods to solve a multi-label classification problem, namely: Problem Transformation, Adapted Algorithm, Ensemble approaches.

Problem transformation: the basic idea of this kind of method is to transform the multi-label classification problem into other known learning problems for solution. Representative algorithms include LP[26], Binary Relevance[27], Calibrated Label Ranking[28], Random k-labelsets[29].

Adapted Algorithm: The algorithm adaptation method is different from the problem transformation method. The problem transformation method is to transform the multi-label problem into one or more single class label problems. The algorithm adaptation method is to study the algorithm on the basis of multi label. In recent years, there are more and more adaptive algorithms for multi label problems, such as representative learning algorithms ML-KNN[30], Rank SVM[31], LEAD[32], etc.

Ensemble approaches: Ensemble approaches are machine learning methods that integrates a variety of basic models to complete the final decision. Ensemble always produces better results.

2.2.2 Ensemble Learning

Ensemble learning combines several individual models to obtain better generalization performance. Ensemble learning has been experimentally proven to be effective for multi-label classification problems [33].

Bagging: Bagging is a widely used method for constructing ensemble-based algorithms that can improve classifier performance. The fundamental concept behind Bagging is to generate a series of independent observations with the same size, and distribution as that of the original data. Given the series of observations, generate an ensemble predictor which is better than the single predictor generated on the original data. Figure 2.2 shows the diagram of the bagging technique. Here, D_i represents the bagged datasets, M_i represents the algorithms and F_{Output} calculates the final outcome [33]. Random Forest (RF) is a machine learning algorithm that enhances the performance of decision trees by utilizing a Bagging technique [34]. The primary objective of this approach is to reduce the correlation between trees and mitigate the issue of overfitting. This method effectively leverages the diversity among the trees to improve the overall quality of the predictions.

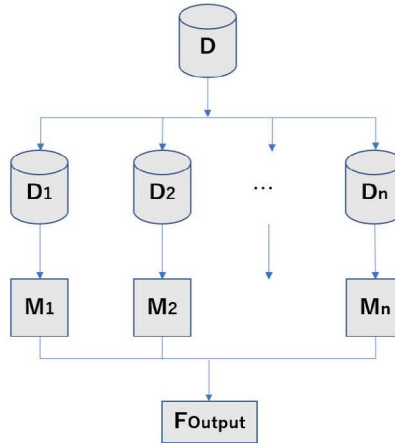


Figure 2.2: Structure of Bagging.

Boosting: Boosting technique is used in ensemble models for converting a weak learning model into a learning model with better generalization. Figure 2.3 shows the diagram of the boosting technique. This algorithm first trains a base learner on an initial training set, then adjusts the distribution of training samples based on the performance of the base learner, so that the previously misclassified training samples receive more attention in subsequent rounds. The next base learner is then trained based on the adjusted sample distribution. This process is repeated until the number of base learners reaches a pre-specified value T , and the final output is a weighted combination of these T base learners [33]. XGBoost belongs to the Boosting ensemble learning algorithm. XGBoost is an optimized distributed gradient

enhancement library designed to achieve efficiency, flexibility, and portability. It implements machine learning algorithms under the Gradient Boosting framework [35].

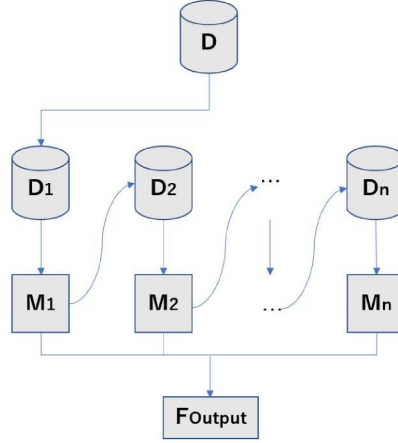


Figure 2.3: Structure of Boosting.

Stacking: Stacking is one of the integration techniques wherein the meta-model is trained on the out-of-fold predictions by base models through Cross Validation (CV) [33]. First, the training set is split into k folds. Each base model is fitted using $(k-1)$ folds and predictions are made with the remaining fold. That procedure is repeated for each of the k folds. The predictions from all base models, along with the outcomes from the training set, are then utilized to train the meta-model. Figure 2.4 shows the stacking technique.

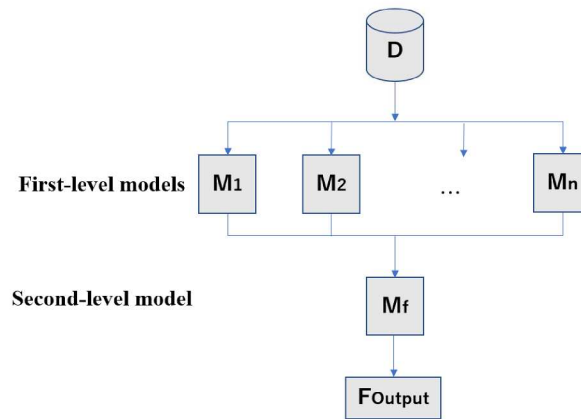


Figure 2.4: Structure of Stacking.

2.2.3 Deep Learning

Deep learning is a special kind of machine learning, which is a new research direction in the field of machine learning. Deep learning can be applied in various fields. According to different applications, the form of deep neural network is also different. Common deep learning models mainly include Fully Connected (FC) network, Convolutional Neural Network (CNN) and Recurrent Neural Network (RNN). The fully connected network structure is the basic deep neural network layer. Each node of the fully connected layer is connected with all nodes of the upper layer. Because all the outputs and inputs of the full connection layer are connected, the parameters of the full connection layer are the most, which requires a considerable amount of storage and computing space. CNN is a neural network used to process data with grid structure, which is often used in the field of computer vision. The difference from FC is that the upper and lower neurons of CNN cannot be connected directly, but through the “convolution core” as the intermediary, and the parameters of the hidden layer are greatly reduced through the sharing of “core”. RNN is also one of the commonly used deep learning models, which is a kind of neural network used to process sequence data. It is often used in the field of natural language processing, and can also be used in the field of computer vision.

Hierarchical Classification is a specialized task in machine learning where labels are organized in a multi-level hierarchy. Each label in this hierarchy can be seen as a node, which may have one or more child nodes, reflecting a parent-child relationship. This hierarchical structure plays a pivotal role in designing effective classification methods, as it captures the dependencies between labels that flat classification approaches fail to consider [36]. One of the earliest studies addressing Hierarchical Text Classification (HTC) is by Koller and Sahami [37], who emphasized the limitations of flat classifiers in hierarchical contexts and highlighted the prevalence of topic hierarchies in HTC tasks. Building on this foundation, Silla and Freitas [38] provided a precise definition of HTC and introduced a unifying framework to classify tasks across various domains, extending beyond textual data. Most recently, Zangari et al. [36] reviewed the theoretical foundations and applications of hierarchical classification, further enriching the understanding of this field. Among these studies, Huang et al. [15] proposed the Hierarchical Attention-based Recurrent Neural Network (HARNN), a novel framework that can automatically annotate documents with the most relevant categories level by level.

2.2.4 Interpretability in Machine Learning for AMT

While machine learning holds great promise for advancing AMT, its application in healthcare often encounters challenges due to the lack of transparency and interpretability of many models, especially deep learning-based approaches. In the medical domain, interpretability is critical not only for gaining the trust of practitioners and patients but also for ensuring the validity and reliability of the predictions made by the model. To address these challenges, interpretable machine learning methods have been increasingly introduced in AMT-related studies. In the domain of acupuncture-assisted decision support, existing work can be broadly categorized into two types.

First, structure-based interpretability methods integrate domain knowledge or specific architectural components, such as knowledge graphs and attention mechanisms[39], into model design in order to align learning processes with traditional medical reasoning. Representative studies model acupuncture knowledge as a structured network that links symptoms, syndromes, meridians, and acupoints, and employ graph-based or attention-based models to explicitly capture these relationships. For instance, recent approaches have utilized knowledge graphs combined with prompt learning or association analysis to systematically map the complex relationships between diseases and acupoints [40, 41]. Similarly, advanced architectures such as multi-graph residual attention networks have been proposed to fuse semantic knowledge for TCM recommendation [42]. These methods provide explicit, structure-driven interpretability that is tightly coupled with the model architecture itself.

Second, post-hoc feature attribution methods aim to explain model predictions after training by estimating the contribution of individual input features. Model-agnostic techniques such as LIME [16] and SHAP [43] have been adopted in several AMT-related applications to provide instance-level explanations. For instance, recent studies have utilized SHAP to identify key diagnostic indicators for diabetic peripheral neuropathy [44] and to visualize clinical factors influencing eye acupuncture treatment outcomes [45], explicitly revealing the variables driving model decisions. Furthermore, Zhu et al. demonstrated the utility of LIME by applying it to an ensemble learning model for ulcerative colitis syndrome differentiation, visualizing how specific symptoms contributed to the diagnostic classification [46].

Overall, the incorporation of interpretable machine learning techniques in AMT enables the diagnostic rationale and acupoint prescription logic learned from data to be explicitly visualized and examined. This not only facilitates clinical understanding and validation, but also enhances the reliability and practical applicability of machine learning-based acupuncture decision support systems.

Chapter 3

Application of Machine Learning Methods to AMT

This chapter transitions from the theoretical foundations reviewed in Chapter 2 to a foundational exploration of applying machine learning to acupoint prescription in AMT. This endeavor addresses three core methodological questions: how to process raw clinical records into a structured dataset suitable for machine learning, which modeling paradigms are best suited for this task, and how to effectively evaluate their performance.

To facilitate this investigation, we developed a purpose-built acupoint prescription database. Using this dataset, we then benchmarked the performance of a diverse set of models spanning three major paradigms: traditional machine learning, ensemble methods, and deep learning.

The outcomes of this foundational study are twofold. On one hand, they demonstrate the feasibility and potential of applying machine learning to acupoint prescription. On the other hand, they reveal the inherent limitations of conventional models in this domain. These insights collectively provide the empirical groundwork and direct motivation for the novel hierarchical model developed in Chapter 4.

3.1 Materials and Methods

In this work, we are to train a model by learning relevant data through machine learning, and provide acupoints prescriptions for treating patients by the trained model based on patients' symptoms. For this purpose, database, algorithms, data preprocessing and evaluation metrics are essential components. 1) Database: As far as we know, there is currently no database available for providing acupoint prescriptions based on patients' symptoms, so we need to establish a database of symptoms and acupoint prescriptions. 2) Algorithms: We regard the problem of acupoint selection as a multi-label classification problem and use 11 algorithms for multi-label classification including 8 classifiers provided by Scikit-learn and Feedforward neural network (FNN) [47], TextCNN [48], and Seq2seq with attention [49]. 3) Data preprocessing: Different algorithms have different requirements for data representation, so we have different preprocessing for data when using different algorithms. 4) Evaluation Metrics: The traditional evaluation criteria in multi-label classification problems of accuracy, precision, and recall are not applicable to the evaluation of the model in this study due to the characteristics of the AMT prescription data. We take Intersection over Union (IoU) [50] as the main evaluation criterion of the model.

3.1.1 Database construction

The text information of AMT collected from books is preprocessed to extract the symptoms used to judge diseases, and then these symptoms are standardized and unified. On this basis, we build a database of symptoms and corresponding acupoint prescriptions. Figure 3.1 shows the construction process of disease cases in the database through "chronic fatigue syndrome with stagnation of liver-Qi" as an example. The text in the green box is the description text of the disease, and the words marked with green underline are the corresponding symptoms. The words marked in the red box are used in TCM to reflect one cause of the sample case, and the circled numbers, e.g. 559, show the corresponding symptoms. The text in the blue box is the acupoint prescription of the disease case, and the words marked with blue underline are the corresponding acupoints which are represented by acupoint numbers. One of the symptom names in the figure, "rib-side and abdominal distention and pain", is expressed in a word in TCM, but it actually consists of four different symptoms. Therefore, the symptom name is divided into four symptom names, which is numbered as (148,149,158,159) in the database. In addition, the word with the meaning "or" often appears in the text describing the case to indicate that the same disease

may have multiple symptoms, which may or may not occur at the same time. For such case, we record it in the database as different cases for the same disease name. In this case, although the symptom combination of the disease is not completely the same, the corresponding acupoint prescriptions are the same. In this way, we build our database of symptoms and corresponding acupoint prescriptions.

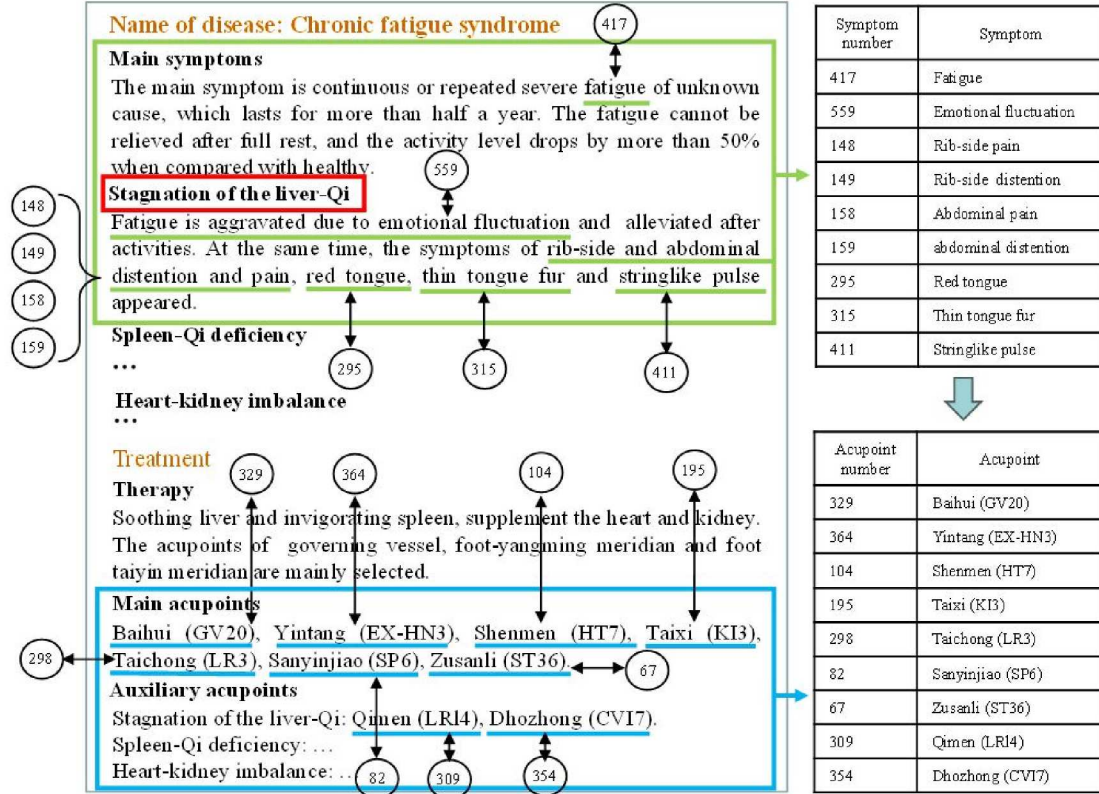


Figure 3.1: Take “chronic fatigue syndrome” as an example to illustrate the construction process of the database. Here the text of this figure is a translation of the Chinese text on page 845 of reference [2].

Our database includes the data shown in Table 3.1 and Table 3.2. Table 3.1 shows symptoms and acupoints with their related numbers. Table 3.2 shows disease case number, name of disease case, corresponding symptoms represented by symptom number and corresponding acupoints represented by acupoint number. In Table 3.2, the symptoms and acupoints are represented by their numbers according to Table 3.1 and the orders of symptom numbers and acupoint numbers are respectively the same as the symptoms and acupoints appear in texts of AMT prescription. One of the key factors in applying text data of books and clinical data to actual AMT through machine learning is symptom. Symptoms are the bridge between patient’s

condition and text data. The patient’s condition can be accurately described by giving the combination of symptoms. At the same time, these symptoms can be diagnosed through various medical technologies. Different from traditional AMT, our method requires the presence or absence of various symptoms to accurately describe the patient’s condition, so our naming method for symptoms is not exactly the same as that of TCM. Luo [51] gave the classification basis of 399 TCM symptom names by analyzing the attributes of symptoms. We plan to use these 399 symptom names as the symptom names in our database. However, 399 symptoms are not enough to accurately describe some diseases. Hence we have expanded the symptom names to 734 up to now. As to the names of acupoints, 409 acupoints certified by WHO and Ashi acupoint, totally 410 acupoint names are included in our database. Currently we have collected 3000 diseases cases of acupoint prescriptions from books [2, 52, 53]. We choose these three sources as our data materials because they are currently the only ones we have managed to compile, and we prioritized them due to the credibility of textbooks. In fact, additional high-quality data sources are essential before proceeding with further clinical trials.

Table 3.1: Database-1.

Symptom number	Symptom	Acupoint number	Acupoint
1	Irritable	1	LU1
2	Feel depressed	2	LU2
3	Oversensitive	3	LU3
4	Sentimental	4	LU4
5	Like to be quiet and lazy	5	LU5
6	Phobia	6	LU6
7	Hyperarousal	7	LU7
...	...	...	...
734	sticky mouth	410	Ashi

3.1.2 Algorithm selection

We regard the problem of acupoint selection as the problem of determining which acupoints are used for a series of symptoms. In machine learning, this kind of problem can be classified as multi-label classification problem and belong to supervised learning. The combination of symptom numbers is used as features of the model input, while the combination of acupoint numbers is used as labels of the model output. Meanwhile, it is obvious that the data type of this study is numeric. Based

Table 3.2: Database-2.

Disease case number	Name of disease case	Symptom numbers	Acupoint numbers
1	Cerebral apoplexy (syndrome of middle viscera closure)	400, 123, 38, 401, 396, 225, 402, 374, 337, 218, 403	225, 335, 11, 228, 106, 12, 229, 107, 77, 296, 193, 76, 295, 192
2	Cerebral apoplexy (syndrome of middle viscera closure) + upper extremity insufficiency	400, 123, 38, 401, 396, 225, 402, 374, 337, 218, 403, 206, 419	225, 335, 11, 228, 106, 12, 229, 107, 77, 296, 193, 76, 5, 15, 395, 233
3	Cerebral apoplexy (syndrome of middle viscera closure) + Shoulder arthralgia	400, 123, 38, 401, 396, 225, 402, 374, 337, 218, 403, 454, 456	225, 335, 11, 228, 106, 12, 229, 107, 77, 296, 193, 76, 295, 192, 28, 26, 115, 121, 120, 410
...	...	...	...
3000	Sequela of encephalitis	20, 505, 297, 315, 312, 411, 192	324, 360, 323, 322, 5, 225, 15, 109, 285, 67, 298

on the above observations, this study considers 11 different algorithms as potential candidates. We divide these 11 algorithms into the following 4 categories.

- Traditional machine learning algorithms: Traditional machine learning algorithms can be used to solve multi-label classification problems by employing problem transformation techniques, including Binary Relevance, Classifier Chains, and Label Powerset [13]. We used 8 classifiers for multi-label classification provided by Scikit-learn, a library in python that provides many unsupervised and supervised learning algorithms, which includes Naive bayes, DecisionTree classifier, ExtraTree classifier, ExtraTrees classifier, Kneighbors classifier, RandomForest classifier, Ridge classifierCV, Neural network. At the same time, each classifier is experimented with the three problem transformation techniques.

- FNN: FNN is a unidirectional multilayer network. Information is transmitted from the input layer to one direction layer by layer until the end of the output layer. FNN is composed of three parts, namely input layer, hidden layer and output layer. The hidden layer can be one layer or multiple layers. FNN can handle more complex nonlinear relationships in data and has stronger generalization than traditional machine algorithms. We hope that these advantages can achieve better results than traditional machine learning algorithms.

- TextCNN: CNN is usually used in the direction of computer vision (CV). Yoon [48] made some modifications to the input layer of CNN and proposed a text classification model TextCNN. The core of TextCNN is to convert words into word vectors through word2vec, and then turns sentences into matrices composed of word vectors, which are calculated through CNN. We hope to enhance the learning of the feature parts of the model through the use of CNN, thereby improving the effectiveness of the model.

- Seq2seq with attention: Seq2seq is an algorithm of encoding-decoding structure, originally proposed by Google, mainly for solving sequence-to sequence tasks such as machine translation and text summarization. However, this approach has also been applied to multi-label classification problems, with previous studies [54, 55, 56] demonstrating the effectiveness of encoder-decoder models and achieving promising results. Seq2seq has two modules, namely encoder and decoder. The encoder encodes the input data and the decoder decodes the encoded data. Simple RNN, GRU, LSTM [57], etc. can be used inside the encoder and decoder. The addition of Attention mechanism (hereinafter referred to as attention) can greatly improve the performance of Seq2seq. The fundamental principle of "Attention" lies in simulating human information processing by assigning different weights to different parts of the

data, thereby selectively focusing on key information. This mechanism calculates the relevance weights of input parts, enabling the model to concentrate on crucial information and ignore irrelevant details, thereby improving the accuracy and efficiency of handling complex tasks [39].

3.1.3 Data preprocessing

Different algorithms have different requirements for data representation, so we have different preprocessing for data when using different algorithms. For the 8 classifiers of traditional machine learning algorithms and FNN, the data can be preprocessed in the same way. For TextCNN and Seq2seq with attention, we use different data preprocessing methods separately.

For the 8 classifiers of traditional machine learning algorithms and FNN, numerical data can be directly used as features and labels of machine learning. Therefore, we encode the symptom of each case as a vector with a length of 734, the total number of symptoms in our database. Each element of the vector corresponds to a symptom name in the database. We encode the acupoints prescription of each case as a vector with a length of 410. Each element of the vector corresponds to the acupoint name in the database. As shown in Figure 3.2, the left part of the table represents symptoms, “0” means there is no such symptom in the disease case, and “1” means there is such symptom in the disease case. The right part of the table shows the acupoints in the acupoints prescription corresponding to the symptoms on the left part of the table. “0” means that the acupoint is not stimulated, and “1” means that the acupoint is stimulated.

For TextCNN, the core of this algorithm is to convert words into word vectors through word2vec, and then turns sentences into matrices composed of word vectors, which are calculated through CNN. Therefore, if symptoms are coded as grid data, the information of the data can be obtained through CNN. Since the number of symptoms in the database is 734 at present, which is not very large, and there is no obvious relationship between the symptoms, we do not need to use the word2vec used in TextCNN. As shown in Figure 3.3, our approach is to expand the symptom vector of 734 into a vector of 756 in length with “0”, and then transform the vector into a matrix of (27, 28). In this way, both one-dimensional convolution kernel or two-dimensional convolution kernel can be used for operation.

For Seq2seq with attention, the embedding layer (which is already included in its structure) handles the conversion of the discrete inputs into a continuous vector representation, thus obviating the need to manually convert the symptom numbers

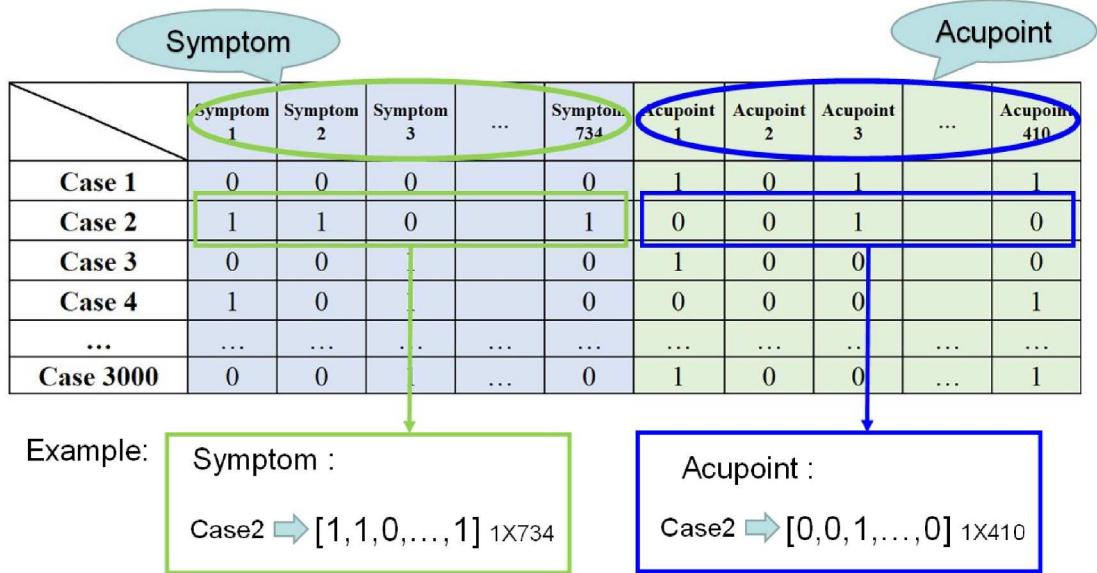


Figure 3.2: Coding method of the traditional machine learning algorithms and FNN.

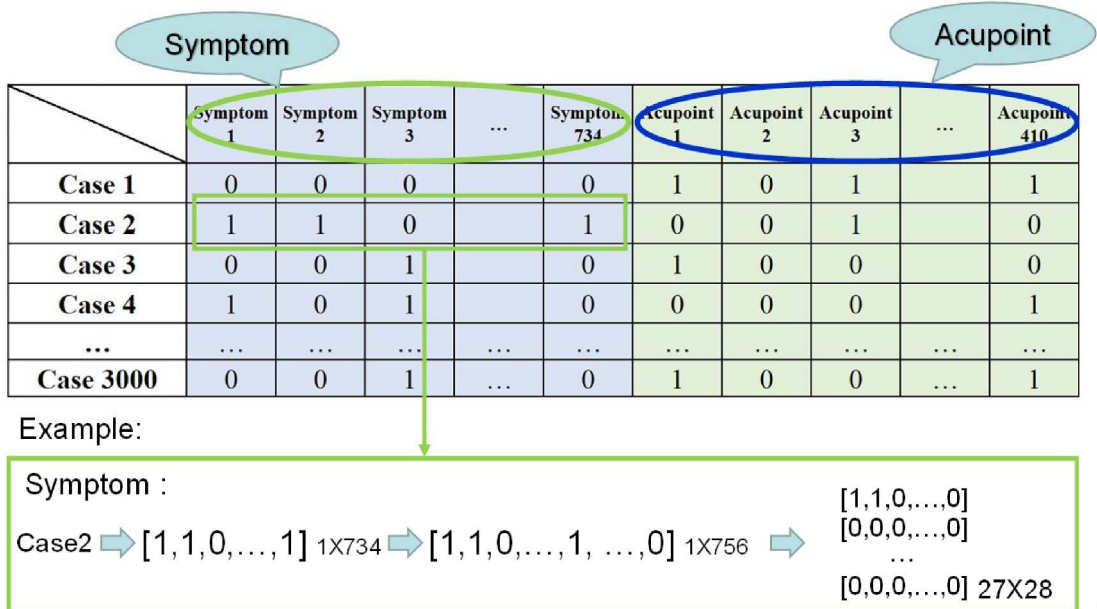


Figure 3.3: Feature encoding of TextCNN.

and acupoint numbers into vectors. Therefore, the symptom numbers and acupoint numbers in the disease cases shown in Table 3.2 can be directly used as input features and corresponding labels for training the model.

3.1.4 Evaluation Metrics

There are 4 possible situations for the results of classification problems, as shown in Table 3.3. In multi-label classification problems, the indicators commonly used to evaluate models are accuracy, precision, recall, for which the calculation formulas are shown in Eq. (3.1)–(3.3) [13]. Accuracy measures the overall correctness of the model’s predictions. Precision measures the proportion of correctly predicted positive labels among all the predicted positive labels. Recall measures the proportion of correctly predicted positive labels among all the actual positive labels. However, the above three evaluation criteria are not applicable to the evaluation criteria of the model in this study because of the characteristics of AMT prescription data.

IoU [50], also known as the Jaccard Index, is a metric utilized to compare the similarities and differences between finite sample sets. A higher IoU value signifies a higher degree of similarity between samples. As an evaluation metric, IoU not only comprehensively assesses the model’s accuracy in pinpointing effective acupoints but also effectively penalizes the incorrect prediction of non-effective acupoints as effective, thereby effectively mitigating the challenges caused by imbalanced labels. The calculation formula for IoU is shown in Eq. (3.4). Table 3.4 shows the results of different evaluation indicators in several possible situations for predicting acupoint prescriptions with two acupoints. When the prediction result is a wrong acupoint, the accuracy is still as high as 99.3%, which is obviously not suitable for this problem. It can be seen from Table 3.4 that IoU is even more strict than other evaluation criteria, such as accuracy, precision and recall. Hence, IoU is taken as the main evaluation criterion of the model in this study.

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}} \quad (3.1)$$

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}} \quad (3.2)$$

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}} \quad (3.3)$$

$$\text{IoU} = \frac{\text{TP}}{\text{TP} + \text{FP} + \text{FN}} \quad (3.4)$$

Table 3.3: Four possible situations for the results of classification

No.	Name	Description
1	True Positive (TP)	Number of positive classes predicted as positive classes ($1 \neq 1$)
2	True Negative (TN)	Number of negative classes predicted as negative classes ($0 \neq 0$)
3	False Positive (FP)	Number of negative classes predicted as positive classes ($0 \neq 1$)
4	False Negative (FN)	Number of positive classes predicted as negative classes ($1 \neq 0$)

Table 3.4: The evaluation results of an acupoint prescription with two acupoints using different evaluation criteria.

Prediction results		Accuracy	Precision	Recall	IoU
A wrong acupoint	TP=0 TN=407 FP=1 FN=2	99.3%	0%	0%	0%
Two wrong acupoints	TP=0 TN=406 FP=2 FN=2	99.0%	0%	0%	0%
A correct acupoint and a wrong acupoint	TP=1 TN=407 FP=1 FN=1	99.5%	50%	50%	33.3%
A correct acupoint and two wrong acupoints	TP=1 TN=406 FP=2 FN=1	99.3%	33.3%	50%	25%
Two correct acupoints and a wrong acupoint	TP=2 TN=407 FP=1 FN=0	99.8%	66.7%	100%	66.7%
Two correct acupoints	TP=2 TN=408 FP=0 FN=0	100%	100%	100%	100%

3.2 Experiments

In this section, we report the experimental results of 11 machine learning algorithms on our database and identify the best performing algorithm Seq2seq with attention. Additionally, we analyze the impact of factors such as the presence or absence of attention and the order of data in the Seq2seq with attention algorithm on the results.

3.2.1 Comparative experiments

In order to reduce the uneven distribution of training data and test data and more objectively evaluate the effects of the trained model, we use 5-fold cross validation to verify the model, and conducts random processing before data segmentation. We take 90% of the data in the database are used for 5-fold cross validation to select the optimal algorithm, and 10% of the data as test data to evaluate the performance of the final model.

Table 3.5 shows the average results of 5-fold cross validation experiments for the 11 machine learning algorithms, obtained by adjusting their respective hyperparameters to achieve optimal performance. For Seq2Seq with attention, we set vocab_size to 735, wordvec_size to 128, hidden_size to 256, and batch_size to 64. Other hyperparameter settings are consistent with those documented in reference [39]. It can be seen from the results that the IoU of Seq2seq with attention algorithm is the highest, that is, 95.72%, while the IoUs of all other algorithms are less than 41.62%.

Table 3.5: Experimental results of different machine learning algorithms.

Evaluating indicator	Naive baves	Decision-Tree classifier	Extra-Tree classifier	Extra-Trees classifier	Kneigh-bors classifier	Random-Forest classifier	Ridge classifier-CV	Neural network	FNN	TextCNN	Seq2seq with attention
Accuracy	84.00%	80.22%	67.11%	82.66%	73.33%	82.88%	78.90%	84.44%	99.82%	99.81%	99.91%
Precision	33.24%	36.06%	31.53%	36.29%	30.12%	33.10%	33.72%	33.61%	49.32%	43.50%	97.23%
Recall	32.66%	35.12%	31.3%	32.11%	23.78%	32.70%	29.74%	33.00%	49.08%	39.70%	96.83%
IoU	30.13%	33.19%	27.84%	31.77%	23.11%	29.78%	28.62%	30.17%	41.62%	38.80%	95.72%

Table 3.6 provides a detailed breakdown of the results and average values obtained from 5-fold cross validation specifically for Seq2seq with attention. We retain the best model in the training process as trained model, and evaluate the effectiveness of the trained model through test data. We input the symptom numbers of 300 disease cases not participating in the training model into the trained model, and then the trained model gives the corresponding acupoint numbers. IoU of the test data is 95.33%.

Table 3.6: Experimental results of the 5-fold cross validation of Seq2seq with attention.

Evaluating indicator	First	Second	Third	Fourth	Fifth	Average value
Accuracy	99.82%	99.96%	99.98%	99.97%	99.83%	99.91%
Precision	94.17%	98.69%	99.22%	99.02%	95.04%	97.23%
Recall	93.03%	98.83%	99.29%	98.81%	94.20%	96.83%
IoU	90.76%	98.30%	98.89%	98.50%	92.14%	95.72%

3.2.2 Further Experimental Analysis of Seq2seq with attention

To investigate the impact of the order of symptoms and acupoints numbers on the experimental results, we conduct the experiment by arranging the symptoms and acupoints numbers in ascending and descending order in the original data. Table 3.7 shows the results of 5-fold cross validation for both ascending and descending orders. Comparison of Table 3.6 and Table 3.7 reveals that label ordering has little effect on experimental outcomes, which is consistent with the findings reported in [54]. Specifically, the results suggest that if models are trained with appropriate regularization techniques, the order of symptoms and acupoints numbers is not a major factor influencing final performance.

Table 3.7: Experimental results of the 5-fold cross validation of Seq2seq with attention on ascending and descending data.

Order of data	Evaluating indicator	First	Second	Third	Fourth	Fifth	Average value
Ascending data	Accuracy	99.70%	99.93%	99.93%	99.96%	99.73%	99.85%
	Precision	90.94%	98.06%	98.66%	99.17%	92.22%	95.81%
	Recall	90.94%	97.94%	98.26%	98.70%	91.78%	95.52%
	IoU	87.98%	97.20%	97.79%	98.39%	89.38%	94.15%
Descending data	Accuracy	99.68%	99.82%	99.98%	99.97%	99.83%	99.85%
	Precision	90.70%	95.46%	99.22%	99.02%	95.04%	95.89%
	Recall	89.91%	95.28%	99.29%	98.81%	94.20%	95.50%
	IoU	86.73%	93.36%	98.89%	98.50%	92.14%	93.92%

In order to analyze the impact of attention on model performance, we conduct ablation experiment. Table 3.8 shows the 5-fold cross validation results of Seq2seq without attention. Comparison of Table Table 3.6 and Table 3.8 reveals that the IoU of Seq2seq without attention is significantly lower than Seq2seq with attention. In addition, the IoU of the third and fourth folds are higher than 95%, while the IoU of the first, second, and fifth folds are all below 70%, indicating that the model’s

generalization ability without attention is poor. The experimental results show that the attention not only improves the IoU of the model, but also enhances the robustness of the model.

Table 3.8: Experimental results of the 5-fold cross validation of Seq2seq without attention.

Evaluating indicator	First	Second	Third	Fourth	Fifth	Average value
Accuracy	99.06%	99.12%	99.93%	99.95%	99.07%	99.42%
Precision	72.77%	75.34%	98.13%	98.45%	73.56%	83.65%
Recall	76.56%	80.21%	98.30%	98.54%	78.06%	86.34%
IoU	63.54%	67.34%	97.18%	97.93%	64.58%	78.11%

To evaluate the generalization and stability of the algorithm under increased data volume, we conducted experiments on seq2seq with attention using augmented data. For the symptoms of a case, some of them are not “and” relationship, but “or” relationship. Therefore, one or two secondary symptoms can be removed for cases with more than 10 symptoms. In this way, the amount of data can be increased and the influence of the main symptoms in the model can be enhanced. We expanded 3000 cases to 6000 in this way. Table 3.9 provides a detailed breakdown of the results and average values obtained from 5-fold cross validation specifically for Seq2seq with attention using 6000 augmented data. Comparing Table 3.6 and Table 3.9 reveals a slight improvement in the model’s IoU for the dataset of 6000, indicating that the algorithm has good scalability. It’s important to note that AMT is a medical practice, which demands exceptionally high data quality standards. The augmented data in this paper is solely for testing the algorithm’s scalability, not for training the final model.

Table 3.9: Experimental results of the 5-fold cross validation of Seq2seq with attention using 6000 augmented data.

Evaluating indicator	First	Second	Third	Fourth	Fifth	Average value
Accuracy	99.85%	99.97%	99.98%	99.96%	99.86%	99.92%
Precision	95.60%	99.10%	99.23%	98.98%	96.04%	97.79%
Recall	95.75%	99.30%	99.26%	98.72%	96.21%	97.85%
IoU	93.93%	98.77%	98.92%	98.37%	94.84%	96.97%

3.2.3 Case Study

We showcase our model’s practical performance through an example. We input a test case’s symptom combination into our model, and it provides an acupoint prescription, as illustrated in Figure 3.4. The model’s prescription corresponds to the acupoint prescription described in reference [53] for Yin-Yang deficiency hypertension. Despite this particular combination of symptoms in the disease case not being included in the model’s training data, the model provides an accurate acupoint prescription, which shows that our model has a good generalization ability.

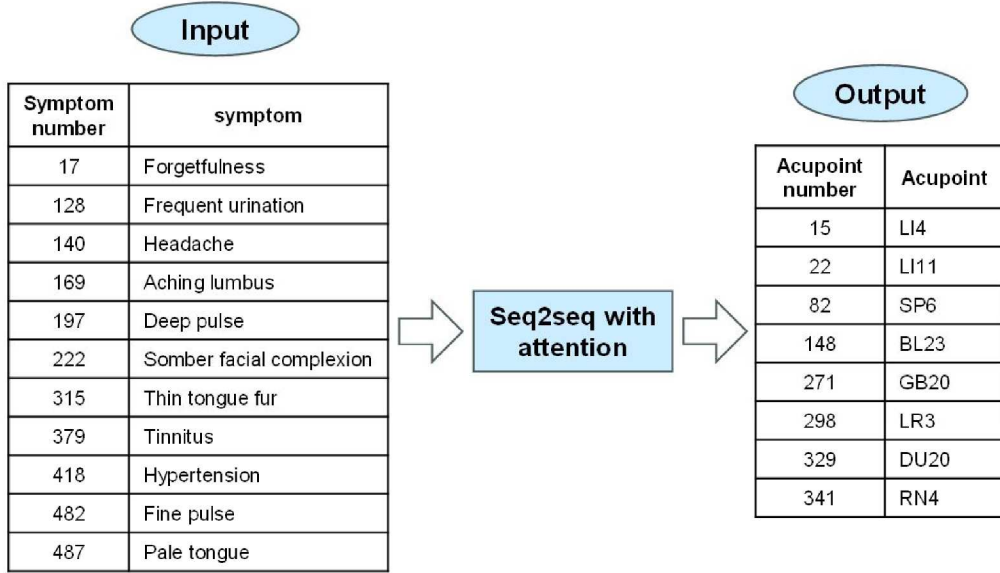


Figure 3.4: Exemplary Acupoint Prescription by Seq2seq with Attention.

3.2.4 Discussion

For traditional machine learning algorithms, we attribute their poor performance to the following three reasons. (1) Traditional machine learning algorithms, such as Decision Trees, Random Forests, and Naive Bayes methods, become ineffective in capturing higher order correlations as they can only capture first or second order correlations [56]; (2) For multi-label classification problems with excessive numbers of labels (e.g., 410 labels in this study), transformation techniques that include Binary Relevance, Classifier Chains, and Label Powerset are ineffective; (3) The imbalance of labels in AMT data is also one of the important reasons for the poor performance of traditional machine learning algorithms. It is difficult to ensure that the positive and negative sample sizes of each label in the training data are consistent in AMT

data. For FNN and TextCNN, although higher-order correlations in features can be captured, their generalization ability is limited due to ignoring the correlations between labels.

For Seq2seq with attention, the superior performance in deciding acupoints compared to other algorithms can be attributed to 3 key advantages. Firstly, Seq2seq can capture higher order relationships among features. Secondly, due to the encoding-decoding structure of Seq2seq, it can accurately capture and predict the relationships between multiple labels while learning high-order correlations of features, thereby enhancing the model's generalization ability and overall performance. Furthermore, the addition of attention further improves the performance and generalization ability of the model.

Based on the experimental results and the analysis mentioned above, Seq2seq with attention is sufficiently applicable to determine acupoints.

3.3 Conclusion

This chapter presented a series of exploratory studies that established a systematic framework for applying machine learning to AMT acupoint prescription. Through comparative experiments, we benchmarked a range of models, finding that the Seq2seq architecture with an attention mechanism was particularly effective at capturing the sequential dependencies in treatment protocols. Despite its predictive strengths, however, its clinical utility was hampered by poor interpretability and sensitivity to data variations. These findings crystallized the core requirements for a clinically viable model: it must not only predict accurately but also reflect the hierarchical reasoning of AMT and offer transparent outputs. The hierarchical and interpretable model proposed in Chapter 4 is designed specifically to meet these requirements.

Chapter 4

Hierarchical and Interpretable Machine Learning Model for Acupoints Determination

Building on the insights gained from the exploratory studies in Chapter 3, this chapter introduces a novel hierarchical and interpretable machine learning model for acupoint determination in AMT. While earlier approaches demonstrated the feasibility of applying machine learning to AMT, they also exposed several limitations, such as difficulties in handling the hierarchical nature of acupoint selection and the lack of interpretability required for clinical applications. To address these challenges, we propose a model that integrates hierarchical classification with mechanisms designed to enhance interpretability, thereby aligning computational predictions more closely with the principles of AMT and the practical needs of clinicians. The chapter is organized as follows. Section 4.1 details the materials and methods, including data analysis and database construction, the design of the hierarchical and interpretable model, and the evaluation metrics employed. Section 4.2 presents the experimental results, demonstrating the effectiveness of the proposed approach in capturing the structured relationships among acupoints while providing interpretable outputs. Finally, Section 4.3 concludes the chapter by summarizing the contributions of the proposed method and setting the stage for its integration into a practical decision-support system discussed in Chapter 5.

4.1 Materials and Methods

The process of selecting acupoints can be regarded as establishing the corresponding relationship between the combination of a series of symptoms and the combination of acupoints in the acupoint prescriptions, rather than the corresponding relationship between a single symptom and a single acupoint. In machine learning, this kind of problem can be classified as multi-label classification problem. The combination of symptom numbers is used as features of the model input, while the combination of acupoint numbers is used as labels of the model output. To achieve the prediction of acupoint prescriptions from symptoms using machine learning models, various approaches are evaluated, including traditional machine learning methods (such as Naive bayes [58], Decision trees [59]), ensemble learning technique (such as Extra tree [33]), and deep learning models (such as FNN [47], TextCNN [48], TextGCN [60], ML-GCN [61], Seq2seq with attention [49], MacBERT [62]). However, these methods do not fully leverage the hierarchical structure inherent in AMT data. This structure plays an important role in AMT practices. To overcome these shortcomings, this study adopts hierarchical learning techniques to better utilize the structural characteristics of AMT data. This section introduces the database construction, presents the hierarchical and interpretable machine learning model, and finally describes the evaluation metrics.

4.1.1 Data analysis and database construction

We take the AMT prescription data for multiple sclerosis (from Reference [2]) as a case study, analyzing it from a machine learning perspective. Through this case, we illustrate the process of database construction and methods for data augmentation.

4.1.1.1 Data sources and analysis

For thousands of years, extensive clinical experience in AMT has been systematically documented in written records. These textual records of prescription information constitute the primary scope of our data sources. In this study, we have currently collected 5,000 acupoint prescription cases from books [2, 52, 53]. These classical TCM books and textbooks were selected as the primary data sources for our database for two main reasons: (1) their standardized terminology for diseases and symptoms provides a consistent foundation for future dataset expansion; and (2) the case descriptions generally follow a clear structure of “symptom—syndrome—acupoint prescription”, which facilitates data extraction and organization.

We use multiple sclerosis, as described in reference [2], as an example to illustrate the data collection process. Figure 4.1 presents the classification of multiple sclerosis and its corresponding acupoint prescriptions. The left part of the figure provides an overview of the primary symptoms of multiple sclerosis, which are further categorized into four syndromes based on symptom differences: lung heat damaging liquid, spreading damp-heat pattern, spleen-stomach vacuity, and liver-kidney depletion. The right part of the figure presents the acupoint prescriptions. First, two primary sets of acupoints suitable for multiple sclerosis patients are identified based on the affected body regions, namely the upper limb and lower limb. Subsequently, specific acupoints are added based on the respective syndromes to address individual conditions, enabling more targeted and precise treatment.

Name of disease: Multiple Sclerosis	
【Key points of syndrome differentiation】	
Main symptoms: Loss of use of the limbs. Severe cases may present with muscle atrophy or deformities.	
Accompanying Symptoms:	
1. Lung heat damaging liquid: Symptoms include fever, profuse sweating, limb paresis, lack of strength, agitation, dry mouth, short urine, light yellow urine, pale red tongue, yellow tongue fur, fine and rapid pulse.	
2. Spreading damp-heat pattern: Symptoms include flaccid and weak muscles, particularly in the lower limbs, which feel heavy. There may be slight swelling or numbness, or a sensation of damp-heat. Additional symptoms include reddish urine, a red tongue with a yellow greasy coating, and a slippery rapid pulse.	
3. Spleen-stomach vacuity: Symptoms include prolonged muscle weakness and lack of strength, poor appetite, abdominal distension, pale complexion, mental fatigue, a thin or tooth-marked tongue, greasy coating, and a weak thready pulse.	
4. Liver-kidney depletion: Symptoms include a gradual onset of weakness in the lower limbs, soft and weak muscles in the lower back, and soreness in the waist. Patients may have difficulty standing, experience dizziness or tinnitus, or even progress to severe disability in walking. Muscle atrophy may occur in the hips and legs. Additional symptoms include a red tongue, little moss, and a deep, thready pulse.	
【Treatment】	
Primary Acupoints	
Upper limb: Jianyu (LI15), Quchi (LI11), Shousanli (LI10), Hegu (LI4), Waiguan (SJ5), Jiaji (EX-B2)	
Lower limb: Biguan (ST31), Futu (ST32), Zusanli (ST36), Fenglong (ST40), Fengshi (GB31), Yanglingquan (GB34), Sanyinjiao (SP6), Jiaji (EX-B2)	
Supplementary Acupoints	
1. For lung heat damaging liquid: Add Yuji (LU10), Chize (LU5), Feiyu (BL13)	
2. For spreading damp-heat pattern: Add Yinlingquan (SP9) and Zhongji (RN3)	
3. For spleen-stomach vacuity: Add Pishu (BL20), Weishu (BL21), Zhangmen (LR13), Zhongwan (RN12)	
4. For liver-kidney depletion: Add Ganshu (BL18), Shenshu (BL23), Taichong (LR3), Taixi (KI3)	

Figure 4.1: Take “multiple sclerosis” as an example to illustrate the construction process of the database. Here the text of this figure is a translation of the Chinese text on page 863 of reference [2].

By observing the prescription for multiple sclerosis, it can be seen that acupoint prescription consists of the following key components: the disease name, primary symptoms, different syndrome classifications, and the acupoint prescriptions corresponding to each syndrome. Furthermore, the following pattern can be summarized from the analysis of extensive AMT data: 1. For the same disease, patients generally share consistent or similar primary symptoms. However, due to differences in etiology, the disease can be further classified into multiple syndromes; 2. The core acupoints for treating the same disease are typically identical or similar. However, based on specific syndrome classifications or particular symptoms, additional acupoints are selected or certain acupoints are adjusted to achieve specific treatment.

Based on the above observations, we use Figure 4.2 to illustrate the characteristics of AMT data. A disease case is composed of multiple symptoms, and different combinations of these symptoms can be categorized under a specific disease name. The same disease name may be further subdivided into multiple subcategories based on variations in syndromes or certain specific symptoms. Here, the subcategories are defined by us to better capture the nuances in acupoint prescriptions. A subcategory is a classification used to distinguish differences in treatment plans caused by variations in syndromes or specific symptoms. Each Subcategory is defined by a combination of the disease name, syndrome type, and symptom names that influence the treatment prescription. Cases within the same subcategory share the same treatment prescription. In terms of classification tasks, a combination of symptoms corresponds to a single disease name and its associated subcategory, making the classification of disease names and subcategories a multi-class problem. In contrast, acupoint prescriptions, which involve selecting multiple acupoints from a set of several hundred acupoints, can be approached as a multi-label classification task.

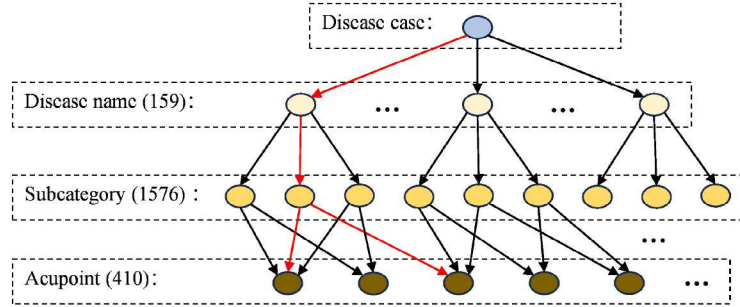


Figure 4.2: Hierarchical label structure for disease cases in acupoint prescriptions. The numbers in the figure represent the statistical information of our database.

4.1.1.2 Database construction

Due to differences in how identical symptoms are described across various data sources, we adopted a manual way to ensure the quality of the database. Specifically, we followed the symptom naming guidelines proposed in reference [63] as a foundation. Unlike traditional AMT, we categorized additional terms, including body parts, symptom severity, and gender, that are closely related to the description of the condition into “symptoms”, resulting in a total of 770 “symptoms” in our dataset. As to the names of acupoints, our database includes 410 acupoint names, consisting of 409 acupoints certified by WHO [18] and the Ashi acupoint. Ashi points refer to areas that lack a specific name or fixed position but are identified by tenderness or

other reactions when stimulated [2]. To date, our dataset has collected 5,000 cases, each defined as a distinct set of symptoms and its corresponding acupoint prescription. While some examples reflect actual clinical experience, most were formulated according to textbook guidance. In our construction, any variation in the symptom combination or acupoint selection is treated as a separate case. As shown in Table 1, each case consists of the following components: disease case number, disease name, subcategory, symptoms, and acupoints. To meet the requirements of machine learning, we assigned numerical IDs to each disease name, subcategory, symptom, and acupoint (as shown in Figure 4.3), and digitized all data accordingly. The results are presented in Table 4.2, which corresponds to Table 4.1 with all original terms replaced by their respective IDs defined in Figure 4.3. The acupoint order in Table 4.2 follows the original prescription sequence in the data. However, since the acupoints are treated as labels in a multi-label classification task, their order does not affect the model’s learning or prediction results.

Each disease case in the database is labeled across three hierarchical levels: The first level is the disease name layer, which includes 159 categories; The second level is the subcategory layer, which consists of 1,576 categories; The third level is the acupoint layer, comprising 410 categories, corresponding to 410 acupoints names. It should be noted that the statistical data presented here reflects the current state of our database, rather than the categorization practices in TCM field. In addition, since manual annotation is time-consuming, we applied data augmentation to expand the dataset to 10,138 samples for large-scale performance evaluation.

4.1.1.3 Data expansion

Among the 5,000 original samples, there are 1,576 subcategories, with an average of only 3.2 samples per subcategory. In addition, due to limitations in data sources, some subcategories have a large number of samples, while many have only one sample. This severe data imbalance poses challenges for model training. Thus, in the absence of rapidly expanding metadata, data augmentation becomes essential to address this issue.

Our data augmentation method is based on the following principle: In acupoint prescriptions, each disease case typically includes multiple symptoms, but these symptoms vary in their impact on diagnosis and treatment plans. Some symptoms do not affect the diagnosis of a disease or the corresponding treatment, while others may significantly influence the prescription. Based on this, we categorized symptoms into

Disease case number	Disease name	Subcategory	Symptoms	Acupoints
1	apoplexy	apoplexy (organ strike block pattern)	dizzy, inhibited urination, difficult defecation, curled tongue, retracted genitals, skin color red, thick breath, phlegm rale in the throat, bad breath, fever, agitation	PC6, DU26, LU11, PC9, HT9, LJ1, SJ1, SP1, LR1, KJ1, ST45, GB44, BL67
2	apoplexy	apoplexy (organ strike block pattern): upper limb, hemiplegia	dizzy, inhibited urination, difficult defecation, curled tongue, retracted genitals, skin color red, thick breath, phlegm rale in the throat, bad breath, fever, agitation, loss of use of the limbs, upper limb	PC6, DU26, LU11, PC9, HT9, LJ1, SJ1, SP1, LR1, KJ1, ST45, GB44, BL67, LJ15, LU5, L14, EX-UE9, SJ5
3	apoplexy	apoplexy (organ strike block pattern), omarthralgia	dizzy, inhibited urination, difficult defecation, curled tongue, retracted genitals, skin color red, thick breath, phlegm rale in the throat, bad breath, fever, agitation, shoulders, shoulder pain and soreness	PC6, DU26, LU11, PC9, HT9, LJ1, SJ1, SP1, LR1, KJ1, ST45, GB44, BL67, LJ17, LJ15, S19, SJ15, SJ14, ASH1
5000	postencephalitis	postencephalitis	... aphasia, sequelae period, red or crimson tongue, thin fur, yellow tongue fur, stringlike pulse, rapid pulse	DU15, RN23, DU14, DU13, LU5, PC6, LJ4, SJ3, GB34, ST36, LR3

Table 4.1: Rotated disease case table.

Disease number	Disease name	Symptom number	Symptom	Acupoint number	Acupoint
1	Apoplexy	1	Irritable	1	LU1
2	Headache	2	Feel depressed	2	LU2
3	Dizziness	3	Oversensitive	3	LU3
...	...	...	...	...	...
159	Postencephalitis	770	Urinary retention	410	Ashi

Subcategory number	Subcategory
1	apoplexy (organ strike block pattern)
2	apoplexy (organ strike block pattern), upper limb, hemiplegia
3	apoplexy (organ strike block pattern), omarthralgia
...	...
1576	postencephalitis

Figure 4.3: Diseases, symptoms, acupoints, and subcategories with assigned IDs.

Original data			
Disease name	Subcategory	Symptoms	Acupoints
multiple sclerosis	multiple sclerosis (lung heat damaging liquid), upper limb	loss of use of the limbs, muscle atrophy, fever, profuse sweating, limb paresis, lack of strength, agitation, dry mouth, short urine, light yellow urine, pale red tongue, yellow tongue fur, fine pulse, rapid pulse, upper limb	LI15, LI11, LI10, LI4, SJ5, EX-B2, LU10, LU5, BL13

↓

Augmented data			
Disease name	Subcategory	Symptoms	Acupoints
multiple sclerosis	multiple sclerosis (lung heat damaging liquid), upper limb	loss of use of the limbs, muscle atrophy, fever, profuse sweating, limb paresis, lack of strength, agitation, dry mouth, short urine, light yellow urine, pale red tongue, yellow tongue fur, fine pulse, rapid pulse, upper limb	LI15, LI11, LI10, LI4, SJ5, EX-B2, LU10, LU5, BL13
multiple sclerosis	multiple sclerosis (lung heat damaging liquid), upper limb	loss of use of the limbs, muscle atrophy, fever, profuse sweating, limb paresis, lack of strength, agitation, dry mouth, short urine, light yellow urine, pale red tongue, yellow tongue fur, fine pulse, rapid pulse, upper limb	LI15, LI11, LI10, LI4, SJ5, EX-B2, LU10, LU5, BL13
multiple sclerosis	multiple sclerosis (lung heat damaging liquid), upper limb	loss of use of the limbs, muscle atrophy, fever, profuse sweating, limb paresis, lack of strength, agitation, dry mouth, short urine, light yellow urine, pale red tongue, yellow tongue fur, fine pulse, rapid pulse, upper limb	LI15, LI11, LI10, LI4, SJ5, EX-B2, LU10, LU5, BL13
multiple sclerosis	multiple sclerosis (lung heat damaging liquid), upper limb	loss of use of the limbs, muscle atrophy, fever, profuse sweating, limb paresis, lack of strength, agitation, dry mouth, short urine, light yellow urine, pale red tongue, yellow tongue fur, fine pulse, rapid pulse, upper limb	LI15, LI11, LI10, LI4, SJ5, EX-B2, LU10, LU5, BL13

Figure 4.4: Example of data augmentation: from original to augmented data.

Table 4.2: Numerical representation of database elements.

Disease case number	Disease name number	Subcategory number	Symptom numbers	Acupoint numbers
1	1	1	400,123,38,401,396,225,402,374,337,218,403	225,335,11,228,106,12,229,107,77,296,193,76,295,192
2	1	2	400,123,38,401,396,225,402,374,337,218,403,206,419	225,335,11,228,106,12,229,107,77,296,193,76,295,192,26,5,15,395,233
3	1	3	400,123,38,401,396,225,402,374,337,218,403,454,456	225,335,11,228,106,12,229,107,77,296,193,76,295,192,28,26,115,121,120,410
...	...	...	...	...
5000	159	1576	20,505,297,315,312,411,192	324,360,323,322,5,225,15,109,285,67,298

three types: primary symptoms (which specifically indicated in the raw data), special symptoms (such as the location of symptoms, severity of disease, etc.) that affect treatment plans, and less significant symptoms. In the process of data expansion, we retained the primary and special symptoms while randomly removing one or two less significant symptoms, and keeping the acupoint prescription unchanged, thus achieving the goal of data augmentation. Figure 4.4 illustrates the process of data augmentation using a case of multiple sclerosis as an example. In the table at the upper part of the figure, symptoms are categorized by font color for illustrative purposes, where green indicates primary symptoms, blue represents special symptoms, and black denotes less significant symptoms. To expand the dataset, one less significant symptom (such as “lack of strength”, “agitation”, “dry mouth”) is randomly removed from the original sample while retaining the original data. This process expands the dataset from one sample to four samples. We applied this approach to expand the database to 10,138 samples.

To evaluate the quality and distributional consistency of the augmented data, we employed t-Distributed Stochastic Neighbor Embedding (t-SNE), a non-linear dimensionality reduction technique designed to visualize high-dimensional datasets in a low-dimensional space [64]. As illustrated in Figure 4.5, where blue points represent original clinical samples and orange points denote generated ones, the visualization demonstrates that the original data forms distinct semantic clusters. The generated samples are distributed in close proximity to these clusters, filling the sparse regions within the data manifold without forming isolated outliers. This observation suggests that the generated data maintains distributional consistency with the original dataset, making it suitable for subsequent model training.

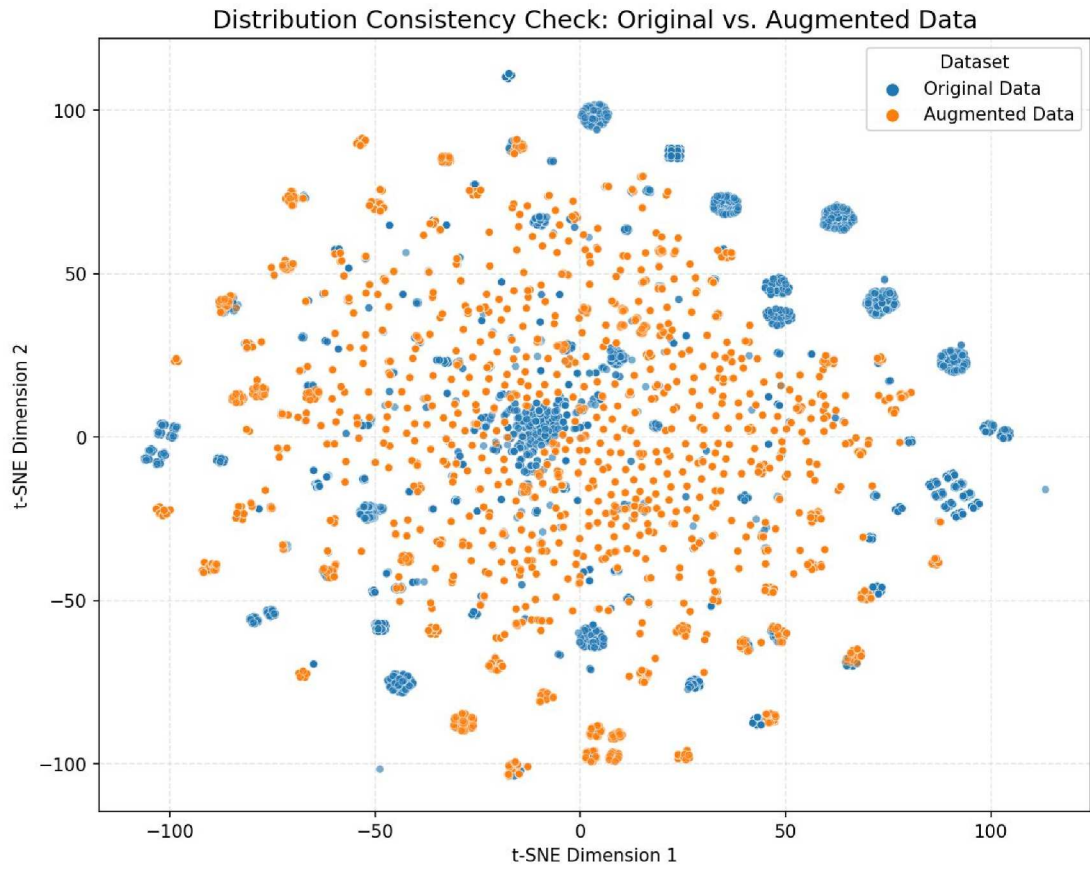


Figure 4.5: t-SNE visualization of feature distributions for original and augmented data.

4.1.2 Hierarchical and Interpretable Machine Learning Model

4.1.2.1 Overview of research framework

Figure 4.6 provides an overview of the proposed hierarchical and interpretable model framework. As shown in Figure 4.6, our model consists of three main components: the data preparation module, the hierarchical classification module, and the result interpretation module. The model takes the symptoms from the disease case D and the corresponding hierarchical category structure γ as input. Then, the hierarchical classification module captures the relationships between symptoms and each category in the hierarchy, models the dependencies across different levels, and predicts the hierarchical categories of the disease case. Finally, we employ LIME to visualize the prediction results at each level, to enhance the interpretability of the results.

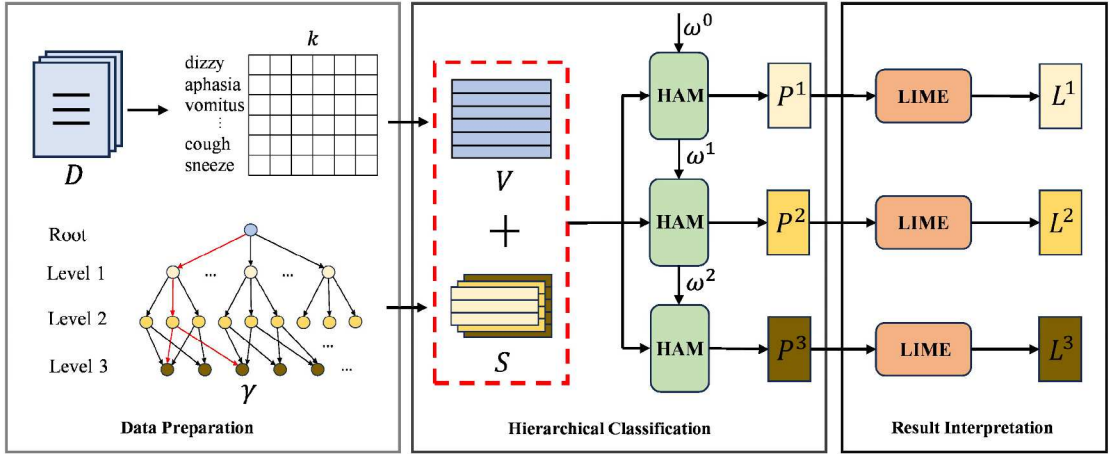


Figure 4.6: Framework of the hierarchical and interpretable machine learning model. D : Textual data of a disease case. k : The dimension of the embedding vector, defining the length of each symptom’s representation. γ : Hierarchical category structure corresponding to the disease cases. V : The complete semantic representation of a disease case’s symptoms. S : The vector matrix of the hierarchical category structure γ . HAM: the hierarchical attention-based memory unit. ω^{h-1} : the information of previous level h . P^i : The prediction result of the i -th layer. L^i : The interpretability results of LIME in the i -th layer.

4.1.2.2 Hierarchical classification module

The hierarchical classification module utilizes the HARNN model proposed in [15], an established method for hierarchical classification tasks. HARNN generates unified representations for both document texts and hierarchical category structures, models dependencies between hierarchical levels, and predicts categories in a top-down

manner. Unless otherwise stated, we follow the configurations outlined in the original paper.

After representing the disease case text $\mathbf{D}$ as a sequence of $\mathbf{N}$ symptoms $(\mathbf{w}_1; \mathbf{w}_2; \dots; \mathbf{w}_N)$, where each symptom is encoded using contextualized embeddings derived from the pre-trained BERT model (bert-base-Chinese) [62, 65]. Similarly, the hierarchical category structure γ is embedded into a matrix $\mathbf{S} = (\mathbf{S}^1; \mathbf{S}^2; \mathbf{S}^3)$ with randomly initialized embeddings for each hierarchical level. A recurrent architecture, as illustrated in Figure 4.6, is employed to model dependencies across different levels, processing the hierarchical structure in a top-down manner. At each level, the Hierarchical Attention-based Memory (HAM) unit functions as the primary mechanism. It captures associations between texts and categories, generates unified representations and predictions for each category level, and transfers hierarchical semantic representations to subsequent layers. The input to an HAM unit consists of the whole semantic representation of the disease cases $\mathbf{V}$, the representation of the current category level $\mathbf{S}^h$, and information from the previous level ω^{h-1} . All elements in ω^0 are initialized to 1.0, as level 0 represents the root of the hierarchical structure and does not contain any additional information. The hierarchical classification module generates predictions at each hierarchical level, represented as $\mathbf{P}^1$, $\mathbf{P}^2$, and $\mathbf{P}^3$.

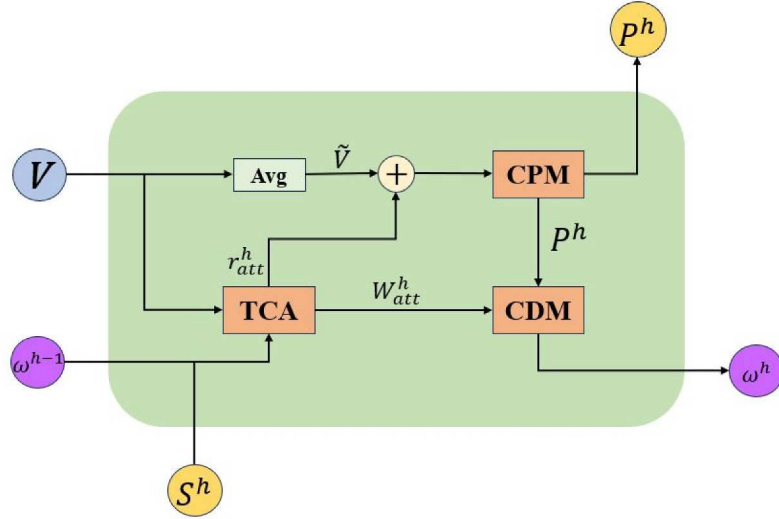


Figure 4.7: Hierarchical attention-based memory unit [15].

In the HARNN architecture, the Attention mechanism is implemented as the Text-Category Attention (TCA) module embedded within the Hierarchical Attention-based Memory (HAM) unit, as illustrated in Figure 4.7. This Attention mechanism

serves as a hierarchical attention mechanism that allocates attention weights according to hierarchical dependencies, thereby enabling the model to focus on semantic features that are most relevant to the classification task at each level. The core component of HARNN is the Hierarchical Attention-based Recurrent Layer (HARL), which models inter-level dependencies in a recursive, top-down manner. The overall Attention process is realized through the collaboration of three sub-modules within each HAM unit:

- **Text-Category Attention (TCA) — Precise Feature Extraction:** TCA captures associations between textual content and category semantics. Specifically, it incorporates hierarchical information from the previous level $\mathbf{!}^{h-1}$ as prior knowledge to filter the original text representation $\mathbf{V}$ via an element-wise product, referred to as contextual filtering. This operation allows the model to re-evaluate textual content under a level-specific contextual background. Subsequently, TCA computes the correlation between category embeddings $\mathbf{S}^h$ and text features to generate the attention matrix $\mathbf{W}_{\text{att}}^h$, enabling the model to emphasize words that contribute most to the current classification level.
- **Class Prediction Module (CPM) — Integrated Decision-making:** The local feature representation $\mathbf{r}_{\text{att}}^h$ refined by TCA is concatenated with the global semantic representation $\mathbf{V}$. Based on this combined representation, the CPM produces the predicted probability vector $\mathbf{P}^h$ for the h -th classification level, integrating both local attention-driven evidence and global contextual information.
- **Class Dependency Module (CDM) — Knowledge Inheritance:** The CDM aggregates the attention-informed predictions by weighting them with the probability vector $\mathbf{P}^h$ to construct an updated hierarchical information vector $\mathbf{!}^h$. This vector is subsequently passed to the next HAM unit, ensuring continuity of hierarchical reasoning, where decisions at deeper levels are constrained by higher-level predictions.

4.1.2.3 Result interpretation module

The result interpretation module in this study employs LIME, a model-agnostic and locally interpretable machine learning method. LIME explains the predictions of complex black-box models by approximating them with simple interpretable models, such as linear regression or decision trees. Its core mechanism involves generating

new samples by perturbing the input around a given instance and using the black-box model’s predictions on these perturbed samples to train a local interpretable model. This interpretable model provides a localized approximation of the black-box model’s behavior for the specific instance being analyzed.

LIME explains the predictions of a pre-trained black-box model, which in our case is HARNN. The process begins by selecting an instance from the test dataset. Perturbed samples are then generated around this instance, with weights assigned based on their similarity to the original instance. Using these weighted samples, a simple interpretable model is trained to locally approximate the decision behavior of the black-box model. Finally, corresponding explanations L^1 , L^2 , L^3 are generated for each of the predictions P^1 , P^2 , P^3 made by the Hierarchical classification module.

4.1.2.4 Evaluation metrics

For multi-label classification, common evaluation metrics include accuracy, precision, recall, and IoU (Intersection over Union, also known as the Jaccard Index) [13, 50]. Accuracy measures overall prediction correctness, precision evaluates the proportion of correct positive predictions, and recall assesses the proportion of actual positives correctly identified. IoU comprehensively evaluates the model’s accuracy in identifying effective acupoints while effectively penalizing incorrect predictions that misclassify non-effective acupoints as effective. This ensures a reliable assessment even when the distribution of effective and non-effective labels is imbalanced. Hence, IoU is taken as the main evaluation criterion of the model in this study. In addition, we incorporated Top-3 and Top-5 accuracy metrics, which evaluate whether the correct acupoint appears among the top three or top five predicted candidates, to provide a more comprehensive assessment of the model’s performance.

4.2 Experiments

In this section, we evaluate the model using both quantitative metrics and case studies. First, we present the overall model performance using metrics such as Accuracy, Precision, Recall, and IoU. Then, we introduce two case studies: the first demonstrates the interpretability of the model using LIME visualizations, and the second analyzes a misclassification case to explore the possible causes of error.

4.2.1 Model evaluation and performance comparison

We first present the performance of the hierarchical classification model on the original dataset of 5,000 samples for each classification layer. Next, we analyze the changes in performance trends across layers during the data augmentation process to investigate its impact on model performance. Finally, we compare the experimental results of different machine learning models using the augmented dataset to further validate the effectiveness of the proposed approach.

4.2.1.1 Hierarchical model results on original dataset

The hierarchical classification task in this study is divided into three levels: the first level predicts disease names, the second level predicts disease subcategories, and the third level predicts acupoints prescriptions. In this experiment, 5,000 original data were randomly divided into 80% for training and 20% for testing. Table 4.3 presents the prediction results for each level.

Table 4.3: Performance metrics for three hierarchical levels on 5,000 original data.

Level	Evaluation Metrics					
	Accuracy	Precision	Recall	IoU	Top-3	Top-5
Level-1	0.9996	0.9613	0.9690	0.9326	0.9860	0.9890
Level-2	0.9997	0.7267	0.7340	0.5752	0.7550	0.7600
Level-3	0.9977	0.9578	0.9189	0.8832	0.9960	0.9970

The experimental results indicate that the prediction performance of the model at the second level is significantly lower than that at the first level and third level. Although the ultimate goal of the model is to predict acupoint prescriptions, it can be reasonably inferred that enhancing the performance of the second level predictions may lead to improved performance at the third level. The primary reason for the suboptimal performance at the second level is attributed to data imbalance. Among

the 4,000 training samples, the second level encompasses as many as 1,576 categories, with an average of only 2.5 samples per category. Furthermore, the majority of categories contain only a single sample. This severe data imbalance limits the model's learning capacity and leads to reduced performance the second level. Based on the above analysis, it is reasonable to posit that data augmentation represents an effective strategy to alleviate data imbalance and enhance overall model performance.

4.2.1.2 Performance improvement through data augmentation

Figure 4.8 consists of three charts, each illustrating the trends in key metrics across different levels of the model during the data augmentation process. It is evident from the figure that as the data volume increases, the predictive performance of the model at the second level improves significantly, while the results for the first level and third level also show varying degrees of enhancement.

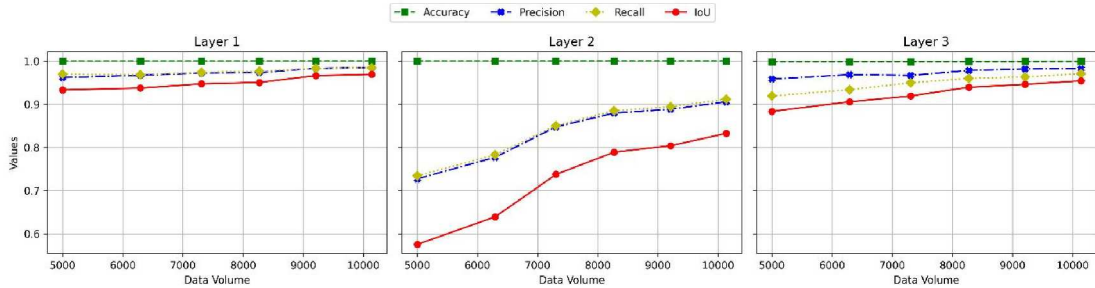


Figure 4.8: Trend analysis of performance metrics for three hierarchical levels with data expansion.

This trend demonstrates that data augmentation effectively mitigates the issues of insufficient samples and class imbalance, thereby improving the overall classification performance of the model. In particular, for the second level, which involves a large number of categories, data augmentation serves an important function in improving the model's predictive capabilities, further demonstrating its effectiveness.

4.2.1.3 Model comparison on expanded dataset

To evaluate and compare the performance of the hierarchical classification model with other machine learning models in predicting acupoints prescriptions, several algorithms were selected as baseline methods. These methods encompass traditional machine learning, ensemble learning, and deep learning approaches, as summarized in Table 4.4.

Table 4.4: Summary of baseline models for extended dataset.

Category	Model	Description
Traditional Machine Learning	Naïve Bayes	Implemented using the Scikit-learn library.
	Decision Tree	
Ensemble Learning	Extra Tree	Implemented using the Scikit-learn library.
Deep Learning	Feedforward Neural (FNN)	Classical multilayer perceptron (MLP) architecture.
	TextCNN	Adjusted loss function for multi-label classification.
	Text-GCN	Constructed TCM graph with disease and symptom nodes, used GCN for supervised learning.
	ML-GCN	Built an acupoint graph and applied ML-GCN for multi-label acupoint prediction.
	Seq2Seq with Attention	Used attention mechanism to translate symptoms into acupoints for prescription prediction.
	MacBERT	A Transformer-based pre-trained language model optimized for Chinese.

To address the issue of imbalanced data distribution, we employed 5-fold cross-validation to evaluate the model. 1,000 original samples were reserved as the final test set and used for visualization analysis, while the augmented dataset of 9,138 samples was utilized for the 5-fold cross-validation experiments. The test set (1,000 samples) consists exclusively of original, non-augmented data that were held out prior to training, ensuring a rigorous evaluation of the model’s generalization performance. Table 4.5 presents the average performance of various methods in predicting acupoint prescriptions, obtained through 5-fold cross-validation. The results presented in the table represent the best performance achieved by tuning the hyperparameters, such as the learning rate, number of network layers, and dropout rate. Table 4.6 provides a detailed breakdown of the results and average values obtained from the 5-fold cross-validation specifically for HARNN.

Table 4.5: Experimental results of different machine learning algorithms on the expanded dataset.

Evaluating indicator	Models									
	Naive Bayes	Decision-Tree	Extra-Tree	Text-CNN	Text-GCN	ML-GCN	FNN	Seq2seq with attention	HARNN	MacBERT
Accuracy	0.977	0.998	0.997	0.996	0.944	0.901	1.000	0.999	0.999	0.999
Precision	0.332	0.549	0.496	0.566	0.097	0.022	0.618	0.961	0.982	0.989
Recall	0.551	0.539	0.480	0.461	0.124	0.093	0.610	0.960	0.971	0.984
IoU	0.323	0.514	0.432	0.445	0.038	0.016	0.599	0.942	0.954	0.973

As shown in Table 4.5, MacBERT achieved the highest performance, with an average IoU of 0.973 under 5-fold cross-validation. HARNN followed closely with an IoU of 0.954, and Seq2Seq with attention achieved an IoU of 0.942. In contrast,

Table 4.6: Experimental results of the 5-fold cross validation of HARNN on the expanded dataset.

Evaluating indicator	First	Second	Third	Fourth	Fifth	Average Value
Accuracy	0.999	0.999	0.999	0.999	0.999	0.999
Precision	0.980	0.982	0.982	0.982	0.981	0.982
Recall	0.972	0.968	0.971	0.973	0.971	0.971
IoU	0.953	0.952	0.954	0.956	0.953	0.954

all other models yielded significantly lower IoUs, with none exceeding 0.599. These results demonstrate that HARNN achieves comparable accuracy to MacBERT and Seq2Seq with attention.

Among the tested models, MacBERT, HARNN, and Seq2Seq with attention all achieved satisfactory experimental results. However, hierarchical learning adopts a structured approach that better reflects the theoretical framework of TCM, enabling it to not only predict acupoint prescriptions but also provide disease names and subcategories. This aligns more closely with human diagnostic processes and enhances practical applicability. In addition, Table 4.6 demonstrates the stability of HARNN based on the results of the 5-fold cross-validation conducted on the expanded dataset. To further validate the hierarchical classification model, we test it on 1,000 samples from the original dataset. The results showed an Accuracy of 0.999, Precision of 0.971, Recall of 0.959, IoU of 0.932, Top-3 of 0.990, Top-5 of 0.995, demonstrating its effectiveness and reliability.

Table 4.7: Paired t-test results comparing HARNN with other models.

Model	t-value	df	p-value
MacBERT	-13.988	4	0.00015
Seq2seq with attention	8.344	4	0.00113
ML-GCN	558.832	4	0.00000
TextCNN	98.424	4	0.00000
FNN	154.777	4	0.00000
Naive Bayes	98.577	4	0.00000
Decision tree	152.73	4	0.00000
ExtraTree	145.116	4	0.00000
Text-GCN	228.543	4	0.00000

Table 4.7 presents the paired t-test results comparing HARNN with other models. All p-values are below the 0.05 threshold, indicating that the performance differences are statistically significant (df = 4, based on 5-fold cross-validation). Except for

the comparison with MacBERT, where the t-statistic is -13.988, all other t-values are positive. This suggests that while HARNN performs lower than MacBERT, it consistently outperforms the remaining models in our evaluation.

4.2.2 Case Study

We present two case studies to illustrate the model’s predictive behavior and interpretability. Case 1 demonstrates HARNN’s prediction process with LIME-based explanations, while Case 2 examines a misclassification to identify possible causes of error.

4.2.2.1 Case 1

We selected one case from 1,000 test samples that were not involved in model training to demonstrate the explanatory effect of LIME. Figure 4.9 presents the results of the LIME analysis for the first-level classification task of this case. From the figure, it can be observed that the LIME output consists of three parts. The first part lists all the symptoms of the case; the second part shows the model’s prediction of the disease name; and the third part illustrates the contribution of each symptom to the model’s decision on whether the case belongs to a specific label. This analysis provides an intuitive understanding of the model’s reasoning and the significance of each symptom in its predictions. As shown in the figure, some symptoms do not support the diagnosis of the disease as a common cold. This may be due to the inconsistency between these symptoms and the cold symptoms learned by the model, or the fact that the symptoms from the cold samples the model has learned do not include these symptoms. However, these symptoms are exactly the key symptoms that doctors need to focus on in clinical diagnosis.

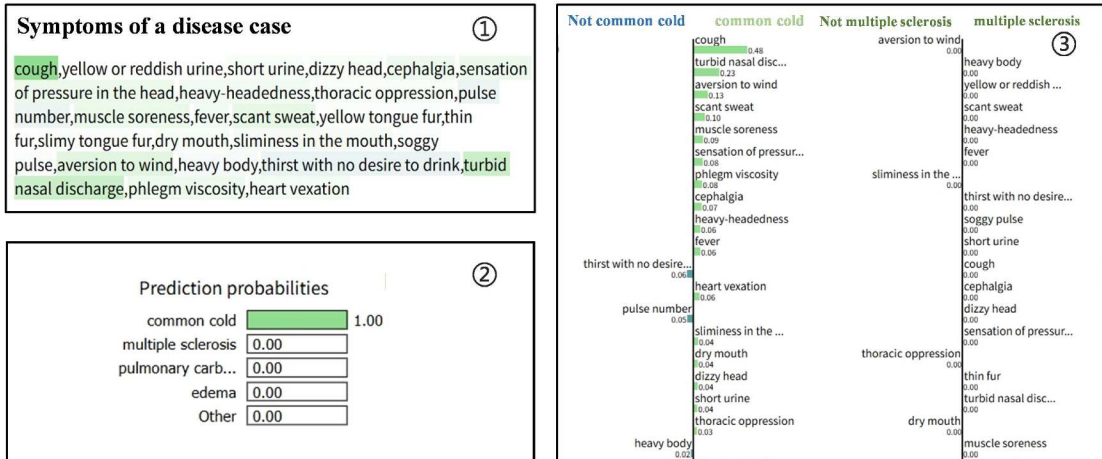


Figure 4.9: LIME visualization of a disease case in disease name classification.

Figure 4.10 illustrates the step-by-step prediction results of LIME for the three-level classification task. At the first level, the model diagnoses the test case as “com-

mon cold” with a very high probability, while assigning very low probabilities to “multiple sclerosis” or other diseases. At the second level, the model further specifies the diagnosis as “common cold with summer heat-dampness assailing the exterior.” The third level presents the model’s predictions for acupoint prescriptions. We visualize the top 15 acupoints ranked by prediction probability to explain the model’s decision-making process and provide intuitive support for practical applications, as shown in Figure 4.11. By observing the weight of each symptom at each acupoint, we can not only validate the reasonableness of the model’s results but also further investigate the potential relationship between acupoints, diseases, and symptoms.

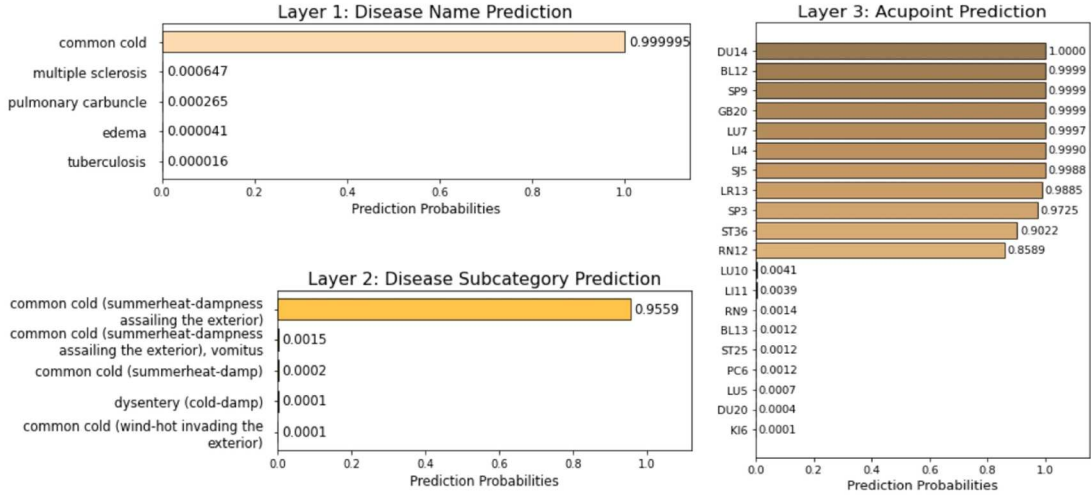


Figure 4.10: LIME visualization of prediction probabilities for disease name, subcategory, and acupoint.

As shown in Figure 4.12, the LIME explanations across three repeated runs for the same case are not entirely consistent. While the overall patterns remain similar, slight variations can be observed in the contribution scores of certain symptoms. This variability arises from the inherent randomness and locality of the LIME algorithm. Nonetheless, the directionality of each symptom’s contribution, whether it is positive or negative, remains largely stable across runs, suggesting that LIME still provides meaningful insights into the model’s decision-making process.

In addition, we leveraged the built-in attention mechanism of HARNN to visualize the model’s focus on each symptom during prediction, as illustrated in Figure 4.13 and Figure 4.14. Unlike LIME, which estimates the contribution of each symptom to a specific prediction outcome, attention visualization reflects how much weight the model assigns to each input symptom during its decision-making process. To investigate the consistency of model interpretability across different training splits,

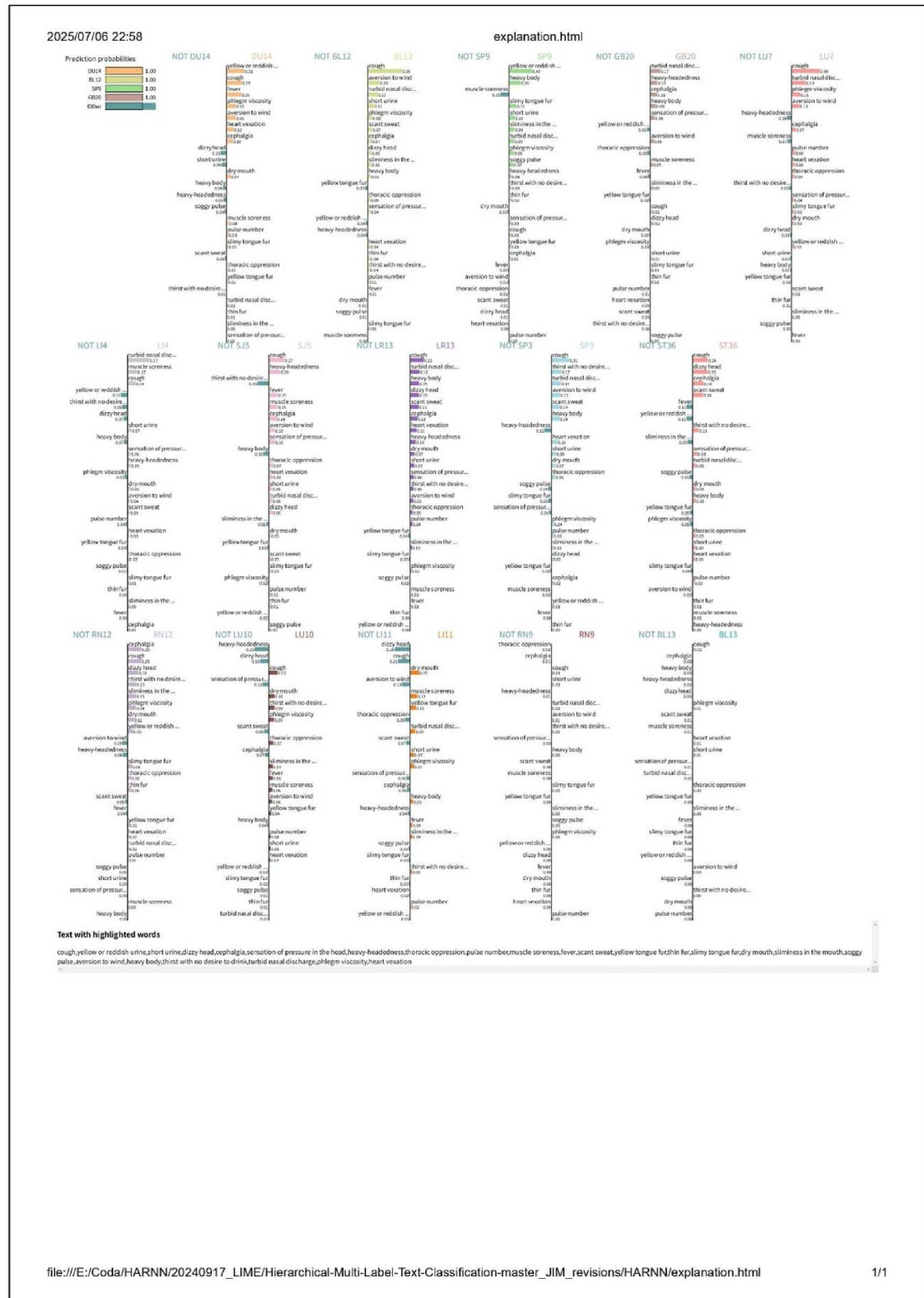


Figure 4.11: LIME visualization of the top 15 acupoints ranked by prediction probabilities to explain the model's decision process.

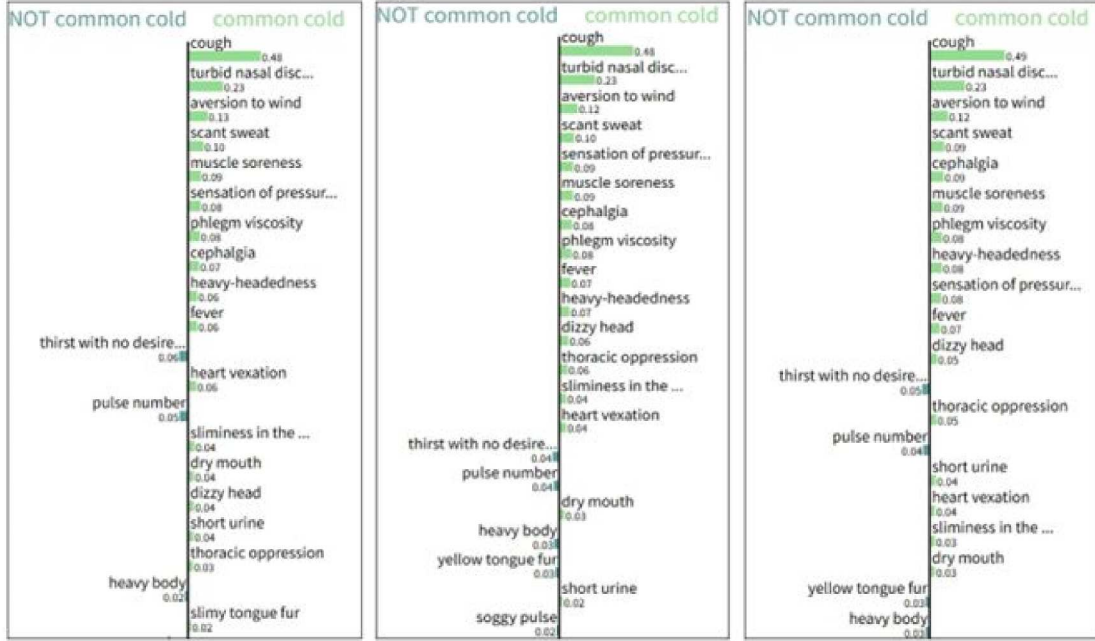


Figure 4.12: LIME visualizations for Level 1 predictions in Case 1 across three repeated runs.

we visualized the attention weights of HARNN models trained on distinct folds of a five-fold cross-validation, as illustrated in Figure 4.13 and Figure 4.14 by the attention maps from two different folds. Despite achieving similar overall classification performance, the attention distributions varied significantly between models trained on different data splits. This suggests that attention, while offering a hierarchical explanation, is highly sensitive to training randomness, making it difficult to provide a stable interpretation framework. In contrast, LIME produces local explanations by perturbing input features and observing their influence on the model’s predictions. Although slight variations exist between different model instances, the core explanatory patterns—such as which symptoms contribute to or against a label—remain largely consistent. This suggests that LIME may offer a more stable and interpretable lens for understanding the decision-making process of the model.

4.2.2.2 Case 2

This case represents a misprediction made by the model during evaluation on 1,000 test samples. We present it here to analyze the factors contributing to the discrepancy between the predicted and target labels.

Input Symptoms:

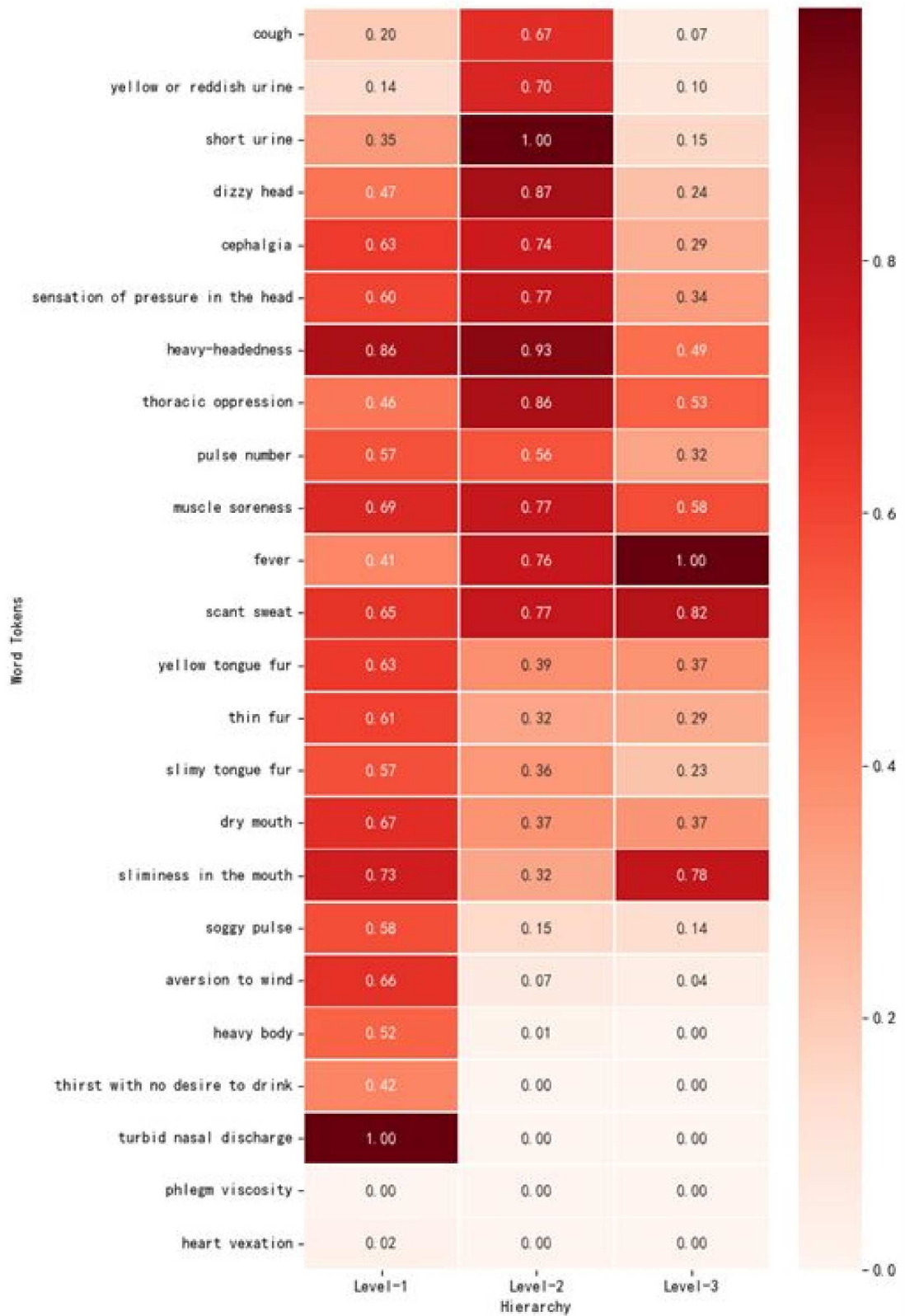


Figure 4.13: Hierarchical Attention on Symptoms for Case 1 (Fold 1 Model).

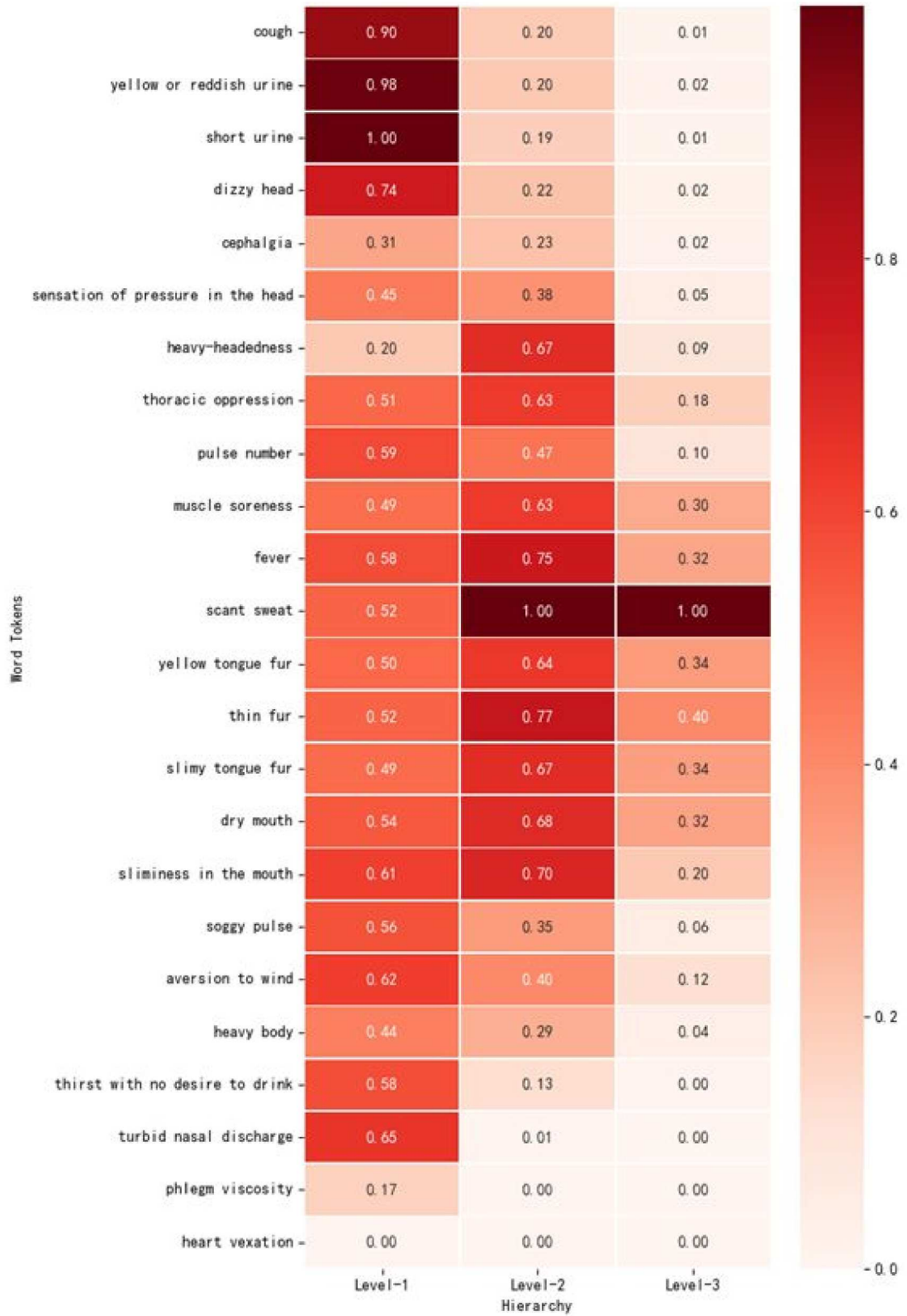


Figure 4.14: Hierarchical Attention on Symptoms for Case 1 (Fold 2 Model).

acute phase, convulsion, rigidity of the neck and nape, bruxism, trismus

Target Labels from Database:

- Level 1 (Disease name): infantile convulsion (Disease number: 68)
- Level 2 (Subcategory): infantile convulsion (acute infantile convulsion), trismus, a (Subcategory number: 1059)
- Level 3 (Acupoints): DU26, EX-HN3, LI4, LR3, PC9, ST6

Model Prediction:

- Level 1 (Disease name): infantile convulsion (Disease number: 68)
- Level 2 (Subcategory): infantile convulsion (acute infantile convulsion), trismus, b (Subcategory number: 1067)
- Level 3 (Acupoints): DU26, EX-HN3, LI4, LR3, ST6

A comparison between the target labels and model predictions shows that the model correctly predicted the Level 1 category but misclassified Levels 2 and 3. To investigate the cause, we examined relevant samples from the training and test sets, as shown in Table 4.8.

The test cases No.1654 and No.3389, as well as the training cases No.7017, No.7018, and No.7019, are all labeled as infantile convulsion (acute infantile convulsion), trismus at Level 2. Among them, cases No.7017, No.7018, No.7019, and No.3389 have minor differences in symptoms but share the same acupoint prescription. They are therefore grouped under the same subcategory. In contrast, case No.1654 has a different prescription and is assigned to a separate subcategory. To distinguish between such cases during data construction, we added suffixes “a” and “b” to the subcategory name.

These two groups of cases come from different sources. Cases No.7017, No.7018, No.7019, and No.3389 are from Reference [2], while case No.1654 is from Reference [53]. The difference in acupoint prescriptions for similar conditions across sources directly led to the mismatch between the model’s prediction and the target label.

Table 4.8: Related Training and Test Samples for Case 2.

	Disease case number	Disease name	Disease number	Subcategory	Subcategory number	Symptoms	Acupoints	Data Source
Train set	7017	infantile convulsion	68	infantile convulsion (acute infantile convulsion), trismus, b	1067	acute phase, hypertonicity, trismus	DU26, EX-HN3, LI4, LR3, ST6	[2]
	7018	infantile convulsion	68	infantile convulsion (acute infantile convulsion), trismus, b	1067	acute phase, myotonia, dysthymic disorder, trismus	DU26, EX-HN3, LI4, LR3, ST6	[2]
	7019	infantile convulsion	68	infantile convulsion (acute infantile convulsion), trismus, b	1067	acute phase, hypertonicity, dysthymic disorder, trismus	DU26, EX-HN3, LI4, LR3, ST6	[2]
Test set	1654	infantile convulsion	68	infantile convulsion (acute infantile convulsion), trismus, a	1059	acute phase, convulsion, rigidity of the neck and nape, bruxism, trismus	DU26, EX-HN3, LI4, LR3, PC9, ST6	[53]
	3389	infantile convulsion	68	infantile convulsion (acute infantile convulsion), trismus, b	1067	acute phase, myotonia, trismus	DU26, EX-HN3, LI4, LR3, ST6	[2]

4.2.3 Discussion

4.2.3.1 Model Performance and Interpretability

This study compares traditional machine learning methods (such as support vector machines, decision trees), ensemble learning method (such as Extra tree), and deep learning methods (such as FNN, TextCNN, TextGCN, ML-GCN, Seq2seq with attention, HARNN, MacBERT) to train models by using acupoint prescription data, providing corresponding acupoint prescriptions based on patients' symptoms. Among all models, the Transformer-based MacBERT achieved the best performance, followed by HARNN and Seq2Seq with attention, all of which have achieved satisfactory results. All three models incorporate attention mechanisms, underscoring the effectiveness of attention in capturing the complex relationships between symptoms and acupoint prescriptions, and highlighting its value in the application of acupuncture and moxibustion treatment [66]. Moreover, although methods such as Text-GCN and ML-GCN did not achieve the best performance in the current experimental setup, they offer a promising alternative for modeling the structured and interrelated nature of TCM data. By constructing knowledge graphs based on theoretical relationships among symptoms, meridians, and acupoints, GNNs can capture multi-relational dependencies that are difficult to model with conventional neural architectures. This direction may better align with the inherent logic of TCM and represents a valuable area for future research beyond the current HARNN framework.

In our experiments, we tested several encoding methods for symptoms, including one-hot encoding, Word2Vec, and BERT-base-Chinese. However, in the case of HARNN, the differences in model performance across these encoding methods were

relatively minor. The best-performing method, BERT-base-Chinese, outperformed the others by an average margin of less than 0.01 in five-fold cross-validation. These results suggest that, for this task, the hierarchical structure may have a greater impact on model performance than the choice of symptom encoding.

From the LIME interpretation of the current model’s prediction results, the acupoint selection logic of the model is largely consistent with the logic described in references [2, 52, 53]. This indicates that the model can preliminarily capture the key principles of AMT. LIME offers intuitive visual explanations that are particularly accessible to non-technical audiences and enable rapid single-sample analysis. However, its reliance on local surrogate models can sometimes result in unstable explanations. On the other hand, SHAP provides both local and global feature attributions grounded in a solid theoretical foundation, satisfying the Shapley axioms and offering higher mathematical precision. However, its model-agnostic implementations, such as KernelSHAP, often come with high computational costs and complexity [43]. Considering these complementary strengths and limitations, we view SHAP as a valuable addition to our interpretability toolkit. Therefore, in future work, we plan to explore SHAP and other alternative explanation methods to achieve more stable and reliable model interpretability.

4.2.3.2 Limitations and Challenges

This study primarily explores data processing strategies and machine learning models suitable for acupuncture prescriptions. While the results are promising, several challenges remain for real-world deployment, mainly concerning data quality, model generalization, and clinical validation:

1. Limitations in Data Quality and Representativeness

Lack of real-world clinical data: The model is trained on datasets derived from classical TCM literature and textbooks, which offer authoritative and standardized information but fall short of capturing the complexity and individual variability present in real clinical cases. Future work should incorporate real-world electronic medical records (EMRs) and involve TCM experts in annotation and validation.

Limitations of data construction methods: The rule-based symptom reduction approach used for data augmentation helps increase dataset size but may introduce subjectivity and bias. Before clinical deployment, all training samples should undergo expert review to ensure data reliability.

2. Constraints in Model Generalization

Class imbalance and data sparsity: The dataset contains many low-frequency subcategories, which reduces the model’s ability to generalize. While the current augmentation strategy provides partial relief, further improvements are needed.

Limited adaptability to novel clinical scenarios: The model struggles to generate prescriptions for symptom combinations or treatment contexts not seen in the training data. This reflects a common limitation of conventional machine learning methods based solely on historical patterns.

3. Lack of Clinical Validation and Practical Evaluation

Inadequate evaluation metrics: Current metrics such as IoU and accuracy assess label alignment but do not reflect the clinical validity of the predicted prescriptions. A more comprehensive evaluation framework is needed, incorporating expert assessment and patient treatment feedback.

Room for performance improvement: Although the proposed method achieves reasonable results, its accuracy still lags behind BERT-based models, indicating the need to explore more advanced model architectures.

4.3 Conclusion

This chapter proposed a methodology for providing acupoint prescriptions based on symptoms for treating patients, through constructing a hierarchical and interpretable machine learning model. Specifically, our model effectively utilizes the hierarchical structure between diseases, subcategories, and acupoints in AMT through the hierarchical learning model HARNN, addressing the issue of poor model performance in multi-label tasks due to the excessive number of labels. IoU is 0.954 in 5-fold cross-validation with augmented data and 0.932 in the test set of 1,000 original samples. Additionally, by using the LIME method to visualize the model's prediction rationale, the credibility of the model is enhanced. Our research advances the application of AI technology in AMT and provided strong support for the future development of data-driven intelligent support systems for AMT. This achievement is expected to play a key role in improving treatment effectiveness, optimizing treatment plans, and enabling personalized healthcare, thereby laying a solid foundation for the intelligent development of AMT.

Chapter 5

System Design and Implementation

This chapter presents the practical culmination of our research: the design and implementation of an interactive AMT Support System. Leveraging the hierarchical and interpretable model from Chapter 4, the system is engineered to bridge the gap between computational analysis and real-world clinical practice. This chapter provides a comprehensive overview of the system's user interface and configuration. The system's capabilities are then showcased through a detailed analysis of two clinical scenarios, demonstrating its potential to assist practitioners in decision-making.

5.1 Overview of the Intelligent Support System

Figure 5.1 illustrates the diagnostic and treatment workflow of AMT using our proposed intelligent AMT support system. This system primarily focuses on optimizing ③, ④, ⑤, ⑨, and ⑩ in the process. We first construct an acupoint prescription database ⑩, which serves as the training data source for a hierarchical and interpretable machine learning model ③. At the beginning of diagnosis and treatment, symptoms are first collected from patients using relevant medical equipment, based on the TCM diagnostic methods of inspection, auscultation and olfaction, inquiry, and palpation. Next, the symptoms ② obtained from the four diagnostic methods are fed into the trained machine learning model ③, which generates a disease classification, syndrome subtype, personalized acupoint prescription ④, and corresponding LIME-based interpretability analysis ⑤. The doctor subsequently examines relevant historical cases from the database in conjunction with recommendations provided by the model, to determine whether it is appropriate to modify the proposed acupoints ⑥. If necessary, the prescription is adjusted based on clinical expertise; otherwise, the recommended acupoint prescription is directly applied. After receiving treatment, the patient provides feedback on their recovery within a specified period. This feedback is recorded in the database ⑩ to further enhance the model's accuracy and improve future predictions.

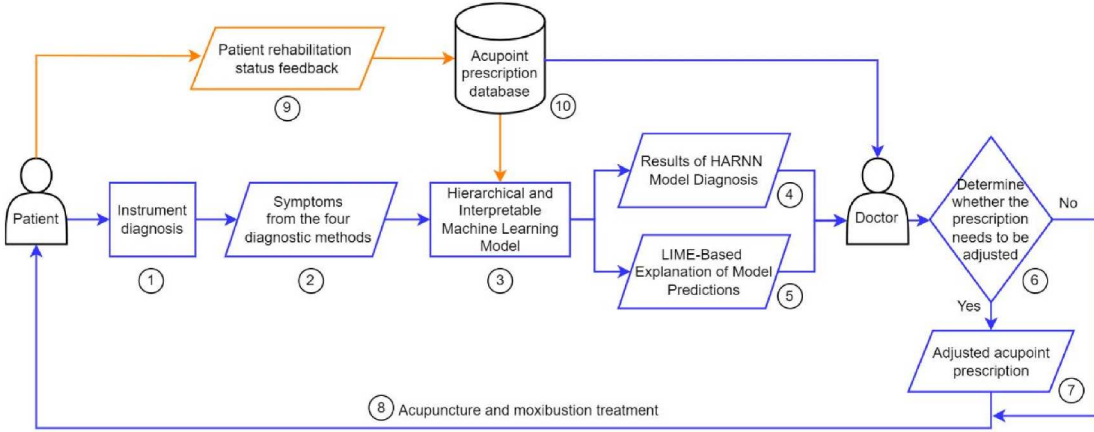


Figure 5.1: Diagnosis and Treatment Workflow with the Intelligent AMT Support System.

5.2 System Implementation

The following details the development of the intelligent acupuncture treatment support system, covering its technology stack, model integration, and user interface design. The system was built using modern, lightweight technologies that enable rapid prototyping and interactive model inference, allowing clinicians to input symptoms and receive acupoint recommendations along with interpretability feedback.

Technology Stack and Model Integration

The system is implemented in Python using TensorFlow for deep learning and Streamlit to build the web-based interface. At startup, the pre-trained HARNN model is loaded in inference mode. User-input symptoms (selected from predefined sets) are processed in real time to generate predictions, including disease classification, subtype identification, and recommended acupoint prescriptions. To enhance transparency, LIME is incorporated into the pipeline to visualize the key contributing symptoms for each prediction.

User Interface Design

The user interface is developed with Streamlit and is organized into the following functional sections:

- Symptom Input Panel: Enables users to select or enter symptom descriptions.
- Prediction Display Panel: Shows the predicted disease, syndrome subtype, and acupoint prescription.
- Interpretability Visualization: Presents LIME-generated visualizations that highlight the symptoms most influential in the decision-making process.
- Case Feedback Section: (For future integration) Allows recording of patient rehabilitation feedback.

Example screenshots of the system interface are provided below. Figure 5.2 and 5.3 provide illustrative screenshots of the AMT Intelligent Support System interface. Figure 5.2 presents the main interface of the system. A language selection menu is available in the sidebar, offering three language options: Chinese, English, and Japanese. On the main panel, clinicians first input the patient's symptoms, then click the "Generate Recommended Prescription" button to initiate the prediction process. The system returns a three-level diagnostic prediction generated by the HARNN model, along with corresponding LIME-based interpretability visualizations. The navigation bar includes three modules: the Home page, "Enter Treatment Record," and "View Past Records." Figure 5.3 shows the treatment input page, where clinicians

can enter patient information, symptom descriptions, and corresponding prescriptions into the system.

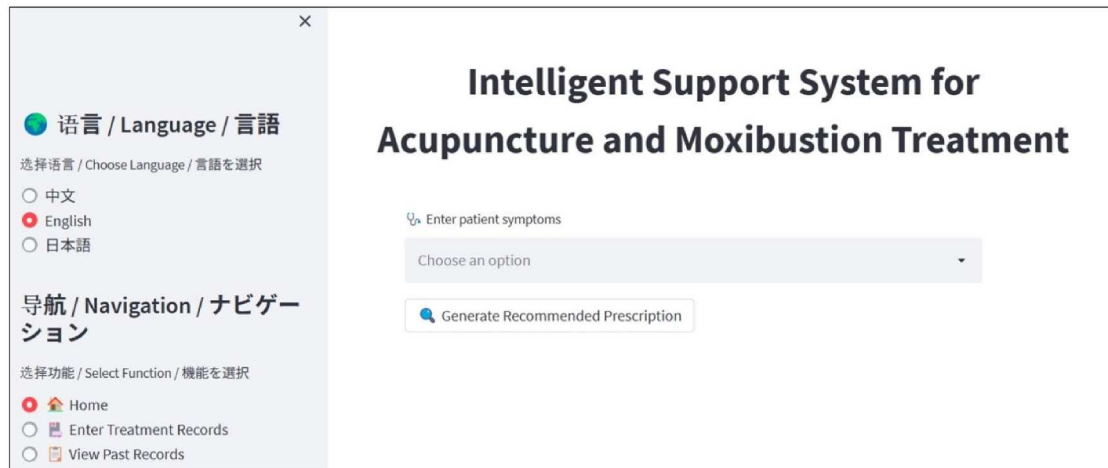


Figure 5.2: Main Interface of the AMT Intelligent Support System.

Notes on Deployment

Currently, the system is deployed and tested in a local environment, suitable for research and demonstration purposes. Future extensions may include cloud-based deployment, database integration, and mobile-friendly adaptations.

Acupuncture Treatment Record System

● 语言 / Language / 言語

选择语言 / Choose Language / 言語を選択

☐ 中文

☒ English

☐ 日本語

导航 / Navigation / ナビゲーション

选择功能 / Select Function / 機能を選択

☐ Home

☒ Enter Treatment Records

☐ View Past Records

👤 Patient Name

🩺 Patient Symptoms

📌 Acupoint Prescription

📅 Treatment Date

2025/06/15

💬 Patient Feedback

💾 Save Record

Figure 5.3: Treatment Record Interface of the AMT Intelligent Support System.

5.3 Case Study

5.3.1 Experimental Setup

To further demonstrate the effectiveness and interpretability of the proposed system, we conducted detailed case studies on two representative test samples. These samples were selected from a held-out test set of 1,000 patient cases that were not seen during training. Each case includes a set of input symptoms and the corresponding predicted acupoint prescription generated by the HARNN model.

The case studies were conducted on a local machine with the following experimental setup:

- CPU: Intel Core i7-1195G7
- Memory: 8GB RAM
- Operating System: Windows 11
- Programming Language: Python 3.6.13
- Framework: TensorFlow 1.15.0
- Visualization & Interface: Streamlit 1.10.0
- Interpretability Tool: LIME 0.2.0.1

In the following case studies, we present the model’s prediction results alongside interpretability visualizations, highlighting how the model arrives at its recommendations and how these insights can aid physicians in understanding the rationale behind them. These examples demonstrate the system’s potential to support clinical decision-making by providing recommendations that are both accurate and interpretable.

5.3.2 Case 1

This case is selected from the reserved test set to demonstrate the system’s prediction and interpretability output for a real-world clinical record.

- Input Symptoms:
myosalgia, muscle swelling, muscle soreness, skin numbness, pale red tongue, yellow tongue fur, slimy tongue fur, dry mouth, bitter taste in the mouth, stringlike pulse, slippery pulse
- Target Labels from Database:
 - Level 1 (Disease Category): restless leg syndrome

- Level 2 (Subtype): restless leg syndrome (damp-heat impeding the channels)
- Level 3 (Acupoint Prescription): BL56, BL57, SP9, ST40

Figure 5.4 illustrates the user interface and system response after entering the symptom descriptions for Case 1. Once the user clicks the “Generate Recommended Prescription” button, the system returns its three-level prediction within approximately one second. The predicted results include the disease name (“restless leg syndrome”), the disease subcategory (“restless leg syndrome: damp-heat impeding the channels”), and the recommended acupoint prescription (“BL56, BL57, SP9, ST40”). Meanwhile, a prompt message—“Generating LIME visualization analysis for you, please wait”—appears on the screen along with an estimated computation time. Under the current system configuration, generating the interpretability results takes approximately 40 minutes. Figure 5.5 to 5.8 present the LIME-based interpretability analysis for each of the three prediction levels.

Enter patient symptoms

myosalgia × muscle swelling × muscle soreness × skin numbness × pale red tongue ×

yellow tongue fur × slimy tongue fur × dry mouth × bitter taste in the m... ×

stringlike pulse × slippery pulse ×

Generate Recommended Prescription

Analysis of Symptoms Based on Models:

The possible disease names for this patient are: restless leg syndrome

Prediction results of syndrome types:

The possible disease subcategory names for this patient are: restless leg syndrome (damp-heat impeding the channels)

Prediction of acupoint prescription:

Based on the analysis of symptoms through the model, the appropriate acupoint prescription for this patient is:
ST40, SP9, BL56, BL57

Generating LIME visualization analysis for you, please wait...

Figure 5.4: System Output for Case 1: Hierarchical Predictions Including Disease Name, Subcategory, and Recommended Acupoints. (Source: Created by the authors)

Figure 5.5 presents the model’s prediction for the disease name associated with this case. The model identifies “restless leg syndrome” with a high confidence of

approximately 0.9965, while also listing other possible disease names along with their respective probabilities. This prediction is consistent with the reference label in the database.

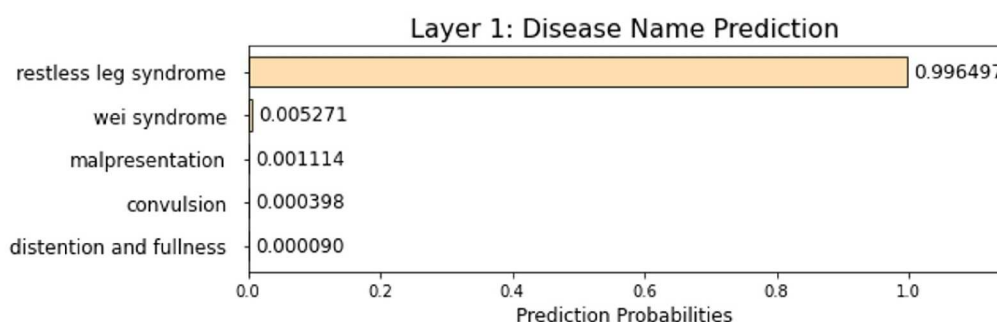


Figure 5.5: Top Prediction Probabilities for Disease Names - Case 1 (Level 1). (Source: Created by the authors)

Figure 5.6 presents the LIME interpretability analysis of the model’s prediction of “restless leg syndrome” for this case. The vertical axis lists the patient’s symptoms, while the horizontal axis shows each symptom’s contribution score toward the diagnosis. Positive scores indicate supportive symptoms, with higher values reflecting stronger influence on the model’s decision; negative scores suggest symptoms that contradict or reduce confidence in the diagnosis. The model ultimately classified the case as “restless leg syndrome” because the total contribution of supportive symptoms outweighed that of the opposing ones. By comparing the model’s reasoning with traditional TCM theories, the interpretability analysis helps judge whether the prediction is reasonable and offers a helpful reference for manual review.

Figure 5.7 and 5.8 present the LIME-based interpretability results for the second-level prediction. The model predicted the patient’s disease subcategory as “restless leg syndrome (damp-heat impeding the channels)” with a high confidence score of 0.9939, which is consistent with the ground truth. The interpretability visualization further reveals the contribution of individual symptoms to this prediction, highlighting which symptoms played a key role in the model’s diagnostic decision.

Figure 5.9 illustrates the predicted probabilities of acupoints recommended by the system for this case. Only acupoints with probabilities greater than 0.5 are considered for use. As shown in Figure 5.9, the top four predicted acupoints match the target prescription. A more detailed interpretability analysis of the top 15 predicted acupoints is presented in Figure 5.10, 5.11, and 5.12 using LIME.

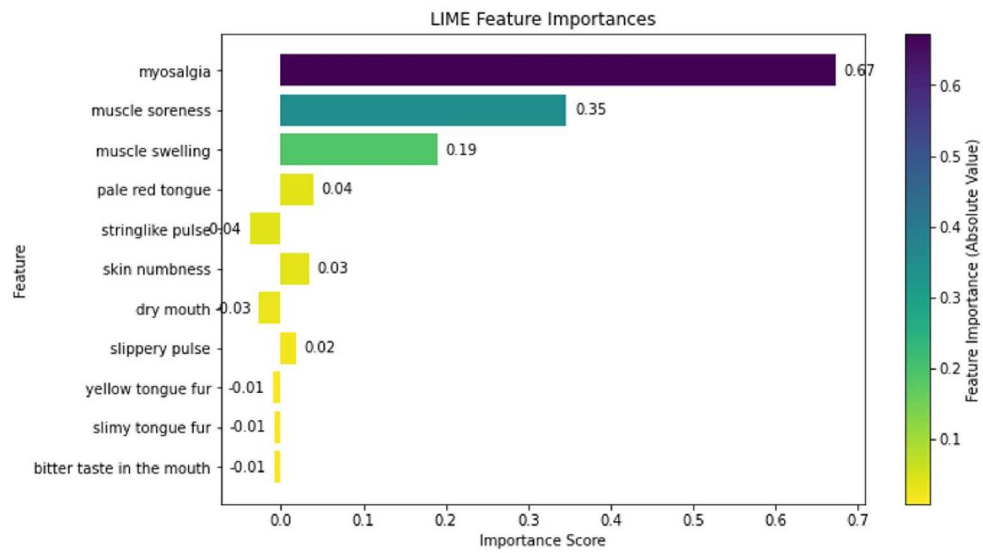


Figure 5.6: LIME Explanation for Top-1 Disease Name Prediction - Case 1 (Level 1).
(Source: Created by the authors)

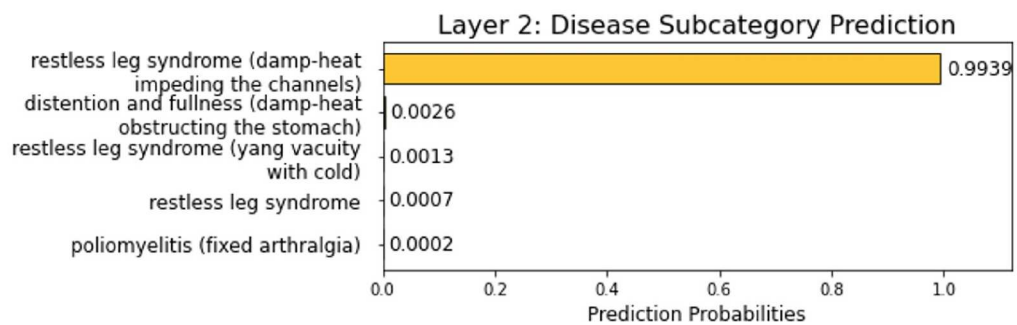


Figure 5.7: Top Prediction Probabilities for Disease Subcategories - Case 1 (Level 2).
(Source: Created by the authors)

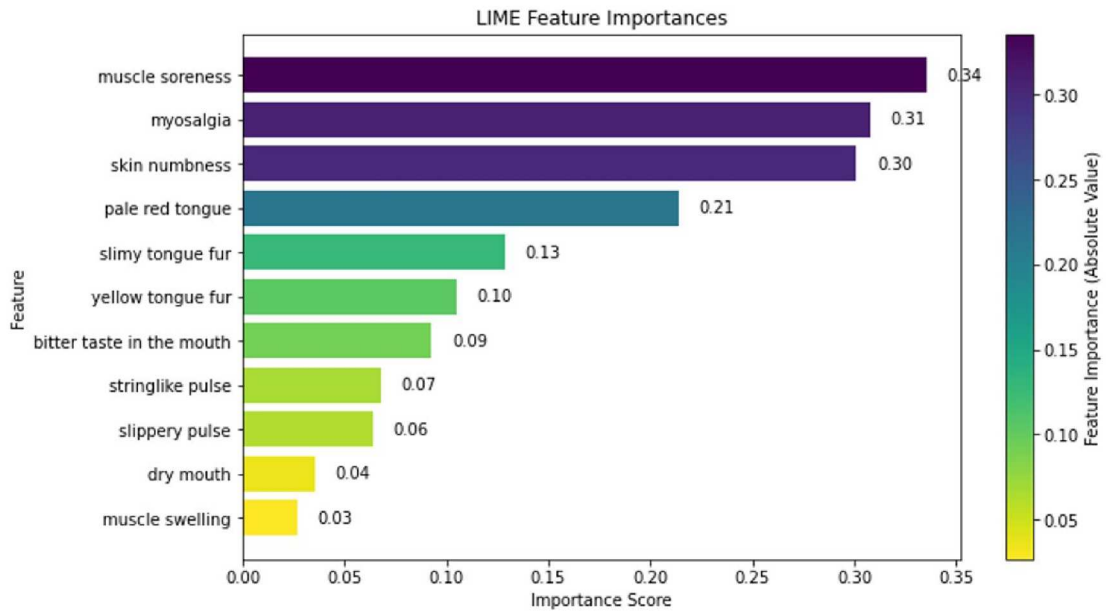


Figure 5.8: LIME Explanation for Top-1 Disease Subcategory Prediction - Case 1 (Level 2). (Source: Created by the authors)

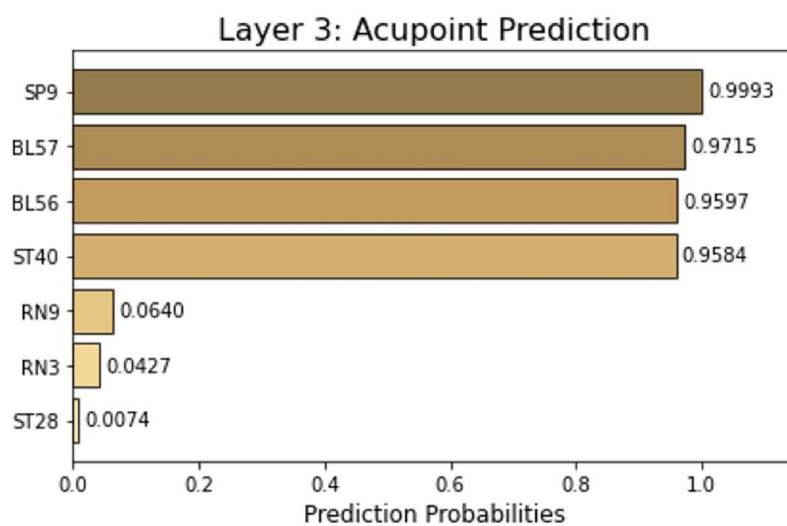


Figure 5.9: Top Prediction Probabilities for Acupoint Prescriptions - Case 1 (Level 3). (Source: Created by the authors)

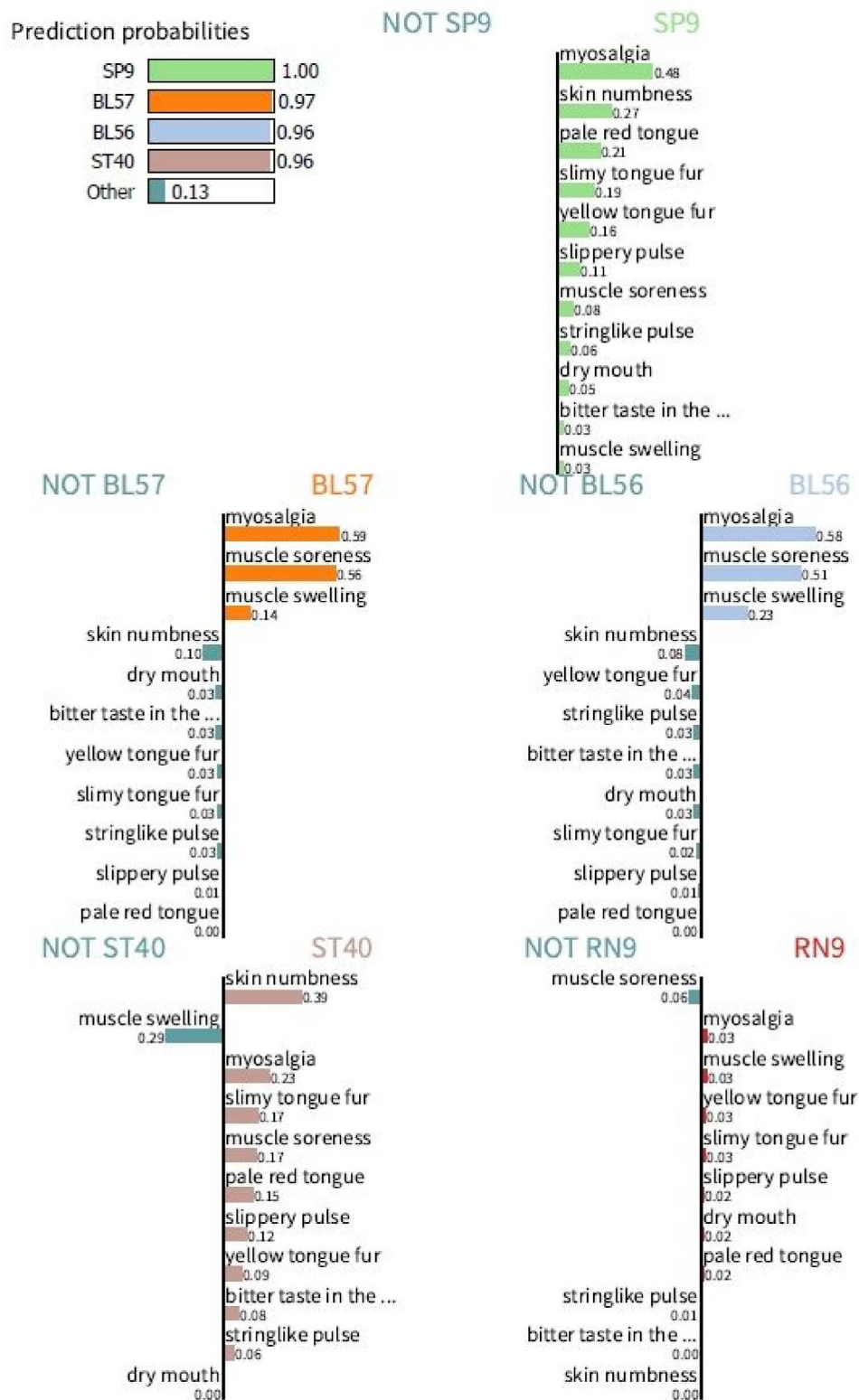


Figure 5.10: LIME Visualization of the Top 15 Predicted Acupoints for Case 1 (Part 1).

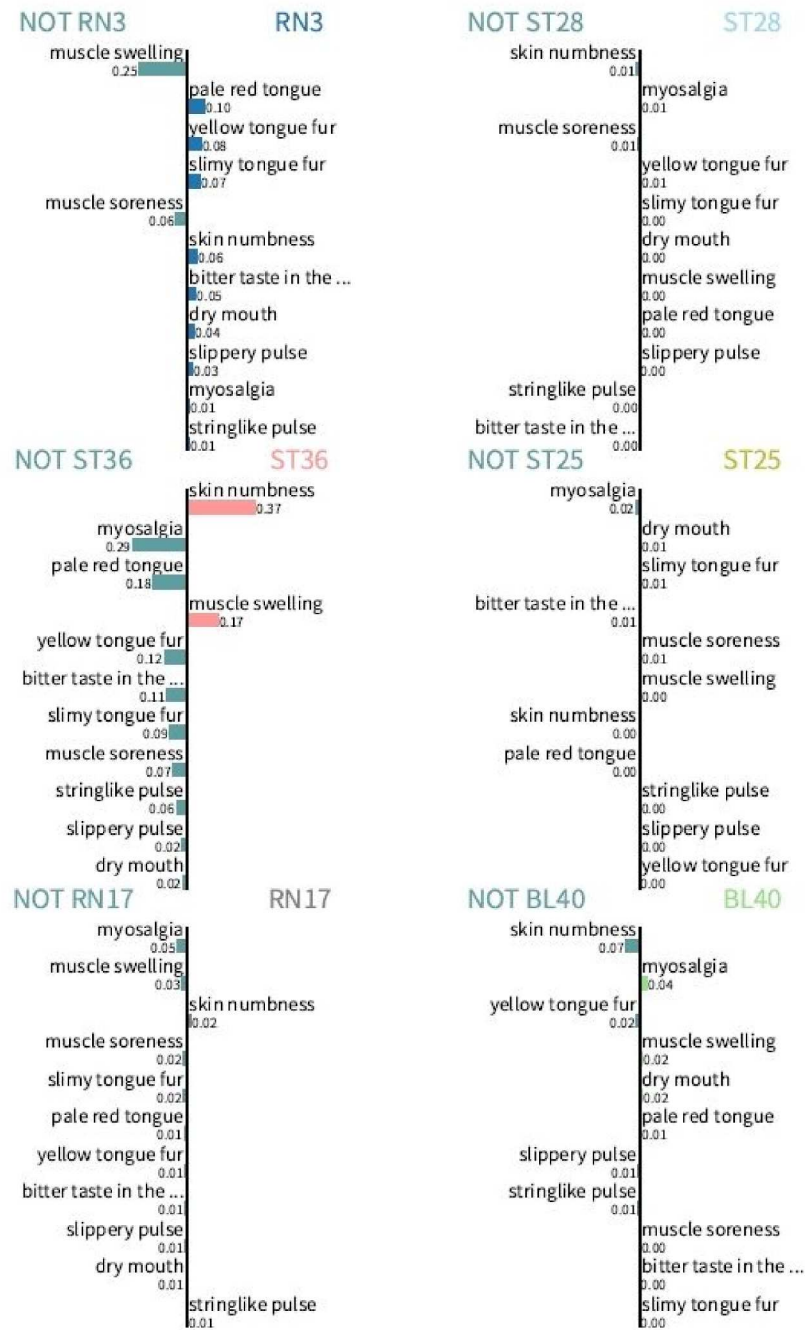


Figure 5.11: LIME Visualization of the Top 15 Predicted Acupoints for Case 1 (Part 2).

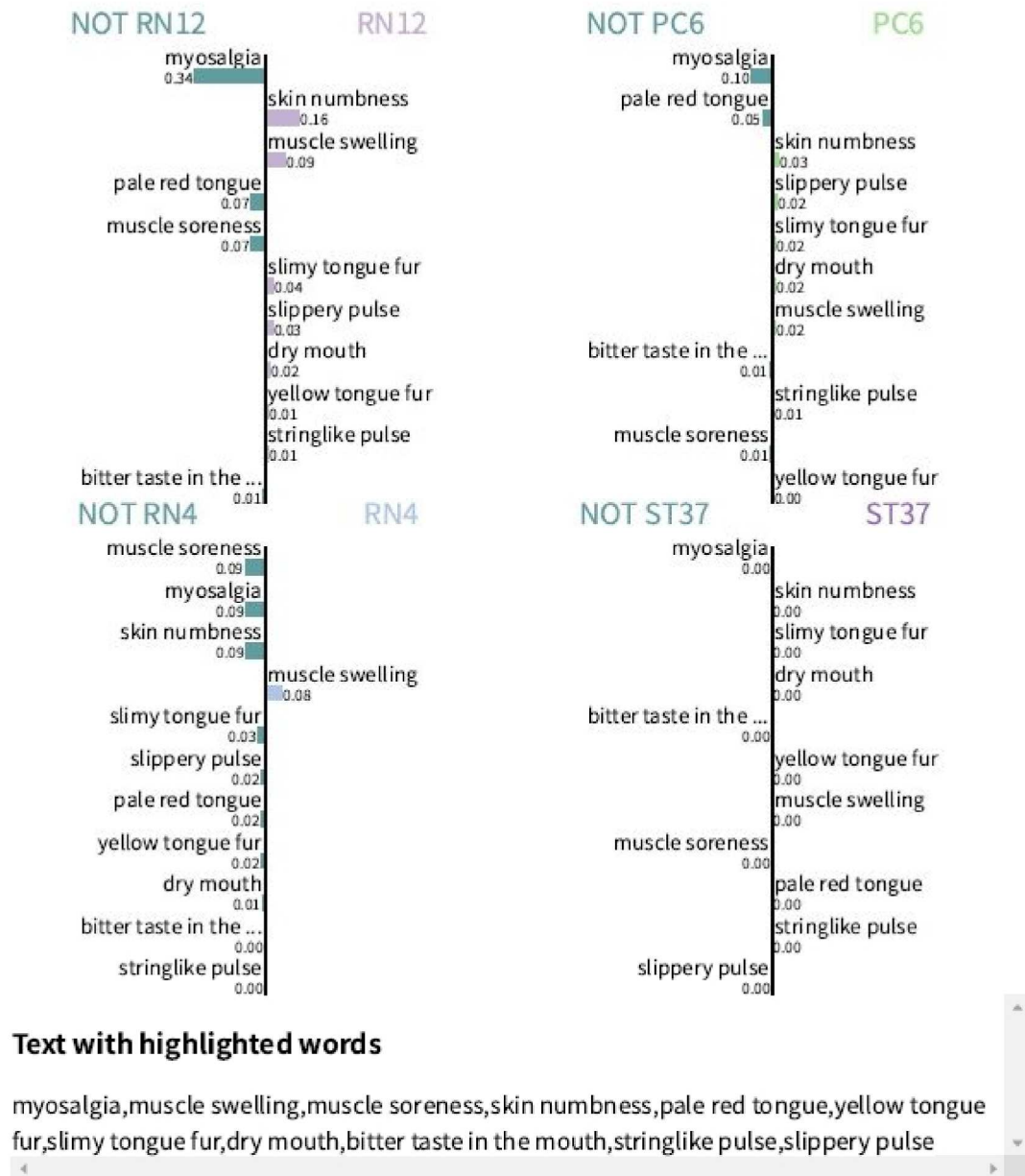


Figure 5.12: LIME Visualization of the Top 15 Predicted Acupoints for Case 1 (Part 3).

5.3.3 Case 2

A second test case is presented to further assess the system's diagnostic accuracy and interpretability.

- Input Symptoms:
cough, rapid pulse, fever, red or crimson tongue, thin fur, dry mouth, dry nose, nosebleed, dry throat
- Target Labels from Database:
 - Level 1 (Disease Category): epistaxis
 - Level 2 (Subtype): epistaxis (evil heat invading the lung)
 - Level 3 (Acupoint Prescription): SP4, SP1, SP3, LU6, LI4, LI20, DU23, LU11

Figures 5.13–5.21 present the system's diagnostic results, recommended acupoints, and corresponding LIME explanations for Case 2. These results further demonstrate the system's robust performance across different clinical scenarios.

Enter patient symptoms

cough × rapid pulse × fever × red or crimson tong... × thin fur × dry mouth × dry nose × nosebleed × dry throat ×

Generate Recommended Prescription

Analysis of Symptoms Based on Models:

The possible disease names for this patient are: epistaxis

Prediction results of syndrome types:

The possible disease subcategory names for this patient are: epistaxis (evil heat invading the lung)

Prediction of acupoint prescription:

Based on the analysis of symptoms through the model, the appropriate acupoint prescription for this patient is: LU6, LU11, LI4, LI20, SP1, SP3, SP4, DU23

Generating LIME visualization analysis for you, please wait...

Figure 5.13: System Output for Case 2: Hierarchical Predictions Including Disease Name, Subcategory, and Recommended Acupoints. (Source: Created by the authors)

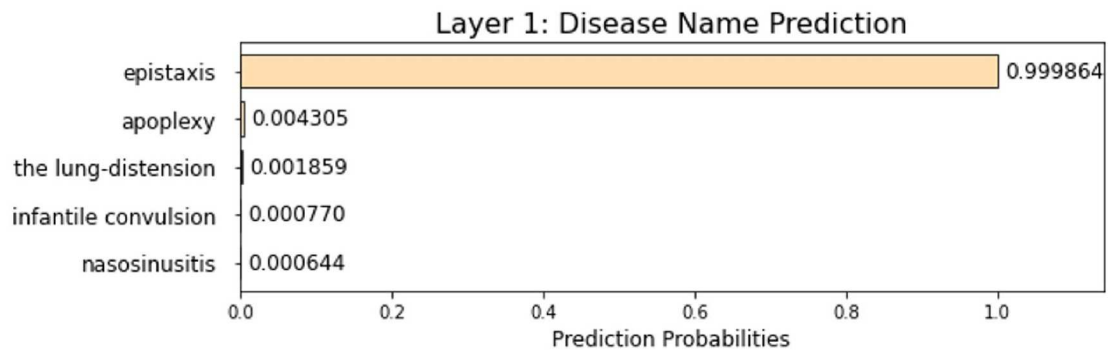


Figure 5.14: Top Prediction Probabilities for Disease Names - Case 2 (Level 1). (Source: Created by the authors)

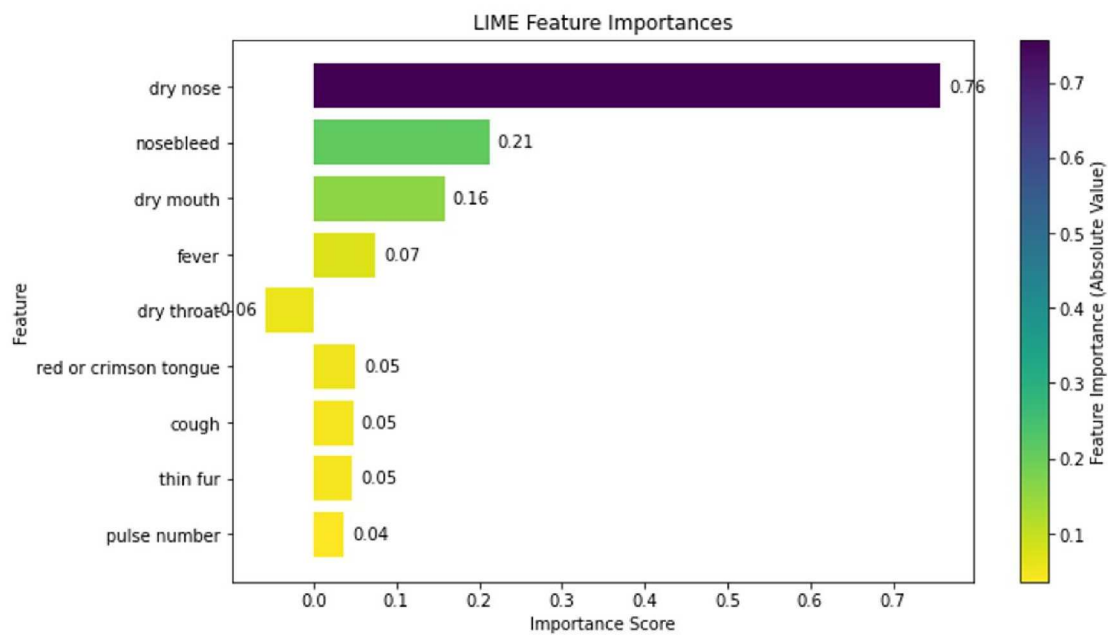


Figure 5.15: LIME Explanation for Top-1 Disease Name Prediction - Case 2 (Level 1). (Source: Created by the authors)

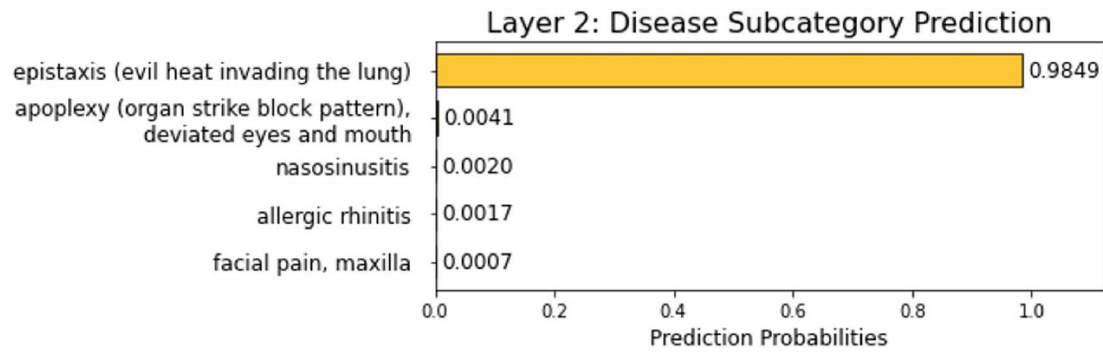


Figure 5.16: Top Prediction Probabilities for Disease Subcategories - Case 2 (Level 2). (Source: Created by the authors)

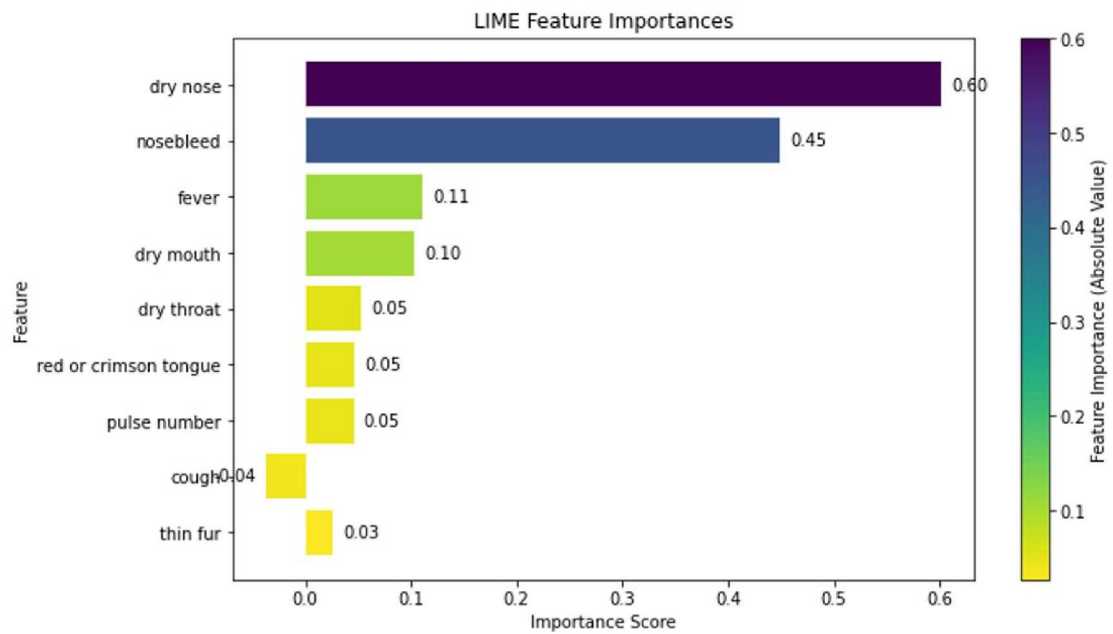


Figure 5.17: LIME Explanation for Top-1 Disease Subcategory Prediction - Case 2 (Level 2). (Source: Created by the authors)

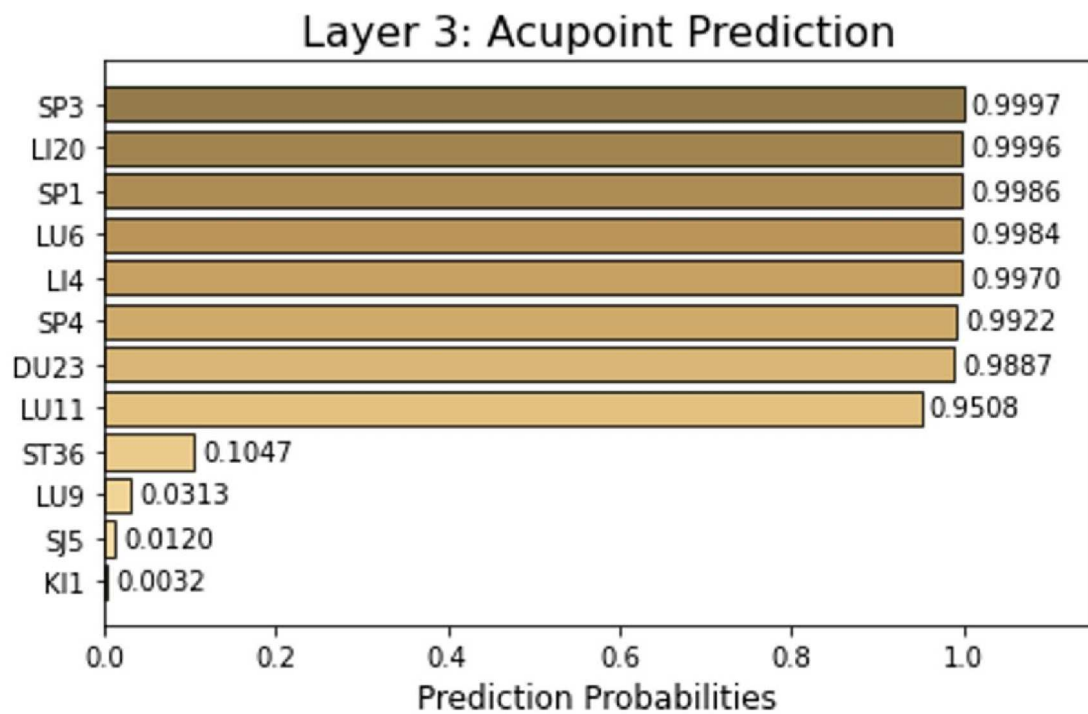


Figure 5.18: Top Prediction Probabilities for Acupoint Prescriptions - Case 2 (Level 3). (Source: Created by the authors)

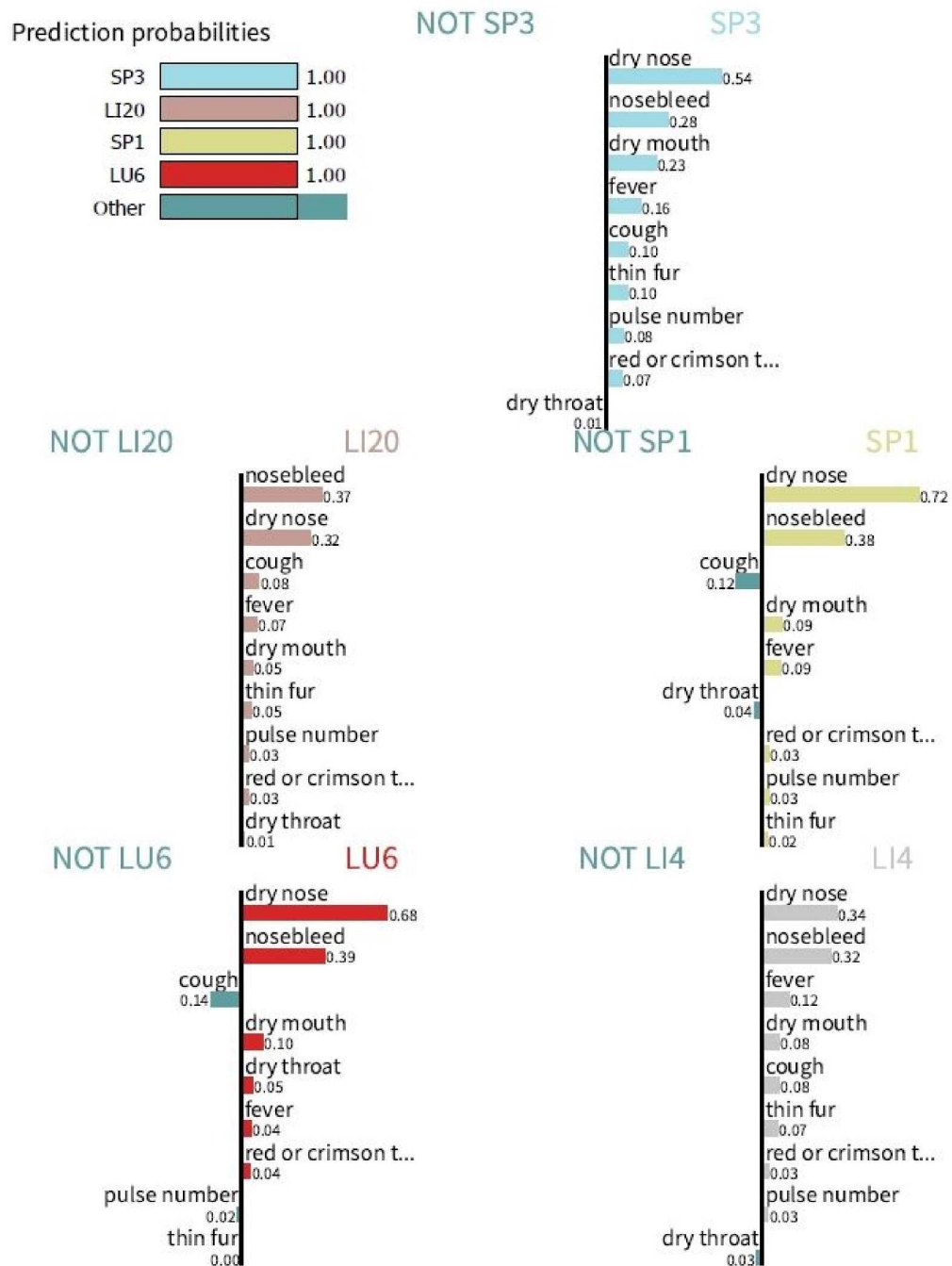


Figure 5.19: LIME Visualization of the Top 15 Predicted Acupoints for Case 2 (Part 1).



Figure 5.20: LIME Visualization of the Top 15 Predicted Acupoints for Case 2 (Part 2).

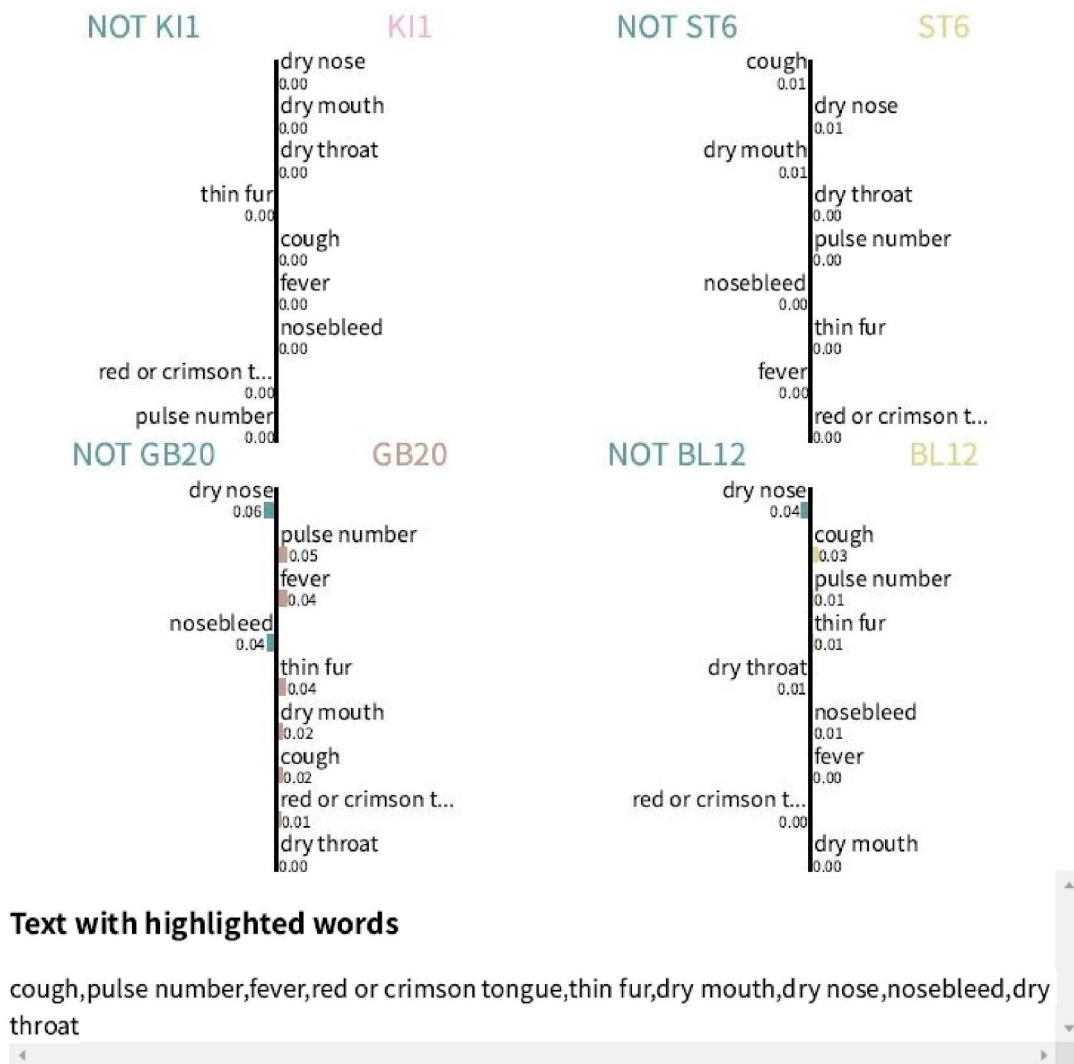


Figure 5.21: LIME Visualization of the Top 15 Predicted Acupoints for Case 2 (Part 3).

5.4 Conclusion

This chapter evaluated the performance of the proposed system using 1,000 original test cases and demonstrated its capabilities through two representative examples. The system predicts disease names, subcategories, and corresponding acupoint prescriptions based on patient symptoms. Experimental results show that it performs well in predicting disease names and acupoint prescriptions on the test set. By integrating explainability techniques such as LIME, the system enhances model transparency and supports clinical interpretability. In practice, the system operates within a physician oversight framework and continuously updates its model based on patient feedback, maximizing the advantages of AI while mitigating potential risks.

Despite these promising results, several limitations remain. The prediction performance for subcategories is relatively lower than that for disease names and acupoint prescriptions, indicating areas for improvement. AI models, especially those trained on historical data, may fail to recognize rare or atypical cases, which can lead to missed or incorrect diagnoses. Moreover, although the model demonstrates high predictive accuracy, the actual effectiveness of treatment recommendations remains heavily dependent on the quality and representativeness of the input data, which requires further refinement and continuous evaluation by clinical experts. Furthermore, regarding the interpretability of the proposed system, this study primarily validated the model’s rationale by comparing the LIME-extracted features with standard TCM diagnostic criteria (i.e., verifying that the model focuses on primary symptoms). However, a direct qualitative or quantitative evaluation of the system’s utility from the perspective of clinicians has not yet been conducted. Future work will involve designing user studies with TCM practitioners to evaluate the practical usefulness of the LIME explanations in real-world diagnostic scenarios. Metrics such as “trust in the model”, “reduction in diagnostic time”, and “interpretability satisfaction” will be assessed to bridge the gap between algorithmic explainability and clinical utility. Finally, in the current experimental environment, generating LIME-based explanations takes over 30 minutes per case, significantly hindering practical deployment. Enhancing the efficiency of interpretability modules is therefore critical for real-time clinical application.

In summary, the proposed system offers a promising step toward intelligent, interpretable support tools for acupuncture and moxibustion treatment planning. With continued validation and technical refinement, it has the potential to assist clinical

decision-making, improve efficiency, and promote more consistent and evidence-based therapeutic practices while maintaining the essential role of physician judgment.

Chapter 6

Conclusions

This dissertation explored the integration of machine learning methods into acupuncture and moxibustion treatment (AMT), aiming to enhance the objectivity, consistency, and clinical utility of acupoint prescription. The research progressed from exploratory applications of traditional machine learning models to the development of a novel hierarchical and interpretable framework, culminating in the design of a prototype decision-support system.

The main contributions of this work can be summarized as follows:

1. Systematic literature review: A comprehensive review of AMT fundamentals and the application of machine learning methods in this field was provided, highlighting both the opportunities and the challenges of bridging traditional Chinese medicine with modern computational approaches.

2. Exploratory applications of machine learning: Early attempts were conducted using traditional machine learning algorithms, ensemble methods, and deep learning. These studies revealed the feasibility of applying machine learning to AMT but also underscored key limitations, including difficulties in modeling the structured relationships among acupoints and the lack of interpretability required for clinical acceptance.

3. Hierarchical and interpretable model: To overcome these limitations, a novel hierarchical and interpretable machine learning framework was proposed for acupoint determination. This approach explicitly captured the hierarchical organization of AMT knowledge while incorporating interpretability mechanisms to provide transparent and clinically meaningful outputs. Experimental results demonstrated that the model achieved improved predictive performance and offered enhanced interpretability compared to baseline methods.

4. System design and implementation: Building upon the proposed model, a prototype decision-support system was designed to integrate the developed methods

into a practical clinical tool. This system highlights the translational potential of the research and provides a foundation for future clinical applications.

Despite these contributions, several limitations remain. The quality and scale of available AMT datasets constrained the scope of the experiments, and further efforts are needed to integrate more diverse and standardized clinical data. Additionally, while the proposed model improves interpretability, its effectiveness in real-world clinical practice requires validation through collaboration with practitioners and rigorous clinical trials.

Looking ahead, future research should focus on three directions. First, the collection and integration of large-scale, high-quality, and multi-source datasets will be essential for improving the robustness and generalizability of models. Second, extending the hierarchical and interpretable framework to incorporate multimodal information—such as patient demographics, textual medical records, and physiological signals—could further enhance its clinical relevance. Finally, strengthening collaboration with clinicians and embedding the system into real-world healthcare workflows will be critical for validating its practical utility and fostering trust in AI-assisted AMT.

In summary, this dissertation contributes to bridging traditional acupuncture and moxibustion practices with modern machine learning methodologies. By advancing both methodological innovation and system development, it provides a foundation for future research toward clinically reliable, interpretable, and personalized decision-support systems in AMT.

Bibliography

- [1] Wang XD, Li B, Ma YP, Cui Y, Yang XM, Li YX, et al. Acupuncture clinical hotspots and trends from 2013 to 2022: a bibliometric and visualized analysis. *Medicine (Baltimore)* 2025; 104(14): 1-14.
- [2] B.L. Zhang, X.M. Shi, and F.R. Liang, *Theory and Practice of Acupuncture & Moxibustion*, China Press of Traditional Chinese Medicine, 2019 (in Chinese).
- [3] Y.L. Wang, X.M. Shi, T. Efferth, and D. Shang, "Artificial intelligence-directed acupuncture: a review," *Chinese medicine*, vol.17, no.1, pp.1-10, 2022.
- [4] R.C. Deo, "Machine Learning in Medicine," *Circulation*, vol.132, no.20, pp.1920-1930, 2015.
- [5] J. Goecks, V. Jalili, L.M. Heiser, and J.W. Gray, "How Machine Learning Will Transform Biomedicine," *Cell*, vol. 181, no. 1, pp. 92-101, 2020.
- [6] A. Esteva, A.Robicquet, B. Ramsundar, et al, "A guide to deep learning in healthcare," *Nature Medicine*, vol. 25, no. 1, pp. 24-29, 2019.
- [7] A. Chaddad, Y. Jiang, and C. He, "OpenAI ChatGPT: A Potential Medical Application," 2023 IEEE International Conference on E-health Networking, Application & Services (Healthcom), Chongqing, China, pp. 210-215, 2023.
- [8] J. Liang, M.Y. Ming, C.B. Wang, X.L. Lv, Z.R. Sun, and H.N. Yin, "Research progress in the integration of machine learning and the science of acupuncture and moxibustion," *Acupuncture Research*, vol.46, no.6, pp.460-463, 2021.
- [9] W. Li, B. Li, Y. Guo, T. Zhao, X. Pan, and B. Chen, "Feasibility Analysis of Applying Data Mining Technology to the Researches of Major Problems in Acupuncture and Moxibustion," *Acta Chinese Medicine and Pharmacology*, vol.46, no.6, pp.6-9, 2018.

- [10] J. Yu, Y. Jiang, M. Tu, B. Liao, and J. Fang, "Investigating prescriptions and mechanisms of acupuncture for chronic stable angina pectoris: an association rule mining and network analysis study," *Evidence-Based Complementary and Alternative Medicine*, vol.10, pp.1-11, 2020.
- [11] X. Zhong, Y. Jia, D. Li, and X. Zhang, "Classification of acupuncture points based on the Bert model," *Journal of Data Analysis and Information Processing*, vol.9, no.3, pp.123-135, 2021.
- [12] P.H. Lu, J.L. Keng, F.M. Tsai, P.H. Lu, and C.Y. Kuo, "An Apriori Algorithm-Based Association Rule Analysis to Identify Acupoint Combinations for Treating Diabetic Gastroparesis," *Evidence-Based Complementary and Alternative Medicine*, vol.17, pp.1-9, 2021.
- [13] M.L. Zhang, and Z.H. Zhou, "A review on multi-label learning algorithms," *IEEE transactions on knowledge and data engineering*, vol.26, no.8, pp.1819-1837, 2013.
- [14] T. Wu, Q. Huang, Z. Liu, Y. Wang, and D. Lin, "Distribution-balanced loss for multi-label classification in long-tailed datasets," *European Conference on Computer Vision (ECCV)*, pp.162-178, 2020.
- [15] W. Huang, E. Chen, Q. Liu, Y. Chen, Z. Huang, Y. Liu, and S. Wang, "Hierarchical multi-label text classification: An attention-based recurrent network approach," *Proceedings of the 28th ACM international conference on information and knowledge management*, pp.1051-1060, 2019.
- [16] M.T. Ribeiro, S. Singh, and C. Guestrin, "Why should I trust you? Explaining the predictions of any classifier," *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*, pp.1135-1144, 2016.
- [17] Streamlit, "Streamlit-A fastest way to build and share data apps," [Online]. Available: <https://streamlit.io>. [Accessed: Apr.15, 2025].
- [18] A. Hyodo, *Traditional Chinese medicine meridians and acupoint textbooks*, Shin-sei Publishing, 2012 (in Japanese).
- [19] Japan-China joint editing, *Acupuncture and moxibustion "Basic"*, Toyo Academic Publishing (2013) (in Japanese).

- [20] G.R. Sun, H.X. Zheng, Basis s of theory of traditional Chinese medicine, Traditional Chinese Medicine publishing (2012) (in Chinese).
- [21] M.E. Maron, Kuhns, J. L., On relevance, probabilistic indexing and information retrieval, Journal of the ACM (JACM), 1960, 7 (3): 216-244.
- [22] T. Cover, P. Hart, Nearest neighbor pattern classification, IEEE transactions on information theory, 1967, 13 (1): 21-27.
- [23] L. Breiman, Friedman, J. H., Classification and regression trees, Routledge, 2017.
- [24] Cover, T. M., Geometrical and statistical properties of systems of linear inequalities with applications in pattern recognition, IEEE transactions on electronic computers, 1965, 3: 326-334.
- [25] McCulloch, W. S., W. Pitts, A logical calculus of the ideas immanent in nervous activity, The bulletin of mathematical biophysics, 1943, 5(4): 115-133.
- [26] G. Madjarov, D. Kocev, D. Gjorgjevikj, et al. An extensive experimental comparison of methods for multi-label learning. Pattern Recognition, 2012, 45 (9): 3084-3104.
- [27] M.R. Boutell, J. Luo, X. Shen, C. M. Brown, Learning multi-label scene classification, Pattern Recognition, 2004, 37 (9): 1757-1771.
- [28] J. Furnkranz, E. Hullermeier, E. Loza Mencia, K. Brinker, Multilabel classification via calibrated label ranking. Machine Learning, 2008, 73 (2): 133-153.
- [29] G. Tsoumakas, I. Vlahavas, Random k-labelsets: An Ensemble Method for Multilabel Classification, European Conference on Machine Learning, Springer, 2007, 406-417.
- [30] M.L. Zhang, Z.H. Zhou, ML-kNN: A lazy learning approach to multi-label learning, Pattern Recognition, 2007, 40 (7): 2038-2048.
- [31] A. Elisseeff, J. Weston, A kernel method for multi-labelled classification, Advances in neural information processing systems, 2001, 14: 681-687.
- [32] M.L. Zhang, K. Zhang, Multi-label learning by exploiting label dependency, In: Proceedings of the 16th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD' 10), Washington, D. C., 2010, 999-1007.

- [33] M.A. Ganaie, M. Hu, A.K. Malik, M. Tanveer, and P.N. Suganthan, "Ensemble deep learning: A review," *Engineering Applications of Artificial Intelligence*, 115, 105151, 2022.
- [34] Breiman and Leo, "Random forests," *Machine learning*, vol. 45, pp. 5-32, 2001.
- [35] T. Chen and C. Guestrin, "XGBoost: A Scalable Tree Boosting System," *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining*, pp. 785-794, 2016.
- [36] A. Zangari, M. Marcuzzo, M. Rizzo, L. Giudice, A. Albarelli, A. Gasparetto, "Hierarchical Text Classification and Its Foundations: A Review of Current Research," *Electronics*, vol.13, no.7, pp.1-57, 2024.
- [37] D. Koller, and M. Sahami, "Hierarchically Classifying Documents Using Very Few Words," *Proceedings of the Fourteenth International Conference on Machine Learning*, pp.170-178, 1997.
- [38] C.N. Silla, and A. Freitas, "A survey of hierarchical classification across different application domains," *Data mining and knowledge discovery*, vol.22, pp.31-72, 2011.
- [39] K. Saito, *Deep Learning from Scratch 2*, O'Reilly Japan, Inc., 2018 (in Japanese).
- [40] X. Li, X. Han, S. Wei, Y. Liang, and R. Guan, "Acupuncture and tuina knowledge graph with prompt learning," *Frontiers in Big Data*, vol. 7, p. 1346958, 2024.
- [41] Y. Liu and Y. Zhang, "Research on the association between acupuncture points and diseases based on the construction of knowledge graph and applied research," *Journal of Combinatorial Mathematics and Combinatorial Computing*, pp. 3355-3373, 2025.
- [42] X. Y. Yang and C. S. Ding, "SMRGAT: A traditional Chinese herb recommendation model based on a multi-graph residual attention network and semantic knowledge fusion," *Journal of Ethnopharmacology*, vol. 315, p. 116693, 2023.
- [43] S. Lundberg, "A unified approach to interpreting model predictions," *Proceedings of the 31st Conference on Neural Information Processing Systems*, pp.1-10, 2017.

- [44] A. Jiang, J. Li, L. Wang, W. Zha, Y. Lin, J. Zhao, Z. Fang, and G. Shen, “Multi-feature, Chinese–Western medicine-integrated prediction model for diabetic peripheral neuropathy based on machine learning and SHAP,” *Diabetes/Metabolism Research and Reviews*, p. 3801, 2024.
- [45] Z. Zhang, D. Jiang, Q. Wang, and P. Wang, “Construction of an interpretable machine learning-based prediction model for the clinical effect on ischemic stroke in treatment with eye acupuncture combined with rehabilitation therapy,” *Zhongguo Zhen Jiu*, vol. 45, no. 5, pp. 559-567, 2025.
- [46] L. Zhu, S. He, W. Zheng, Y. Tong, and F. Yang, “Predictive model of ulcerative colitis syndrome with ensemble learning and interpretability methods,” *Scientific Reports*, vol. 15, no. 1, pp. 1-12, 2025.
- [47] M. Frean, “The Upstart Algorithm: A Method for Constructing and Training Feedforward Neural Networks,” *Neural Computation*, vol. 2, no. 2, pp. 198-209, 1990.
- [48] Y. Kim, “Convolutional Neural Networks for Sentence Classification,” *Proc. EMNLP2014*, pp. 1746-1751, 2014.
- [49] I. Sutskever, O. Vinyals, and Q.V. Le, “Sequence to sequence learning with neural networks,” *Advances in neural information processing systems*, pp. 1-9, 2014.
- [50] H. Rezatofighi, N. Tsoi, J.Y. Gwak, A. Sadeghian, I. Reid, and S. Savarese, “Generalized intersection over union: A metric and a loss for bounding box regression,” *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 658-666, 2019.
- [51] Z. Luo, *The Classification Research of the Symptomatic Units of TCM*, M.S. thesis, Shandong University of Traditional Chinese Medicine, 2012 (in Chinese).
- [52] R.M. Yan, *Essence of Yan Runming’s 60 years of clinical experience in acupuncture and moxibustion*, China Press of Traditional Chinese Medicine, 2013 (in Chinese).
- [53] S.Z. Gao and J. Yang, *Therapeutics of Acupuncture and Moxibustion*, China Press of Traditional Chinese Medicine, 2016 (in Chinese).

- [54] J. Nam, E.L. M., H.J. Kim, J. F., “Maximizing subset accuracy with recurrent neural networks in multi-label classification,” *Adv. Neural Inf. Process. Syst.*, pp.5413-5423, 2017.
- [55] P.C. Yang, X. Sun, W. Li, S. Ma, W. Wu, H.F. Wang, “SGM: Sequence Generation Model for Multi-Label Classification,” *Proceedings of the 27th International Conference on Computational Linguistics*, pp. 3915-3926, 2018.
- [56] W.Z. L, Y. Wang, Y.C. Yin, X.B. Zhang, P. Ma, “Improved sequence generation model for multilabel classification via CNN and initialized fully connection,” *Neurocomputing*, vol.382, pp.188-195, 2020.
- [57] Greff, Klaus, et al., “LSTM: A Search Space Odyssey,” *IEEE Transactions on Neural Networks & Learning Systems*, vol. 28, no. 10, pp. 2222-2232, 2016.
- [58] Maron ME, Kuhns JL, “On relevance, probabilistic indexing and information retrieval,” *Journal of the ACM (JACM)*, vol. 7, no. 3, pp. 216-244, 1960.
- [59] Quinlan JR, “Induction of decision trees,” *Machine learning*, vol. 1, pp. 81-106, 1986.
- [60] Yao L, Mao CS, Luo Y, “Graph Convolutional Networks for Text Classification,” In *Proceedings of the 33rd AAAI Conference on Artificial Intelligence (AAAI 2019)*, pp. 1809.05679, 2019.
- [61] Chen ZM, Wei XS, Wang P, Guo Y, “Multi-label image recognition with graph convolutional networks,” In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 5177-5186, 2019.
- [62] Cui Y, Che W, Liu T, Qin B, Yang Z, “Pre-training with whole word masking for Chinese BERT,” In *Proceedings of the IEEE/ACM Trans Audio Speech Lang Process*, pp. 3504-3514, 2021.
- [63] Zhang QM, Zhang ZZ, Ki M, “Symptom names of Chinese medicine as scientific and technological terms,” *Journal of Beijing University of Traditional Chinese Medicine*, vol. 30, no. 12, pp. 797-799, 2007 (in Chinese).
- [64] L. van der Maaten and G. Hinton, “Visualizing data using t-SNE,” *Journal of Machine Learning Research*, vol. 9, pp. 2579–2605, 2008.

-
- [65] Devlin J, Chang M-W, Lee K, Toutanova K, “BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding,” NAACL-HLT, vol. 1, pp. 4171-4186, 2019.
 - [66] Yang H, Wu R, Nakata M, Ge QW, “On Application of Machine Learning for Deciding Acupoints in Acupuncture and Moxibustion Treatment,” IEIE Transactions on Smart Processing and Computing, vol. 14, no. 2, pp. 152-164, 2025.