

The Coevolution of Human and Machine learning

by Siqi Zhang

Thesis submitted in fulfilment of the requirements for
the degree of

Doctor of Philosophy

under the supervision of Prof. Yang Wang, A/Prof. Zhidong
Li and Prof. Richard Xu

University of Technology Sydney
Faculty of Engineering and Information Technology

06/2023

CERTIFICATE OF ORIGINAL AUTHORSHIP

I, Siqi Zhang, declare that this thesis is submitted in fulfilment of the requirements for the award of Doctor of Philosophy, in the Faculty of Engineering and information Technology at the University of Technology Sydney.

This thesis is wholly my own work unless otherwise referenced or acknowledged. In addition, I certify that all information sources and literature used are indicated in the thesis.

This document has not been submitted for qualifications at any other academic institution.

This research is supported by the Australian Government Research Training Program.

Signature: Production Note:
 Signature removed prior to publication.

Date:05/30/2023

ABSTRACT

Machine learning has made remarkable progress in the past decade, and its application scope and depth continue to expand. However, it also faces significant crises and challenges. Firstly, although machine learning has high accuracy for model output, they rely heavily on large amounts of annotated data. In situations with limited data or no annotations, ensuring the accuracy for the model output becomes a significant goal for specific industries. Secondly, machine learning applications often encounter scenarios with insufficient data or low quality data. When such data is used to train models, it can lead to significant deviations, resulting in inaccurate output and a decline in people’s trust in machine learning. Thirdly, many machine learning models operate in a black box environment, where the models often do not provide explanations or the explanations provided are too complex. Without appropriate feedback, humans cannot understand the learning status of the model and cannot effectively intervene in the model’s learning. Therefore, it is difficult for machine learning to gain human trust. This issue can have a direct impact on the application and advancement of machine learning.

To address these challenges, under the guidance of my supervisors, I conducted research on the coevolution of human and machine learning. Our research goal is to enable human and machine learning models to progress together, achieving better performance, gaining people’s understanding and trust ,and then applying it. To improve the accuracy of output results in situations with limited data, we introduced domain knowledge to jointly train the model. For scenarios with low-quality data, we proposed a multi-model structure that fosters interaction and collaboration between models and experts. This approach allows experts to monitor and enhance model performance, boosting the trustworthiness of machine

learning models. To address situations where there is no explanation or unreasonable explanation, we developed an explainable machine learning framework that uses multimodal methods to clarify the output results of machine learning models for non-machine learning experts. This framework promotes the broader application of machine learning. Our specific work is as follows:

1. We proposed a Bayesian Nonparametric Process(BNP) method for adding rule-based domain knowledge, enabling effective training even with limited and unannotated training data. The utility function of domain knowledge was integrated into the model as a prior, forming a responsive BNP method that can quickly learn from input data and achieve excellent performance in limited data situations. We validated the proposed method on both a simulated dataset and a supermarket dataset, achieving outstanding results.

2. We proposed a framework for interactive collaboration between models and experts, continuously improving the performance of machine learning models, and enhancing trust in machine learning. The framework adopts a multi-model structure that leverages knowledge from source datasets, target datasets, and small datasets annotated by experts, aiding student models in knowledge extraction-based learning. Before releasing its output, the student model’s results are validated by experts, and in case of disagreement, the experts provide guidance to allow the student model to continue learning. The approach incorporates additional domain knowledge provided by experts to guide the model, distinguishing it from other learning methods. We tested our proposed method on a medical dataset, and it demonstrated exceptional performance.

3. We proposed an explainable multimodal information framework to make it easier for non-machine learning experts to comprehend and trust the model. The framework utilizes privileged information from text and expert annotation to aid in training student models, where experts use privileged text information to explain the model during the training process. Due to the privileged information framework, text information, or expert annotation is no longer required during

the testing process, making it more broadly applicable. We tested the framework in music classification and received overwhelming recognition for its music interpretation results, ranking it high among many state-of-the-art models in terms of performance.

Our research has made progress towards achieving the goal of coevolution between human and machine learning, addressing the challenges of trust in machine learning models and improving their performance. By reducing the dependence on training data, fostering human trust in the models, and establishing an interactive ecosystem between human and machine learning models, we can apply machine learning to more fields and advance research on machine learning theory further. Our work paves the way for coevolution between human and machine learning.

ACKNOWLEDGEMENTS

First and foremost, I am extremely grateful to my supervisors, Prof. Yang Wang, for his invaluable advice, continuous support, and patience during my PhD study. His immense knowledge and plentiful experience have encouraged me throughout my academic research and daily life. I also extend my special thanks to Prof. Zhidong Li for his all-round help during my PhD studies. He started to teach me related knowledge before I enrolled and continued to offer help whenever I needed it after I enrolled. In my student career, Yang and Zhidong are the best supervisors I could ask for.

I would also like to express my gratitude to Dr. Feng Zhou and Dr. Ling Luo for their technical support on my study. I am deeply grateful to Zhou Feng for guiding me like a half-teacher, giving me a complete understanding of how to conduct research, and regularly assisting me with revising my paper. When my paper was rejected, they still encouraged me to submit it to the next conference. When I was lost about my future, he encouraged me with his own experiences. He may be the best senior to me in the world.

I would like to thank all the members of the DSI, such as Yongzhe Chang, Yiyuan Zhang, Shuming Liang, Yangyang Shu, Zhilin Zhao, Boyu Li, Bowen Zhang, and Haoxiang Huang. It is their kind help and support that have made my study and life in Australia a wonderful time.

Finally, I would like to express my gratitude to my parents. Without their tremendous understanding and encouragement in the past few years, it would have been impossible for me to complete my studies. I am extremely grateful for the mental and material support they have given me.

Contents

Abstract	2
Acknowledgements	5
List of Figures	10
List of Tables	11
1 Introduction	12
1.1 Background	12
1.1.1 Motivation	12
1.1.2 Definition of the Coevolution of Human and Machine Learning	15
1.1.3 The importance of coevolution	17
1.2 Research Questions	20
1.2.1 Domain knowledge injection issue	21
1.2.2 Human and machine learning model interaction issue	22
1.2.3 Explaniable feedback issue	23
1.3 Thesis Contributions	25
1.3.1 Propose a BNP method for adding rule-based domain knowledge	25
1.3.2 Propose co-teaching accountable learning framework with distilled and domain knowledge	26
1.3.3 Propose an explainable framework for multimodal information	27
1.4 Thesis Outline	27
2 Literature Review	30
2.1 Injecting domain knowledge	30
2.1.1 Bayesian Nonparametric Model	30
2.1.2 Transfer Learning	32
2.2 Obtaining explanatory feedback	36
2.2.1 Explainable Machine Learning	36
2.2.2 Truthworthy machine learning	41
2.3 Facilitating interaction between humans and machine learning	42

2.3.1	Human in the loop for machine learning	42
2.3.2	Active Learning	43
2.3.3	Reinforcement Learning	43
2.4	Literature Conclusion	44
3	A BNP method for adding rule-based domain knowledge	45
3.1	Motivations	45
3.2	Preliminary Knowledge	47
3.2.1	Dirichlet Process	47
3.2.2	Chinese restaurant process	49
3.3	Methodology	51
3.3.1	Problem Definition	51
3.3.2	Preliminary for CRP Model	52
3.3.3	Utility Functions	54
3.3.4	Simultaneous Customer Segmentation and Utility Estimation Model	56
3.4	Inference: Gibbs Sampling for UtSeg model	57
3.5	Experiments	58
3.5.1	Experiment Setup	58
3.5.2	Synthetic Data Set	61
3.5.3	Case Study	61
3.6	Conclusion	63
4	Co-teaching accountable learning framework with distilled and domain knowledge	66
4.1	Motivations	66
4.2	Problem Formulation	68
4.3	Co-Teaching with Dual-Knowledge	70
4.3.1	Perception from Teacher Model	70
4.3.2	From Explanation to Feedback	72
4.3.3	Knowledge Passing from Expert	76
4.3.4	Optimization	76
4.4	Experiments	78
4.4.1	Performance of Co-teaching	79
4.4.2	Co-teaching for Abstruseness	80
4.4.3	Co-teaching for Accessibility	81
4.5	Conclusion	81

5	An explainable framework for multimodal information	87
5.1	Motivations	87
5.2	Emotion Recognition and Musical Domain knowledge	91
5.2.1	Emotion Recognition	91
5.2.2	Musical Domain knowledge:5 Musical Dimension	92
5.2.3	Main Meledy of Music	95
5.3	Preliminary Knowledge	96
5.4	Model	97
5.4.1	Problem Formulation	97
5.4.2	Learning with Explainable Privileged Information via Mul- timodal Distillation	98
5.5	Experiment	99
5.5.1	Experiment Setting	99
5.5.2	Data preprocessing	101
5.5.3	Experiment 1:The level of acceptance for different explana- tions	103
5.5.4	Experiment 2: Verify whether the melody labeling method is effective	105
5.5.5	Experiment 3: Compare performance with other SOTA models	105
5.6	Conclusion	105
6	Conclusion and Future Work	107
6.1	Conclusion	107
6.1.1	A BNP method for adding rule-based domain knowledge .	108
6.1.2	Co-teaching accountable learning framework with distilled and domain knowledge	108
6.1.3	An explainable framework for multimodal information . .	108
6.2	Future Work	109
	CERTIFICATE OF ORIGINAL AUTHORSHIP	112
	References	113
	Appendices	121

LIST OF FIGURES

1.1	Coevolution of human and machine learning	17
1.2	Coevolution framework of human and machine learning	17
1.3	Research question	21
1.4	Thesis Outline	29
3.1	Example of CRP	51
3.2	The flow chart and graphic model of our framework.	53
3.3	Utility functions for different data sets: (a)-(d) are for synthetic data, (e)-(x) are for four product types. 5 utility functions and observed points are shown for each type. The learned parameters α_k and β_k are under the plots.	64
3.4	Comparison results of model efficiency.	65
4.1	The general framework of our co-teaching	69
4.2	Loss function process	69
4.3	The example of using co-teaching framework on iteration j	75
4.4	Algorithm	75
4.5	Co-teaching for Accessibility	82
5.1	Model	91
5.2	Emtional	92
5.3	Musical Dimensions	95
5.4	Learning Algorithm	99
5.5	Loss function	99
5.6	MFCC	102
5.7	Contents of the questionnaire	103
5.8	Result Of Questionnaire	104

LIST OF TABLES

3.1	Evaluation results on synthetic data.	60
3.2	Evaluation results on real data.	62
4.1	Performance of Co-teaching(different part of co-teaching)	84
4.2	Performance of Co-teaching(baseline models)	85
4.3	Co-teaching for Abstruseness	86
5.1	Results of Experiment 2 and Experiment 3	106

Chapter 1

Introduction

In this chapter, we briefly introduce the background of coevolution of human and machine learning, related challenges and questions, thesis contributions, and finally show the framework of the entire thesis.

1.1 Background

1.1.1 Motivation

Machine learning currently plays a crucial role in assisting human decision-making in many fields. However, the theory and technology of machine learning are not mature enough, which causes low accuracy and low reliability in some application scenarios. The current main challenges with machine learning are as follows:

The first one is the theoretical flaws in machine learning decision-making mechanisms: Currently, machine learning often establishes associations between input data and expected results. Due to the limitations or biases that commonly exist in data samples, such association learning inevitably learns a false relationship. Models that rely on this as their decision-making basis may perform well on most test data, but in fact, they have not learned the ability to make reasoning decisions based on correct causal relationships. When faced with situations that are inconsistent with the training samples, their performance will greatly deteriorate. In order to further identify true causal relationships from probabilistic associations that may contain false relationships, it is necessary to introduce human intervention or other methods to clarify and strengthen the intrinsic causal relationships of intelligent decision-making and improve the accuracy and credi-

bility of models [1, 2].

The second one is that defects in the application of machine learning: In practical applications, machine learning systems obtained through data-driven learning have many hidden dangers and may cause serious social problems. Firstly, the limitations and biases in the collection of data samples lead to biased machine learning systems, similar to biases in human society. For example, the COMPAS crime risk assessment algorithm used by the Chicago court was found to discriminate against black criminal suspects [3]. Secondly, machine learning black box models often make some low-level errors that human cannot make, resulting in potential security risks. Finally, from the perspective of decision-making mechanisms, the analysis of current machine learning algorithms is still in an opaque exploration stage, and its decision-making process has not been clearly explained academically, making it temporarily impossible to obtain human understanding and trust, and its application poses potential risks.

The third one is the failure of machine learning systems to meet regulatory requirements: In fields such as finance, healthcare, and law, governments around the world are gradually strengthening legislation for risk prevention and regulation of artificial intelligence applications. For example, the European Union’s ”Ethical Guidelines for Trustworthy Artificial Intelligence” [4] states that trustworthy AI systems must meet seven requirements: human oversight and error correction, technical safety and robustness, privacy protection and data governance, transparency and explainability, algorithmic fairness and non-discrimination, environmental and societal impact, and accountability. The US FDA has issued the “Action Plan for Artificial Intelligence/Machine Learning-Based Software as a Medical Device,” proposing good machine learning practice to improve product transparency and strengthen regulation of algorithm bias and robustness. The regulatory requirements for artificial intelligence in various countries have been gradually improved at the legal and regulatory level, but how to implement these rules, systems, and regulations into feasible technical solutions is the challenge

we face.

The best way to address the above challenges of machine learning at the current stage is to involve human in machine learning system. By leveraging human capabilities in recognizing causal relationships and strong reasoning abilities, the theoretical limitations of machine learning decision mechanisms can be mitigated. Human judge whether the results meet the requirements based on the feedback information from the model. If the requirements are not met, human will provide the appropriate domain knowledge to the model for further learning. If the requirements are met, the model will output the results. This approach effectively avoids result biases caused by limitations and biases in the training data, and prevents certain basic mistakes that black box models can make but human can avoid. By leveraging human sensitivity to rules and regulations, machine learning systems can ensure compliance with regulatory requirements. On the other hand, human intervention in the model [5] can enhance the efficiency, credibility, and user experience of the model, which in turn promotes the application and development of machine learning technology. After human intervention in machine learning, through the interaction between human and the machine learning model, the integrated advantages of both can effectively address the aforementioned challenges. In order to achieve these goals, this paper proposes a coevolution definition and framework for human and machine learning, as detailed in Section 1.1.2.

Additionally, Harari [6] proposes that humanity has entered the era of big data, where we face vast amounts of data (measured in units of at least petabytes, exabytes, or zettabytes), diverse types (including web logs, audio, video, images, geolocation information, etc.), low value density (requiring deep mining), high speed, and time sensitivity. Without the support of machine learning, human are unable to make timely decisions. For example, Macy’s real-time pricing system in the United States requires real-time price adjustments of up to 73 million items based on demand and inventory. If done solely by human, it not only

requires a lot of manpower and time, but also cannot achieve real-time price adjustments. In the era of big data, human need to collaborate with machine learning to accomplish tasks. By utilizing machine learning to make informed decisions from large datasets, human can improve their decision-making abilities even with limited information.

1.1.2 Definition of the Coevolution of Human and Machine Learning

In this thesis, the concept of coevolution is derived from biology. In biology, coevolution is a process in which two or more species interact and evolve together over time. This process occurs when the characteristics of one species affect the evolution of another species, and vice versa [7]. For example, some predators and prey evolve together, with the prey developing better camouflage, speed or agility to avoid being caught, while predators develop sharper teeth, better eyesight or faster running speed to catch their prey.

In this thesis, the coevolution of human and machine learning can be defined in three levels, as follows: (1) The coevolutionary system for human and machine learning consists of human (users, decision makers, developers, etc.), models, input channels, and explanation channels. (2) Human act as the leaders in this process. They input information (domain knowledge, rules, etc.) into the model through the input channels and receive feedback information (reflecting the decision-making process) from the model through the explanation channels. The models act as the executors, receiving information from human through the input channels, running and making decisions, and sending explanations back to human through the explanation channels. (3) Continuous interactive process between human and models occurs through input and explanation information. This interaction aims to enhance human' understanding and trust of the model's decision-making process, while human continually provide suitable information to the model, enabling the model's output to approach the desired values. hu-

man and models work together, efficiently completing tasks. The entire process is depicted in Figure 1.1.

The evolution of the model is that the model gets the appropriate data and outputs results closer to the expected value after learning, and the feedback is easier to understand. The evolution of human is that human receive explainable information feedback by the model, deepen their understanding of the model's decision-making process, and provide more accurate data for the model.

The framework for the co-evolution of human and machine learning is a technical solution for achieving the co-evolution of human and machine learning, as shown in Figure 1.2. This framework is a closed-loop regulation system that ensures that the model output continuously approaches the expected value. In the framework, the human is the leader and regulator. Human adjusts the domain knowledge given to the models by understanding the model explanation information feedback. The model receives this information and let output quickly approaches the expected value through training. Since human are involved, the output of the framework is reliable and can effectively avoid theoretical defects in machine learning decision-making mechanisms, limited or low-quality learning data, and other application defects, and meet government monitoring requirements. The working process of this framework is as follows: human input domain knowledge (industry knowledge, rules, etc.) into the model through the input channels. The model starts learning and produces learning results while also providing model explanations to human. These explanations help human understand the model's decision-making process. human then input suitable domain knowledge to the model, and the model continues learning and producing results that are closer to the expected values. This iterative process continues until the model's output reaches the expected values.

To carry out the application of the Coevolution of human and machine learning, it is necessary to solve three key technologies: domain knowledge injection, model explainability, and human-model interaction. For details, see Chapters 3,

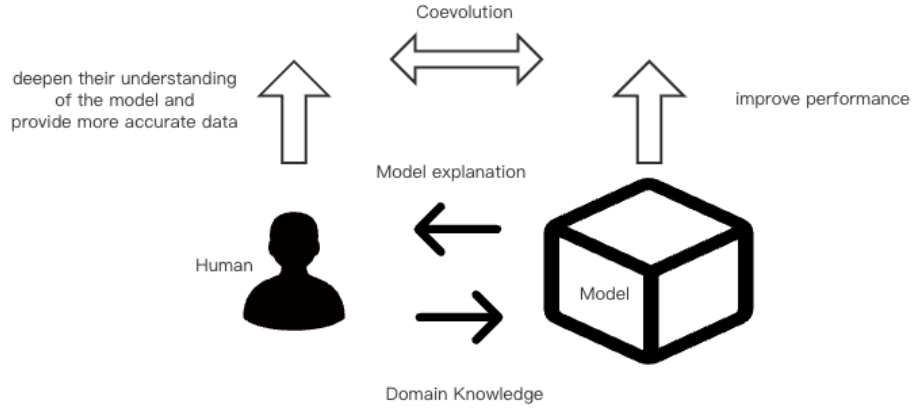


Figure 1.1: Coevolution of human and machine learning

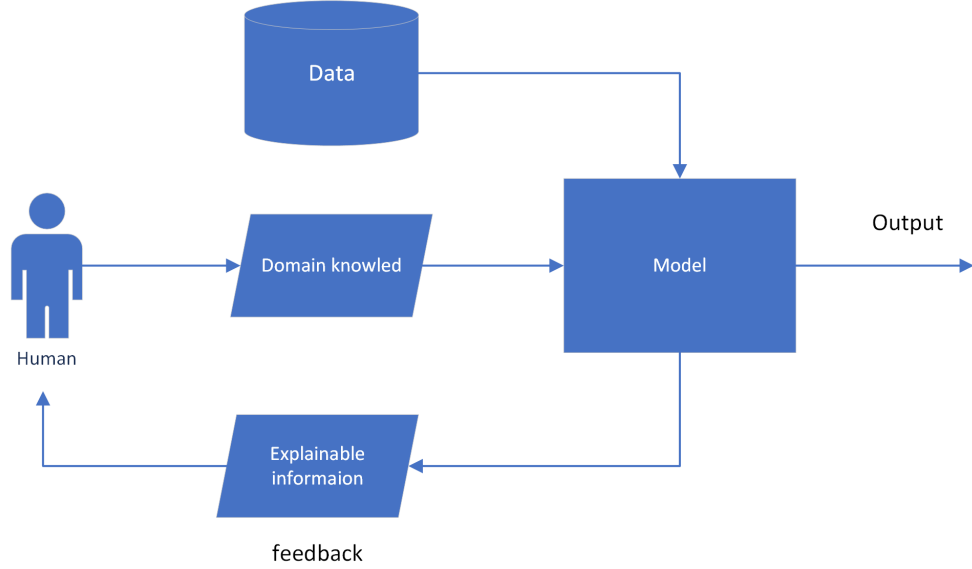


Figure 1.2: Coevolution framework of human and machine learning

4, and 5 of this paper.

1.1.3 The importance of coevolution

In the co-evolution of human and machine learning, human play a major role in: (1)annotating data, providing domain knowledge, rules, etc., which provide high-quality data for the model, and the model output results quickly and accurately based on this data. (2)adjusting the model and its inputs based on the explainable feedback provided by the model, to make the model's results approach the expected value. (3)testing and validating the model by rating its

outputs, especially in areas where the algorithm is not confident in its judgment or is too confident in incorrect decisions. The coevolution of human and machine learning involves feeding back these training, adjustment, and testing tasks into the algorithm. It leverages machine learning’s ability to make informed decisions from vast datasets and human’s ability to make decisions with limited information. This coevolution enhances the algorithm’s intelligence, reliability, and accuracy. This approach is particularly effective when the model needs to determine its next steps and sends data to human annotators for training. The coevolution can be applied to various artificial intelligence projects, where human intelligence is integrated to some extent, benefiting the machine learning-powered AI.

As one of the core areas of AI, machine learning is devoted to studying how to improve the performance of a system by using experience through computational means. Mitchell once proposed a famous definition of machine learning [8]: machine learning problems = tasks + objective functions + experience, where experience refers to the data collected to solve the learning task. This reflects an important characteristic of machine learning methods compared to other computer algorithms, which can automatically improve the performance of computer systems in specific tasks (defined by the objective function) by using external data (experience). Traditional machine learning tasks are usually an iterative process. After the data collection phase is completed to collect suitable training data, the learning algorithm repeatedly alternates between the model training phase and the model evaluation phase, eventually obtaining the model needed to solve the task. This process is closed, and the iterative alternation of model training and model evaluation is a fully automatic process that is often a black box for external users and experts. In other words, pure data-driven machine learning methods are a closed system for human, without a mechanism for human interaction with the machine learning model, and the system relies solely on data for learning.

However, as machine learning applications continue to explore, the difficulty of learning tasks continues to increase, and the closed learning paradigm driven by

pure data faces enormous challenges. First, real-world applications require machine learning to have the ability to adapt to open dynamic environments, which is becoming increasingly demanding. In his keynote report at the AAAI conference in 2016, Dietterich emphasized that AI methods should be more robust, that is, they should improve their ability to deal with dynamic changes in the environment and system failures [9]. In particular, he stressed that AI systems need to be able to deal with 'known unknowns' and 'unknown unknowns' in the environment. However, it is more difficult for a closed black box learning system to have such capabilities. Second, corresponding to the big data learning paradigm based on massive data, the efficient learning ability of human, such as small sample learning and continuous learning, has become a research hotspot [10, 11]. Obviously, the learning paradigm based solely on data will be difficult to get rid of the constraints of statistical rules such as the law of large numbers and will not be able to achieve efficient learning abilities similar to human. Therefore, it is necessary to consider introducing human intelligence into machine learning. Third, in the future machine learning ecology, different learning tasks and the models learned will not be isolated from each other. Knowledge transfer between learning tasks and model reuse across tasks will become a common phenomenon, making learning models like software components. This requires that the model is not a black box learned from a closed process, but an open model with explainability, transferability and human friendliness.

In summary, pure data-driven learning methods will not be able to meet the ever-changing demands of real-world applications. Breaking the closedness of machine learning systems is an inevitable research direction. Among them, the openness of the learning process to human is a crucial factor. Only when the learning system can fully utilize the expert knowledge provided by human can it break the pure dependence on data and achieve efficient learning. Only when the learning system can fully understand user feedback can it better achieve the goal of serving human needs. The history of scientific and technological develop-

ment tells us that the development of a scientific and technological field always accompanies the negation of negation, and is a process of spiral advancement. At the beginning of machine learning research, its original intention was to unleash the power of data to reduce the system’s dependence on expert knowledge. For the application of machine learning in fields with strict risk control such as finance, healthcare, and autonomous driving, trustworthy and explainable machine learning is the trend. Trustworthy and explainable machine learning needs to once again bring human involvement into the process of machine learning, and leverage the advantages of the human-machine interactive learning paradigm. Compared with the pure data-driven learning paradigm, the framework design and theoretical analysis of interactive learning pose unique challenges. Therefore, research on the coevolution of human and machine learning is of great significance for the development of machine learning.

1.2 Research Questions

The thesis focus on how to interact and evolve between human and machine learning models through domain knowledge provided by human and explainable information provided by the model under conditions of limited or poor-quality data, in order to improve the performance of machine learning models, make people understand and trust machine learning models more and be willing to use them to enhance productivity.

In machine learning applications, we often encounter the following challenges:

1. How to utilize accumulated domain knowledge to support model learning and improve model performance.
2. How to use interaction between human and machine learning models to improve the performance of model output results.
3. How to establish model explainability, reveal the mechanism behind the model’s predictions (decisions), promote human-model interaction, and help human understand and trust the model.

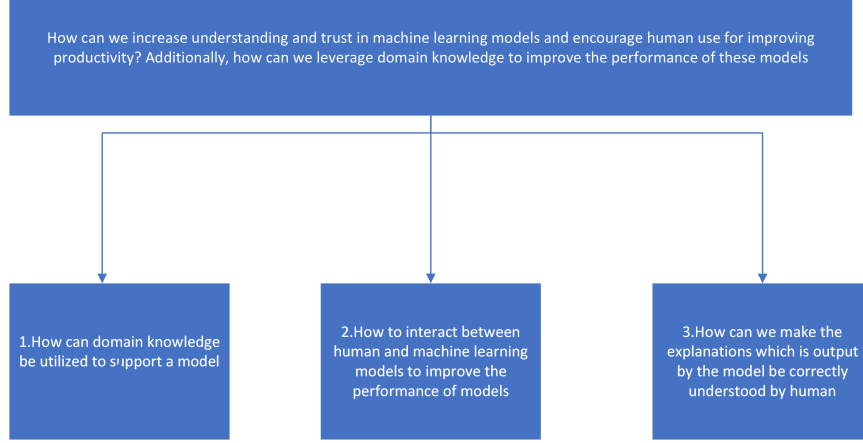


Figure 1.3: Research question

In this thesis, we have specifically proposed three detailed questions in Fig1.3.

1.2.1 Domain knowledge injection issue

Some machine learning applications suffer from the issues of limited data or low quality data. Therefore, relying exclusively on data from machine learning applications for learning purposes often leads to unsatisfactory results. In this case, the expert improves the output of the model by injecting a certain amount of accumulated domain knowledge into the model. In Chapter 3, we conducted a study on how can domain knowledge(Expert-provided knowledge) be utilized to support a model using supermarket customer segmentation as an example. The specific research points are as follows:

1.Collection of Demographic Data: When collecting user data, the more data the better. However, some users are worried about their personal privacy being leaked and are unwilling to provide relevant personal data. This is a challenge in collecting demographic data in a way that balances the need for completeness and the privacy of customers. The lack of key attributes in the demographic data can lead to unreliable segmentation.

2.Joint Estimation of Utility Functions and Customer Segmentation: Traditional machine learning classification either relies on a pre-set number of classes determined by human or depends on the type of utility function, which may not

necessarily align well with the data’s characteristics. The challenge in this setting is to determine the utility functions before customer segmentation, whereas the framework presented in the paper jointly estimates the segmentation of customers and the form and parameters of the utility functions at the same time.

3. Inference in the Bayesian Non-parametric Framework: Optimizing the utility function is a challenging problem that requires modeling a stochastic process. Previous work used a complex mathematical model that was difficult to design for compatibility, robustness, and scalability. To overcome these issues, we proposed an approximation model that maintains flexibility in modeling utility functions with different parameters and types.

4. Inconsistent Parameters: The parameters for different types of utility functions are inconsistent in both their prior and likelihood, due to the diversity of their meanings. This inconsistency in posteriors can cause a significant amount of heuristic work and requires domain knowledge. To address this, we propose a re-parameterization solution that wraps up the identity of the utility function with a parameter-free nonlinear function to maintain consistency of parameters. This design improves the generalizability and explainability of the model.

1.2.2 Human and machine learning model interaction issue

Machine learning is the process of fitting data to model with the aim of not just accurately predicting the finite training set, but also correctly predicting samples that have not appeared in the training set. The ability of a model to make predictions on data outside the training set is referred to as its generalization ability, and pursuing this ability is the goal of machine learning. Due to the mismatch between the learning capacity of models and the complexity of data, problems such as overfitting, underfitting, long training time, and low precision can arise. In the application of machine learning, it is common to encounter limited training sets, insufficient or low-quality training data, which results in unsatisfactory

model outputs [12]. To address this problem, we incorporate human intervention into the model’s learning process, improving its performance through multiple interactions between human and the model. Human utilize the feedback from the model to adjust the learning process and domain knowledge input in a timely manner, leveraging their ability to infer accurate knowledge from limited samples. The aim is to minimize learning time, maximize the accuracy of model outputs, and improve the robustness of the model. The specific research points are as follows:

1. Knowledge Disproportion: We note that the impact of supportive data and domain knowledge needs to be considered evenly. However, the loss-based learning mechanism tends to give more weight to the knowledge from supportive data, which is usually more voluminous, than the limited domain knowledge provided by experts. To address this, we propose a combined loss function of few-shot learning to balance both knowledge sources.

2. Abstruseness: This challenge refers to the issue of experts not knowing how well the model has been trained and not being able to interpret the model’s status during training. We propose the use of an explainable component to address this.

3. Accessibility: The authors highlight that the prevailing autonomous learning process does not provide an interface for experts to provide feedback. The proposed framework is designed to be accessible to experts by allowing them to access the training process and provide feedback easily.

1.2.3 Explainable feedback issue

The explainability of machine learning refers to the feedback information that the model provides along with its output results. The explainable information should reflect the decision-making mechanism of the model, allowing people to understand the results of the model. However, due to the professional nature of this information, it can be too obscure and difficult to understand for non-machine learning experts. This situation can lead to non-machine learning experts not

understanding and not trusting the output results of the model. Good explainable information can help people understand the decision-making mechanism of the model, correctly judge whether the results of the model are good or not, and people can also inject appropriate domain knowledge into the model based on their understanding to improve the results of the model. This directly impacts the application and promotion of machine learning. In Chapter 5, we conducted a study on what strategies can be used to understand and build trust in a model among non-machine learning experts using musical emotion recognition as an example. The specific research points are as follows:

1. Explanation of machine learning: The first research challenge is establishing trust between machine learning models and musicians or music enthusiasts, as it can be difficult to determine which inputs are driving the models' decisions. In traditional machine learning, users can observe and correct the system's predictions, but the predictions are not typically explained to them. Users may have trouble trusting a prediction if it is accurate, but the explanation provided is unclear or inaccurate. Explanation interactive machine learning addresses this challenge by allowing experts to interactively query the system and relabel data based on the prediction and explanation, if needed. However, there is a significant gap between the explanations provided by machine learning models and what musicians or other non-machine learning expert can understand. Machine learning explanations are often too complex for non-specialists to comprehend.

2. The diversity and uncertainty of music: The second challenge is the diversity and uncertainty inherent in the art of music, as described in Smith's work on variability [13]. The diversity of music stems from the fact that people from different backgrounds and cultures can express their unique perspectives and experiences through music [14]. For example, various musical styles can be applied to the same melody, and the same melody can evoke different emotions when played in different keys, such as C major and A minor. The uncertainty in music is especially prominent in absolute music, where there is no explicit story or

meaning conveyed by the music. Consider two pieces of music, one conveying positive emotions and the other conveying negative emotions. Artists can alternate between sad and happy music, and vice versa. Due to the diversity and uncertainty inherent in the art of music, both people and machine learning models face significant challenges.

3. Shortage of annotated data : The third challenge is the shortage of annotated data in the field of emotion recognition, particularly in music sound data and music background information. Annotated data is crucial for training robust predictive models for emotion recognition. The combination of different types of data, including complementary information from different modalities, can help improve the accuracy of the models. However, there may be certain conditions that make it impossible to have access to different modal information at all times, such as the difficulty in finding related sheet music or providing melody for some music.

1.3 Thesis Contributions

This thesis systematically studies the above three research questions regarding coevolution for human and machine learning and makes the following contributions.

1.3.1 Propose a BNP method for adding rule-based domain knowledge

The BNP method incorporates rule-based domain knowledge (utility function) provided by experts to assist a model. This method addresses the challenge of incorporating domain knowledge as prior information into the model. The method has three contributions:

1. We propose an automatic and generalizable framework based on the BNP model, which can simultaneously segment customers based on their behavior, discover their utility type, and determine their purchase behavior is influenced

by product price (while allowing for the direct involvement of additional external factors).

2. We unify the parameter estimation for different utility functions by using the method of derivation, allowing for the use of predefined conjugate priors to simplify the inference drastically. This overcomes the complexity of modeling and the inefficient inference associated with BNP.

3. We design an experimental solution based on the above methods and conducted experiments using both synthetic and real-world supermarket data, achieving good results.

1.3.2 Propose co-teaching accountable learning framework with distilled and domain knowledge

The Co-teaching accountable learning framework with distilled and domain knowledge is the solution for the coevolution of human and machine learning. The learning process of the model can help human gain trust and assist the model in achieving better performance with human guidance. This framework addresses the challenge of a interaction method between human and machine learning models. The co-teaching framework with dual-knowledge has three contributions:

1. We propose the framework to involve both supportive data and domain knowledge in the learning process. This framework can enable the model to produce high-performance results even with limited or low-quality data.

2. We provide a way of interaction between experts and machine learning models, enabling experts to pass critical knowledge for reasoning with ease and feasibility, without the need to the specify inference.

3. We design an experimental solution based on the above methods and conducted experiments using real-world medical data, achieving good results.

1.3.3 Propose an explainable framework for multimodal information

The explainable framework for multimodal information is a solution for making non-machine learning experts understand and trust the model. This framework addresses the challenge of complex model explanations for non-machine learning experts. The explainable framework for multimodal information has three contributions:

1. We propose a multimodal explanation framework by using privileged information. The privileged information not only can provide more accurate information(domain knowledge), but also can provide explanations for experts. This kind of explanation is more friendly to these non-machine learning experts.
2. We propose a fine-grained method for audio classification, which can help improve the learning efficiency of the model by identifying the main melody in music.
3. We propose an application framework based on privileged information in the field of music. By training the model with multimodal information (music, main melody, and text), we can use the model with only music information during testing, without requiring annotated data. This can reduce the cost of using the model.

1.4 Thesis Outline

This thesis systematically studies: ① Incorporate rule-based domain knowledge into machine learning models using the BNP model; ② Adopt a multi-model learning framework that combines domain knowledge with learning models to establish an interactive relationship between human and models; ③ Propose a multimodal framework to ensure that the explainability information of the model is fed back to human. This thesis is organized as follows in Fig1.4:

- Chapter 1: This chapter provides a brief introduction to the research background of the coevolution of human and machine learning, presents three

research questions, introduces the main contributions of this research, and briefly outlines the framework of this study.

- Chapter 2 : This chapter systematically reviews the research status of injecting domain knowledge, obtaining explanatory feedback and facilitating interaction between human and machine learning. It focuses on reviewing representative works, objectively describing and briefly evaluating their contributions and limitations, laying the groundwork for the research work of this paper.
- Chapter 3 : This chapter proposes an A BNP method for adding rule-based domain knowledge, which aims to improve the performance of model output results in situations with limited or unannotated data, in contrast to the reality of machine learning requiring large amounts of data or annotated data. The method is tested on both real supermarket datasets and simulated datasets, and its effectiveness is validated.
- Chapter 4 : This chapter addresses the issue of low-quality data in machine learning applications, which can cause output result biases. To solve this problem, a Co-teaching accountable learning framework with multiple model structures is proposed. In this framework, human intervention in machine learning systems enhances the performance of the model's output through the interaction between human and the model. The effectiveness of the framework is tested on a real medical dataset.
- Chapter 5 : This chapter proposes an explainable framework for multimodal information to address the issue of helping non-machine learning experts understand and trust the results of machine learning models. The framework is used to conduct research on emotion recognition for music, with a brief introduction to emotion recognition and musical domain knowledge. The effectiveness of the framework is tested on two real musical dataset, and a questionnaire survey is conducted to validate its effectiveness.

- Chapter 6: This chapter summarizes the work in this paper and analyzes the limitations of the research. Finally, some research directions with theoretical and practical value in this field are listed for future work.

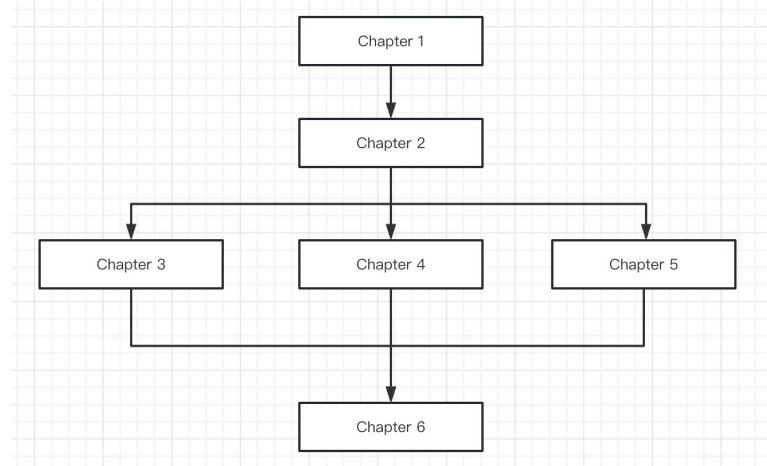


Figure 1.4: Thesis Outline

Chapter 2

Literature Review

Based on extensive reading and comprehension of the literature in the research field related to the coevolution of human and machine learning. Based on the research questions of the thesis, we categorize the literature into three main areas: ① How humans input domain knowledge into model? ② How machine learning models provide explanatory feedback to human? ③ How human and machine learning model collaborate with each other? This chapter conducts a comprehensive analysis and summary of the current research status, main academic viewpoints, research achievements, points of contention, existing problems, and possible causes in this thesis fields.

2.1 Injecting domain knowledge

2.1.1 Bayesian Nonparametric Model

A Bayesian nonparametric model [15] adapts the effective complexity of the model, measured by the number of dimensions used, to the data by choosing only a finite subset of available parameter dimensions that explain a finite sample of observations. The set of dimensions used depends on the sample.

Bayesian nonparametric models [16] are considered explainable models because they provide a probabilistic framework for modeling complex data, while allowing for flexibility and adaptability to the data without overfitting or underfitting. This adaptability arises because the effective complexity of the model is controlled by the data, as it selects the number of dimensions required to explain the data. Furthermore, Bayesian nonparametric models often have intuitive

priors that can be used to incorporate prior knowledge into the model, allowing for further explainability. The models can also be used to identify important variables and their relationships with the response variable, providing a deeper understanding of the underlying structure of the data.

①Non-Homogeneous Poisson Process

A non-homogeneous poisson process(NHPP) [17] is a stochastic process where the arrival rate of events varies with time, modeled as a function of time, and used to model situations where event rates change over time, such as in a telecommunications network or a queue. The process is a point process in which the probability of an event occurring depends on the time since the last event, and is an extension of the homogeneous Poisson process, which assumes a constant arrival rate over time.

The NHPP can naturally fit the problem of describing events (e.g. purchase) based on observations, with different types of utility functions. However, the inference of NHPP under the BNP framework is an extremely complicated task. Previous research focused on the inference of intensity function without considering grouping at the same time. The tractable inference was proposed in[18], then the major research is to find faster ways to infer the intensity function[19, 20]. There are also some models [21] considering the change of customer segmentation over time. The requirement of inference design is also the main obstacle to generalize these models. We propose a effective model to approximate NHPP with a unified framework to generalize.

Exploring purchase behavior based on price sensitivity is a common topic of customer analysis. Similar problem attracts research in a variety of applications, such as tourism and airline industry[22] . The utility function is the main assumption used in such studies to assess customer purchase behavior [23] to describe the relationship between price and purchase. The model determines one utility function for each customer, then analyses/segments customers according to the

assumed utility function.

② Hawkes Processes

Hawkes process [24] is a mathematical model used to model self-exciting processes. It is a counting process that describes a series of events occurring over time, where each event increases the likelihood of the next event (excitation), and the effect of this excitation decays over time. The counting process $N(t)$ can be viewed as an accumulated count of events that have occurred between the start ($T=0$) and the current time ($T=t$). We can use $N(T+h)-N(T)$ to describe the number of events that occur during this time interval. If there exists a random variable sequence $T = t_1, t_2, \dots$ with values in $[0, +\infty)$, then the random variable sequence T is referred to as a point process.

Hawkes processes are used to model events that exhibit clustering behavior, such as customer segmentation and behavior discover. They can also be used for prediction and forecasting, allowing for better understanding and management of systems that exhibit self-exciting or contagious behavior over time [25].

2.1.2 Transfer Learning

Transfer learning [26], in simple terms, refers to using existing knowledge to learn new knowledge. Its core principle is to identify similarities between existing and new knowledge. Due to the high cost of learning from scratch in the target domain, we turn to leveraging relevant existing knowledge to facilitate rapid learning of new knowledge. For example, we can transfer the knowledge of Australian English speech recognition to the speech recognition model for Canadian English. This way, we can achieve good performance without requiring a large amount of Canadian-specific data. Everything in the world has commonalities, and the key challenge of transfer learning lies in how to systematically identify these similarities and utilize this bridge to facilitate learning new knowledge.

①Knowledge Distillation

Knowledge distillation [27] is a technique in machine learning that is used to transfer knowledge from a complex model (the teacher model) to a smaller, simpler model (the student model). The goal of knowledge distillation is to improve the performance and efficiency of the student model, while maintaining its accuracy. The basic idea behind knowledge distillation is to train the student model to produce the same predictions as the teacher model. This is done by using the output of the teacher model, in addition to the actual labels, as the target for the student model during training. The teacher model acts as a teacher, providing the student model with information about how to make accurate predictions. There are several benefits to using knowledge distillation. Firstly, it can reduce the computational complexity of the model, making it faster and more efficient. Secondly, it can lead to improved accuracy, as the student model has access to the knowledge and experience of the teacher model. Finally, knowledge distillation can also be used to transfer knowledge between models that have different architectures, allowing for more flexible and scalable solutions.

②Privileged Information

Privileged information [28] is considered as expert knowledge and it has been combined with distillation knowledge [29]. Additional from X , the features in privilege information is $X' \in \mathbb{R}^{n \times (p+p')}$. Here p' columns are additional features that are only observed in training data. It can also be expert knowledge but only be available in training. Besides, our expert knowledge is added during the training. For the benefits, first, the expert can give critical knowledge without the redundant knowledge that can be learned from data. without specially designed model. Second, our framework does not require a specific design of models to fit the inconsistent feature dimensions in training and testing.

③ Few Shot Learning

Few-shot learning [30] is a type of machine learning that involves learning from a limited number of examples. Prototypical Networks [31] are a type of few-shot learning model that aims to solve this problem. These networks are based on the idea of prototypical examples, where each class is represented by a prototype that is learned from the few examples of that class. The prototypes are then used to make predictions for new examples, by calculating their similarity to the prototypes of each class.

④ Collaborated Learning

Collaborative learning has been widely discussed in human-computer interactions [32]. The aim is to help both *human learner* and *artificial learner* to synchronize knowledge. Machine learning also discusses collaborative learning which refers to the learning with distributed nodes, such as the federal learning [32] where each node will use its own data to tune the model. It does not contain experts' knowledge. These are all different from our setting. However, we can extend our framework with collaborative learning, by considering each expert as a node. They will all provide their own knowledge. However, the complex setting will face more challenges such as synchronizing between experts, as they may not be available at the same time.

⑤ Co-training

Co-training [33] aims to extend the knowledge from an initial small dataset by searching for useful data from a large but unlabeled dataset, by searching for similar data. However, it lacks certainty because of its unsupervised progress, especially when the endogenous space is sparse. When we proposed to use domain knowledge to extend, the student model will become more confident and certain and the learning will converge faster. A metaphor of co-training in real-life study is to let students discuss together to learn, while co-teaching is learning guided

by two teachers. we use teacher to point out what should be learned. We also examined the co-teaching mentioned in [34], which refers to two models learning together, so its essence is also co-training.

⑥Domain knowledge with active Learning

During the learning, some methods, such as active learning [35, 36] and interactive learning [37], can involve human knowledge. Active learning prompts users with uncertain data to label them as queries, but there are natural drawbacks. First, the queried data rely on the model, this could be useless if the model was trapped into local minimum/over-fitting with biased data/low data volume. On the contrary, experts in our framework can confidently provide critical knowledge guiding learning. Second, active learning assumes an omniscient oracle however the oracle is limited by the data pool to be queried. That means active learning does not generate data but can only push the possibly existing data. However, experts can focus on data that could be in an unknown domain. The third issue of active learning is query number and convergence. The expert can provide critical information rather than answering all active learning queries.

⑦Autonomous Machine Learning

Autonomous learning [38], also known as self-directed learning, is a learning process where individuals take charge of their own learning by setting their own learning goals, selecting resources and materials, and evaluating their own progress. This type of learning is characterized by being self-motivated, self-directed, and self-reliant. However, autonomous learning, e.g. Auto Machine Learning (AutoML) that can automatically find optimal hyper-parameters, provides huge benefits to modelers instead of domain users. So they can skip the domain understanding. In many cross-domain collaborations, however, such learning outcomes are often questioned by domain experts. Our co-teaching framework is not the opposite of autonomous machine learning. The student model can still be fit

into AutoML, however, we let domain experts see how the model is formed and provide helps to learn. This help can let the model be more accurate, trusted, accountable, and equality [39].

2.2 Obtaining explanatory feedback

2.2.1 Explainable Machine Learning

①Global Model-Agnostic Methods

Global model-agnostic methods [40] are techniques in explainable artificial intelligence (XAI) that aim to provide a global understanding of a machine learning model’s predictions. Unlike local explanations, which focus on understanding the prediction for a specific instance, global explanations describe the average behavior of a machine learning model. They are designed to provide a high-level view of the model’s behavior and decision-making process, and to uncover any biases or limitations in the model. Below are several significant techniques:

Partial dependence plot [41] demonstrates which variables have the greatest impact on predictions, while partial dependence plots show how features influence model predictions. Partial dependence plots can be used to answer questions similar to the following: What is the impact of latitude and longitude on house prices, while holding all other features constant? In other words, how does the price of houses of the same size vary in different locations? Is the difference in predicted health levels between two different groups primarily influenced by their debt levels or is there another reason? If you are familiar with linear regression or logistic regression, the effect of partial dependence plots is similar to the parameters in these models. For example, partial dependence plots always show a linear relationship when applied to a linear regression model. However, compared to the parameters in simple models, partial dependence plots on complex models can capture more complex patterns. Similar to permutation importance, partial dependence plots can only be computed after fitting the model. The model is

trained on unmodified real-world data.

Accumulated Local Effects [42] plot illustrates the average influence of features on predictions made by a machine learning model. ALE plots provide a faster and impartial substitute for partial dependence plots.

A global surrogate model [43] is an explainable model designed to approximate the predictions of a black-box model. By explaining surrogate models, we can derive insights and draw conclusions about the behavior of black-box models. In fact, understanding the surrogate model does not require much theory, and provides a simple and explainable explanation for the behavior of complex black-box models. It can be a linear model or a nonlinear model, depending on the complexity of the black box model. Global agents can be trained using any explainable machine learning algorithm, such as decision trees, logistic regression, or random forests, among others. We want to approximate our black-box predictor function f with the surrogate model predictor function g as closely as possible, subject to the constraint that g is explainable. For the function g , any explainable model can be used.

②Local Model-Agnostic Methods

There are several benefits to separating the explanations from the machine learning model, also known as model-agnostic explanation methods [44]. One of the key benefits of these methods compared to model-specific approaches is their flexibility. These explanation methods can be applied to any machine learning model, giving developers the freedom to choose the model that best suits their needs. Additionally, any graphics or user interfaces that rely on the explanation of the model become independent of the specific machine learning model used. When evaluating multiple models for a task, it is easier to compare explainability using model-agnostic explanations as the same method can be applied to any type of model. Here are a few important methods:

Local Surrogate: The Local Surrogate is a kind of explainable model, which

can be used to analyze the black box machine learning model. Local explainable model-agnostic explanations (LIME) [44] is a specific implementation of a local Surrogate model, where the LIME is trained to approximate the predictions of the underlying black-box model. LIME focuses on the concept of training a local surrogate model to provide explanations for individual predictions. To understand this approach, let's consider a scenario where we only have access to a black box model. In this case, we can input various numbers and observe the corresponding model outputs. The box can be probed on demand. The purpose of the study is to understand the specific reasons behind the machine learning model's predictions. The LIME examination occurs when you input the general numbers into the machine learning model. LIME generates a new collection of numbers, which is based on the perturbed number collection. Then, in this new numerical collection, LIME trains a model that can be understood, and the model is weighted based on the closeness of the sampled instances to the instances of interest. The explainable model can be any kind of explainable model, for example, a decision tree or a Bayesian model. This local surrogate model is a good approximation of the black box model's predictions locally, but it may not be accurate globally.

Individual Conditional Expectation: Individual Conditional Expectation (ICE) plot illustrates the change in an instance's prediction as a feature changes, by displaying one line per instance. Individual expectations instead of partial dependencies is to gain a more detailed understanding of how a feature affects the prediction for each instance separately. While partial dependence plots can show the average relationship between a feature and the prediction, they can obscure important variations in the relationship between instances. For example, there may be interaction effects between features that create a heterogeneous relationship between a feature and the prediction. In such cases, the ICE plot can provide much more insight because it visualizes the individual relationship between the feature and the prediction for each instance separately.

By examining the ICE plot, one can identify subpopulations of instances with

different response patterns to a feature, which may be due to interactions with other features. This information can be useful in developing more accurate and explainable models and can help to identify areas where the model may be improved.

Overall, individual conditional expectation (ICE) plots are useful because they provide a more detailed understanding of the relationship between a feature and the prediction for each instance separately, which can be important when there are interactions between features.

③ Example-Based Explanations

Example-based explanation methods select instances to explain machine learning models or data distribution [45]. These methods are mostly model-agnostic, making any model more explainable by highlighting specific instances instead of summarizing features. Example-based explanations are effective if data instances are easily understandable, such as in images or structured data. Representing tabular data can be challenging, but listing feature values or summarizing the instance can help make it more effective [46]. The following explanation methods are all example-based:

Counterfactual Explanations: Counterfactual Explanations [47] Counterfactual explanations are a type of explanation used in machine learning to show the impact of small changes to the input data on the prediction made by a model. The goal of counterfactual explanations is to provide insight into the decision-making process of a model by showing how the prediction would change if certain aspects of the input data were altered. Counterfactual explanations are created by computing the prediction of the model for a slightly modified version of the input data, and comparing the results to the original prediction. For example, if a model predicts that a loan applicant will default on a loan, a counterfactual explanation could show what would happen if the applicant had a higher income, or if they had a longer work history. Counterfactual explanations are particularly

useful for identifying the factors that are most important for a prediction made by a model, and can help to uncover biases or limitations in the model. They can also be used to understand how a model is making predictions, and to provide more transparent and trustworthy explanations for its decisions.

Influential Instances: To understand influential instances [48, 49] using explainable machine learning, we can analyze the prediction of the model for each instance in the dataset, and compare the prediction to the prediction made for slightly altered versions of the instance. This allows us to identify which instances have the largest impact on the prediction, and are therefore considered influential. Imagine that you want to estimate the average income of people in your city and randomly ask ten people on the street about their income. To answer this question, you can recalculate the average by omitting individual answers or derive mathematically how the average is affected using an “influence function”. Using the deletion method, we recalculate the average ten times, omitting one each time to calculate a profit and loss statement and measure the degree of change in the average estimate. A large change means that an instance is very influential. Influential instances use an infinitesimal weight to increase a person’s weight, which corresponds to calculating a statistic or model’s first-order derivative. By the way, the answer is that your average estimate may be strongly influenced by a single answer because the average is linearly related to individual answers. A more robust choice is the median (the value where half of people earn more and half earn less) because even if the income of the highest-earning person in the sample increases tenfold, the resulting median will not change. This means influential instances can be applied to the parameters or predictions of machine learning models to understand their behavior better or to explain individual predictions.

④Saliency Maps

Saliency maps [50, 51] are used to visualize which parts of the input data are most important for a particular prediction made by a neural network. The goal

of saliency maps is to provide insight into the decision-making process of a neural network, by highlighting which parts of the input data the network is focusing on when making a prediction. Saliency maps are created by computing the gradient of the prediction with respect to the input data, which shows how much the prediction changes when small changes are made to each part of the input data. The gradient is then used to highlight the regions of the input data that have the largest impact on the prediction. Saliency maps can be useful for identifying which features of the input data are important for a particular prediction, and can help to uncover biases or limitations in a neural network. They can also be used to detect errors or mistakes in a neural network, and to understand how a neural network is learning to make predictions.

Towards Model-Level Explanations of Graph Neural Networks: One important example for feature attribution is Towards Model-Level Explanations of Graph Neural Networks(XGNN). The XGNN is a tool that explains Graph Neural Networks (GNNs) at the model level [52]. It provides a high-level understanding and insight into how GNNs work. The approach used by XGNN involves training a graph generator that maximizes a specific prediction made by the model. The graph generation is framed as a reinforcement learning task, where each step involves the prediction of adding an edge to the current graph. The graph generator is trained using a policy gradient method, which utilizes information from the pre-trained GNN. Additionally, XGNN incorporates various graph rules to ensure that the generated graphs are valid. Results from experiments conducted on synthetic and real-world datasets demonstrate that XGNN effectively helps in understanding and verifying the performance of trained GNNs.

2.2.2 Truthworthy machine learning

Trustworthy Machine Learning [53, 54] brings four concepts to machine learning: explainability [55], fairness, privacy, and robustness. Trustworthy Machine Learning first discusses how to explain ML model outputs and internal workings.

Then, Trustworthy Machine Learning studies how bias and unfairness arise in ML models and learns strategies to mitigate this problem. Next, Trustworthy Machine Learning will study differential privacy and membership inference in the context of models leaking sensitive information when they should not [56]. Finally, Trustworthy Machine Learning will discuss adversarial attacks and methods for providing robustness against adversarial manipulation.

Federated Learning, as discussed in Zhang’s survey [57], is an emerging artificial intelligence technology. It is designed to ensure information security, protect terminal data and personal data privacy, and ensure compliance while conducting efficient machine learning among multiple participants or computing nodes. This technology focuses on preserving privacy in AI by allowing collaborative machine learning without sharing raw data. The machine learning algorithms that can be used in federated learning are not limited to neural networks but also include important algorithms such as random forests. Federated learning is expected to become the foundation of the next generation of artificial intelligence collaborative algorithms and collaborative networks. Federated learning is a powerful tool for privacy-preserving AI because it allows for the training of models on distributed data sources while maintaining the privacy of those sources [58]. With federated learning, the data remains on the devices or servers where it was collected, and only the updated model parameters are transmitted to a central server. This enables companies and organizations to train robust AI models without compromising the privacy of their users or customers.

2.3 Facilitating interaction between humans and machine learning

2.3.1 Human in the loop for machine learning

Humans in the loop integrate human knowledge and experience to train accurate prediction models at the lowest cost. With the help of machine-based methods, humans can provide training data for machine learning applications and

directly complete some tasks that are difficult for computers to complete in the pipeline [59, 60, 61]. Human in the loop is often used to improve model performance (classification, dialogue, and QA, etc.) and generalization ability, as well as to improve model explainability and usability (user-created feature dictionaries, user-generated adversarial Q in QA, etc.) and enhance user experience.

2.3.2 Active Learning

By using some technical means or mathematical methods to reduce the cost of people’s labeling, scholars call this direction active learning . In the entire machine learning modeling process, there are parts and links where humans are involved, and the process of using machine learning methods to screen out suitable candidates for human labeling. Active learning [35] prompts users with uncertain data to label them as queries, but the queried data rely on the model which is problematic when the data cannot be guaranteed. We will show that the experts will provide critical knowledge that will boost the convergence speed. With biased data/low data volume, the model was trapped into local minimum/over-fitting.

2.3.3 Reinforcement Learning

Reinforcement learning [62] is a field in machine learning that emphasizes how to act based on the environment to achieve maximum expected benefits. Its inspiration comes from the behavioral theory in psychology, that is, how organisms gradually form expectations of stimuli under the stimuli of rewards or punishments given by the environment, and produce habitual behaviors that can obtain the greatest benefits. Reinforcement learning is the third basic machine learning method besides supervised learning and unsupervised learning. Unlike supervised learning, reinforcement learning does not require labeled input-output pairs and does not require precise correction of non-optimal solutions. Its focus is on finding a balance between exploration (of unknown domains) and exploitation (of existing knowledge) . The “exploration-exploitation” trade-off in reinforcement learning

has been most studied in the multi-armed bandit problem and finite MDPs. Some technologies and techniques commonly used in reinforcement learning such as q-learning, deep reinforcement learning, policy gradients, actor-critic methods and multi-agent reinforcement learning.

2.4 Literature Conclusion

This chapter reviews the current research on existing methods in three aspects: Injecting domain knowledge into the model, obtaining explanatory feedback from the model, and facilitating interaction between human and machine learning. Under the guidance of my supervisors, I have gradually developed my own research ideas through studying, referencing, discussing, and experimenting with these three aspects of literature.

Chapter 3

A BNP method for adding rule-based domain knowledge

In Chapter 3, we provide a solution that enables experts to provide specific rule-based domain knowledge to models, thereby improving their performance. This method effectively addresses the issue of unsatisfactory model output results in situations with limited or low-quality data. In this work, we conducted relevant research on supermarket customer segmentation and proposed a method that simultaneously performs customer segmentation and behavior discovery by incorporating rule-based domain knowledge to support the model. We validated our solution using public supermarket datasets and synthetic datasets, demonstrating its effectiveness.

3.1 Motivations

Customer segmentation is a common technique that enables businesses to optimize their resources and increase profits by analyzing purchase behavior [22, 63, 64]. The objective of segmentation is to categorize customers into manageable sub-groups, allowing tailored marketing activities for different customer types. Traditionally, customer segmentation is based on demographic data, with different utility functions applied to customer groups. However, this approach has two drawbacks: 1) Collecting demographic data poses challenges in balancing data completeness and customer privacy. Missing key attributes can result in unreliable segmentation. To address this, we utilize purchase data instead, as key demographic attributes can implicitly influence purchase behavior, which is reflected in the purchase patterns. 2) In the traditional setting, utility functions are determined before segmentation. In our proposed framework, we simulta-

neously estimate customer segmentation, as well as the form and parameters of utility functions. Additionally, we employ the Bayesian non-parametric (BNP) framework for segmentation, eliminating the need to explicitly set the number of clusters. Overall, our approach offers a more flexible and data-driven solution to customer segmentation.

There are several challenges involved in jointly estimating utility functions, parameters, and segmentation within the Bayesian non-parametric (BNP) framework. Firstly, optimizing utility functions is a functional problem that requires modeling stochastic processes. Previous work by Luo et al. utilized non-homogeneous point processes (NHPP) to describe customer purchase behavior [19], employing a base measure of polynomial and trigonometric functions. While such models are mathematically complex, necessitating careful design for compatibility, robustness, and scalability, we propose an approximation model that maintains flexibility in modeling utility functions with different parameters for various types.

Secondly, parameters for different types of utility functions exhibit inconsistency in their prior and likelihood, given the diverse meanings associated with each. This inconsistency in posteriors often leads to a significant amount of heuristic work, as the modeling heavily relies on expert knowledge. To address this, we introduce a unified function setting that ensures consistency in the parameter space. We demonstrate its effectiveness by employing three typical types of utility functions. The re-parametric solution involves encapsulating the identity of the utility function within a parameter-free nonlinear function to maintain parameter consistency. This design choice significantly reduces the burden of designing utility functions by assuming that all parameters are generated from the same posterior distribution. As a result, the model exhibits higher generalizability and is easier to interpret.

Overall, our approach tackles these challenges and enhances the inference process in estimating utility functions, parameters, and segmentation within the BNP framework.

There are three main contributions of this work. Firstly, we propose an automatic and generalizable framework based on the Bayesian non-parametric (BNP) model. This framework allows for simultaneous segmentation of customers considering their behavior, discovery of their utility type, and analysis of how their purchase behavior is influenced by factors such as product price (additional external factors can also be incorporated). Previous studies have not explored the application of BNP modeling in purchase analysis. Secondly, we unify the parameter estimation for different utility functions by employing a derivation-based method. This approach enables the use of predefined conjugate priors, greatly simplifying the inference process. In our model, only two priors are required regardless of the number of utility functions considered. Lastly, we design an experimental solution based on the aforementioned methods and conduct experiments using both synthetic and real-world supermarket data. The results of these experiments demonstrate the effectiveness of our approach. Our work makes significant contributions by introducing an innovative framework, unifying parameter estimation, and providing experimental validation using diverse datasets.

3.2 Preliminary Knowledge

3.2.1 Dirichlet Process

The Dirichlet distribution (DP) is the conjugate distribution of the multinomial distribution, and the Beta distribution is a special case of the Dirichlet distribution, specifically in its binary form. According to the definition from Wikipedia, the Dirichlet distribution is a continuous multivariate probability distribution with a positive parameter vector α . It is commonly used as a prior distribution in Bayesian statistics. The Dirichlet process is the infinite-dimensional extension of the Dirichlet distribution.

The core idea behind the DP is as follows: Suppose we have a set of N data points, denoted as $c_N = (c_1, \dots, c_N)$, where each point is independently gener-

ated from an unspecified, generic distribution G . The Dirichlet process allows us to establish a prior belief about generic distributions, denoted as G . Our prior belief suggests that the most probable scenario is G_0 , but we have limited confidence in this prediction. We can represent our confidence level using the parameter α : a small α indicates low confidence, resulting in a prior that encompasses numerous potential distributions G , while a large α indicates a highly concentrated set of possible distributions around G_0 . Therefore, we can express this relationship as follows:

$$\begin{aligned} c_i &| G \sim G \\ G &| G_0, \alpha \sim \text{DP}(G_0, \alpha) \end{aligned} \tag{3.1}$$

The crucial aspect here is that the distribution G is generated in a manner that aligns with the Chinese Restaurant Process (CRP), the stick-breaking process, and the Pólya urn scheme. For example, one approach to sampling c_N would involve generating the values using the Pólya urn. Another equivalent method would be to first sample z_N from the CRP. Then, all customers seated at the k th table are assigned the same color, independently sampled from G_0 . In a less efficient manner, one could construct the complete distribution G by sampling using the stick-breaking process and associating each k value with a color (sampled from G_0). The probability that the i th customer is assigned the k th color is denoted by θ_k .

The fundamental point here is that the Dirichlet process characterizes a "distribution over distributions" and does so consistently with all the aforementioned simpler concepts. However, it is important to note that the Dirichlet process can only generate discrete distributions. It cannot serve as a prior for continuous distributions, and attempting to use it in such a manner can yield peculiar outcomes.

3.2.2 Chinese restaurant process

The Chinese restaurant process (CRP) is a probabilistic model that provides a distribution over partitions. It is used to cluster or partition a collection of observations into groups. In this metaphorical representation of a Chinese restaurant, each potential group is likely to a table within an infinitely large Chinese restaurant. Every observation is compared to a "customer" entering the restaurant and choosing a table to sit at. In this analogy, customers are assumed to have a preference for popular tables, but there is still a non-zero probability that a new customer will select an unoccupied table.

To illustrate the mechanism, let's consider a scenario where there are currently N customers seated in the restaurant. We can use the indicator variable z_i to determine the table at which the i -th customer is seated. We can represent the table assignments of the customers as a vector, $\mathbf{z} = (z_1, z_2, \dots, z_N)$, where each z_i indicates the table number assigned to the i -th customer. Note that $\sum_k 1^k n_k = N$. If there are currently N customers seated in the restaurant, the probability that customer $N + 1$ sits at the k -th table is proportional to the popularity of that table:

$$P(z_{N+1} = k \mid \mathbf{n}, \alpha) = \frac{n_k}{N + \alpha} \quad (3.2)$$

The hyperparameter α in the CRP can impact the likelihood of a new customer choosing a new table. After summing Equation 3.1 over all K tables, the resulting sum is $N/(N + \alpha)$. This is because there is a probability that customer $N + 1$ chooses to sit at a new table, which we label as table $K + 1$. The probability of customer $N + 1$ sitting at a new table is:

$$P(z_{N+1} = K + 1 \mid \mathbf{n}, \alpha) = \frac{\alpha}{N + \alpha}. \quad (3.3)$$

Here, it is generated a set of assignments from the CRP through sampling:

$$\mathbf{z} \mid N, \alpha \sim \text{CRP}(\alpha) \quad (3.4)$$

For example, in the figure 3.1, circles represent dining tables, and numbers represent customers. Assuming there are an infinite number of tables, each customer entering the restaurant chooses a table to sit at, representing an experiment with an infinite number of possible outcomes. Initially, all tables are empty, and when customer 1 enters, they directly sit at table 1.

When customer 2 enters, they have a probability of $\frac{1}{1+\alpha}$ to sit at table 1 and a probability of $\frac{\alpha}{1+\alpha}$ to sit at a new empty table. The result is that they sit at table 1.

When customer 3 enters, they have a probability of $\frac{2}{2+\alpha}$ to sit at table 1 and a probability of $\frac{\alpha}{2+\alpha}$ to sit at a new empty table. The result is that they sit at a new empty table, table 2.

...

When customer 8 enters, they have a probability :

$$\frac{3}{7+\alpha}, \frac{1}{7+\alpha}, \frac{2}{7+\alpha}, \frac{1}{7+\alpha}, \frac{\alpha}{7+\alpha}$$

They are the probability of sitting at table 1, table 2, table 3, table 4, and a new empty table, and the result is sitting at 3 tables;

Therefore, the probability of occurrence of the figure 3.1 is

$$\begin{aligned} p(z = (1, 1, 2, 3, 1, 3, 4, 3)) \\ &= p(z_1) p(z_2 \mid z_{z_1}) p(z_3 \mid z_{1:2}) \dots p(z_8 \mid z_{1:7}) \\ &= 1 \cdot \frac{1}{1+\alpha} \cdot \frac{\alpha}{2+\alpha} \dots \frac{2}{7+\alpha} \end{aligned}$$

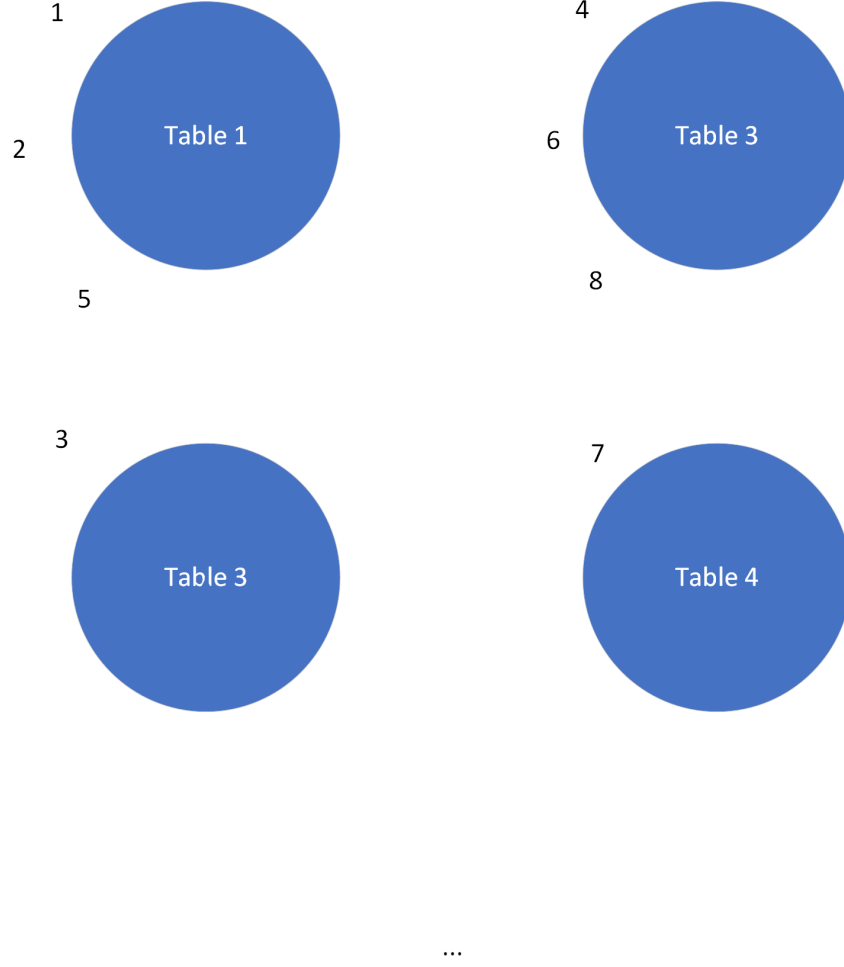


Figure 3.1: Example of CRP

3.3 Methodology

3.3.1 Problem Definition

Given a customer $i = \{1 \dots N\}$, we know a series of their purchase events \vec{y}_i , where $\vec{y}_i = \{y_{i,j} | y_{i,j} \in \mathbb{N}^+\}$ is the number of products that customer i purchased for their j^{th} purchase. The corresponding price is $\vec{x}_i = \{x_{i,j} | x_{i,j} \in \mathbb{R}^+\}$. We assume that there are M_i observations for i in total. Here we use discount rate as the price to normalize the price value of different types of products. Details will be shown in the data pre-processing in Section 3.5.3. Our target is to segment customers into unknown number of groups, so that a group index k_i needs to be obtained for i . The grouping is based on the function f_i that can map price $x_{i,j}$ with purchase behavior $y_{i,j}$.

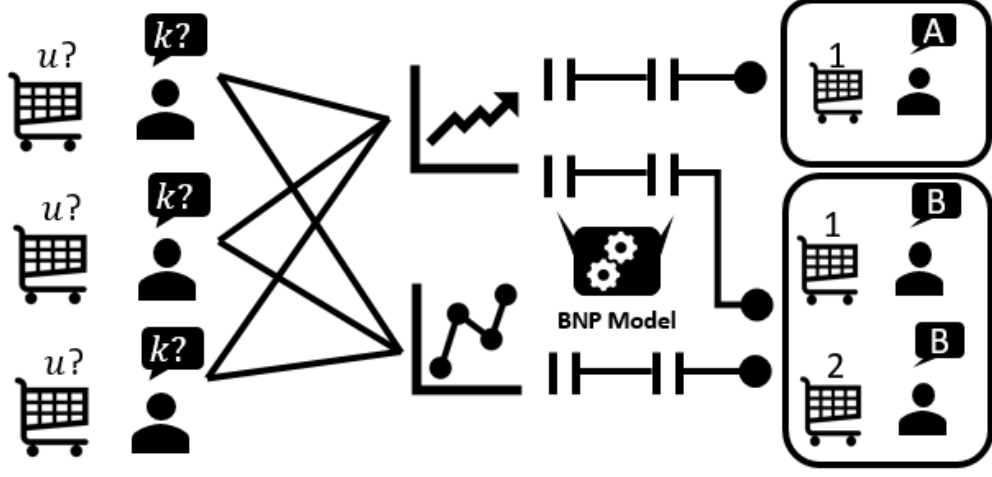
Determining and grouping functions is mathematically a functional problem. Traditionally, this can be modeled using the BNP framework with stochastic processes as the mixture components. The generalization of such methods is hampered by its complexity since dedicated inference must be designed. The MCMC method is the most often used inference algorithm, which could become inefficient for stochastic process due to the high dimension.

Therefore, considering the efficiency and generalizability, we convert the problem into the semi-parametric model but it can approximately cover the space that a full BNP model can cover. Here we assume that three utility functions can represent most types of relationships between customers' purchase behavior and price. As we do not know which function should be used for each customer, we use a latent variable $u_i \in \{1, 2, 3\}$ to denote the function selected for customer i . We hope to jointly estimate group index k_i and utility function type u_i , without heuristically selecting for each customer, by the integration of a Bayesian semi-parametric model for utility function selection and BNP model for grouping.

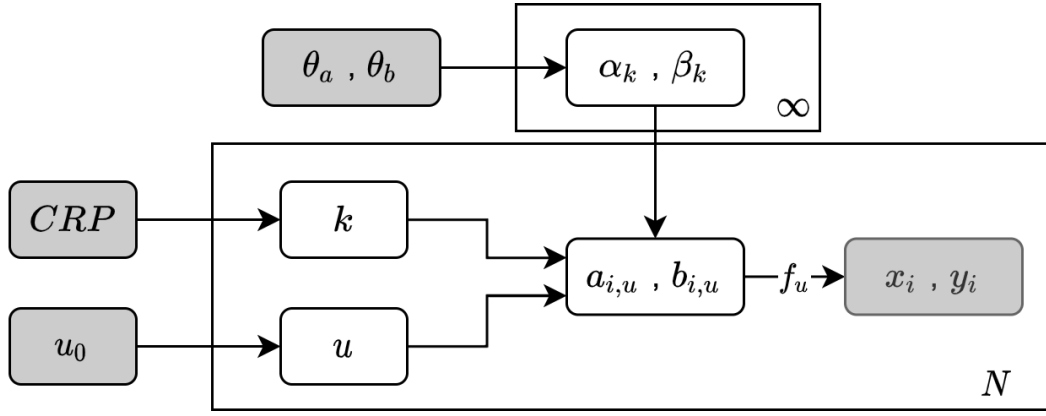
The flow of our work is shown in fig: sub-shiyitu. As shown in the figure, all the customers will be compared with different utility functions, then our algorithm will determine the best latent utility function used to describe the customer behavior and customer group. In fig: sub-graphic-model, two people are associated with utility function 1, and one person is associated with utility function 2. As to grouping results, one person is in Group A, while two people are in Group B. The customers in the same group have the same parameters as the utility functions. Since we have the unified utility function, the parameters for different forms of utility functions can still have the same parameter values.

3.3.2 Preliminary for CRP Model

The CRP (Chinese Restaurant Process) model is a manifestation of the Dirichlet process. It is a discrete random process to mimic the allocation of customers to different tables in a Chinese restaurant, with an unlimited number of tables



(a) Select utility functions and groups



(b) Graphic model of UtSeg, shaded nodes are known (hyperparameters and observations)

Figure 3.2: The flow chart and graphic model of our framework.

in assumption. Customers can be allocated to either an existing table or a new table. Due to the exchangeability property of CRP, the allocating result for all customers is equivalent to a process to allocate them one by one[65]. The probability of customer i sitting at a table m is defined as follows:

$$p(k_i = m) = \begin{cases} \alpha_0 / (i - 1 + \alpha_0), & \text{if table } m \text{ has no customers} \\ c_m / (i - 1 + \alpha_0), & \text{if table } m \text{ has } c \text{ customers} \end{cases} \quad (3.5)$$

where α_0 is a prior parameter of Dirichlet, c_m is the number of customers already on the table m . When the customer chooses the table, it is not only related to

the number of existing customers on the table but also related to the customer's behavior at the table. The customers who sit in the same table have similar behavior, and the new customer may choose to sit at a table with similar behavior. The basic CRP does not consider the information which is brought by the customers. That is why we need to consider the likelihood.

3.3.3 Utility Functions

The utility functions are used to describe the relationship between the price and demand [66, 67]. In economics, there are many utility functions. We define three different utility functions to describe typical behavior with price changes. These three functions can explain most of the relation between purchase behavior and price.

The traditional way to represent different functions is based on different parameters. For example, for customer i the functions could include:

$$f_i^1 : y_i = a_1 x_i + b_1, f_i^2 : y_i = \log_{a_2} x_i + b_2, f_i^3 : y_i = a_3^{x_i} + b_3 \quad (3.6)$$

However, based on such a trivial setting, 6 parameters must be considered for posterior. This will cause many heuristic settings, which requires domain knowledge to determine what priors are needed for each different setting, and what likelihood function is suitable for each parameter, otherwise, it could cause intractability with non-conjugated priors. This setting limits the generalizability and scalability of the system.

To overcome such an obstacle, we propose the utility function setting with unified parameters, which means that the parameters from different utility functions can be generated from the same distribution. Meanwhile, they can cover all the possible utility functions that we are going to use to model customer behavior. All three utility functions can be described in the unified form as follows:

$$\vec{y} = \vec{a}g(\vec{x}) + \vec{b}, \quad (3.7)$$

Eq. (3.7) gives a general representation of any relation between demand and price change. The parameters a and b are unified, which can be interpreted in the same way in different utility functions, such as a as the coefficient and b as adjustment. Then we can represent them with fixed prior functions with any such setting. If the parameters of any utility functions can be reduced to Eq.(3.7), they can be directly used in our model. Such a setting provides the generalization capability so that it can be easily extended to incorporate a variety of utility functions. Specifically, the utility functions explored in this work are:

$$\begin{aligned} f_i^1 : \quad y_i &= a_{i,1}x_i + b_{i,1} \\ f_i^2 : \quad y_i &= a_{i,2} \log(x_i) + b_{i,2} \\ f_i^3 : \quad y_i &= a_{i,3}e^{x_i} + b_{i,3} \end{aligned} \tag{3.8}$$

The first utility function is a linear equation in its standard form, which means that the quantity of demand is a linear function of price. The second utility function is a logarithmic equation. The Fechner's law [68] is a principle law widely used in psychology. The purchase behavior can be considered as a psychological behavior, in which the subjective price scale for the buyer follows a logarithmic scale and there is a range of acceptable prices for certain products. The third utility function is an exponential equation. Weber's law [68] states that a just-noticeable change in a given stimulus appears as a constant ratio of the original stimulus. This can be applied to pricing by identifying the point at which a price change is sufficient to be 'noticed' by the customers to change their response.

The utility function is selected based on the negative log-likelihood loss function. It is a common way to measure if the utility function can fit the data points well or not [18].

3.3.4 Simultaneous Customer Segmentation and Utility Estimation Model

This section describes the details of the Simultaneous Customer Segmentation and Utility Estimation (UtSeg) model and introduces the generative process of parameters, latent variables, and observations of UtSeg. The model is mainly used for cluster customers into groups and infer their utility functions. The graphic model of UtSeg is shown in fig: sub-graphic-model.

In the UtSeg, α_0 is the hyperparameter for the CRP. Then, each customer i will get $\vec{a}_i = [a_{i1}, a_{i2}, a_{i3}]$ and $\vec{b}_i = [b_{i1}, b_{i2}, b_{i3}]$ by using curve fitting function based on price and purchase information. We assume that \vec{a}_i and \vec{b}_i follow Gamma distribution. This is because of the Poisson distribution used to estimate the purchase number, given by $N_i(x_i) \sim Poi(y_i(x_i))$. The Poisson distribution can be decomposed as a superposition of multiple Poisson distributions with the summation of frequencies as the overall frequency. Therefore, b_i is also the parameter of a Poisson distribution.

This is the first property that can drastically reduce the complexity of inference, because the conjugated prior can be set for both gamma distribution, without sampling in high dimension space for stochastic process.

For each customer i , the generative process of UtSeg can be represented as follows:

$$\begin{aligned}
 k_i &\sim CRP(\alpha_0), u_i \sim Mul(u_0), \alpha_{k_i} \sim Gam(\theta_a), a_{i,u_i} \sim Gam(\alpha_{k_i}), \\
 \beta_{k_i} &\sim Gam(\theta_b), b_{i,u_i} \sim Gam(\beta_{k_i}), l_{i,u_i} \sim N(\mu_{i,u_i}, \sqrt{M_i}\sigma_0) \\
 \mu_{i,u_i} &= \sum_{j=1 \dots M_i} a_{i,u_i} g_{u_i}(x_i) + b_{i,u_i} - y_i
 \end{aligned} \tag{3.9}$$

- We generate table index k_i based on CRP , using the hyperparameter α_0 ;
- For utility function selection, an function index u_i is generated for customer i with Multinomial distribution parameterized by u_0 ;

- We generate a latent variable for each table k_i , for both coefficient variable α_{k_i} and offset variable β_{k_i} with the base measure parameterized by $\text{gamma}(\theta_a)$ and $\text{gamma}(\theta_b)$ ¹;
- a_i and b_i are generated based on α_{k_i} and β_{k_i} using Gamma distributions;
- The selected function should fit the observations, so the minimised loss l_{iu_i} can be learned based on Section 3.3.3. l_{iu_i} is assumed to be Gaussian distributed² loss, with variance σ_0 , mean $\mu_{i,u_i} = \sum_j a_{i,u_i} g_{u_i}(x_i) + b_{i,u_i} - y_i$, based on a_i and b_i as a_{i,u_i} and b_{i,u_i} respectively.

Therefore, the joint probability of the model is:

$$\begin{aligned}
& P(k_{1\dots n}, u_{1\dots n}, \alpha_0, \theta_a, \theta_b, l_{1\dots n, 1\dots 3}, a_{1\dots n, 1\dots 3}, b_{1\dots n, 1\dots 3}) \\
& \propto \prod_i P(k_i | \alpha_0) P(\alpha_{k_i} | \theta_a) P(\beta_{k_i} | \theta_b) P(u_i | u_0) P(a_{i,u_i} | \alpha_{k_i}, k_i) \cdot \\
& \quad P(b_{i,u_i} | \beta_{k_i}, k_i) P(l_{i,u_i} | u_i, a_{i,u_i}, b_{i,u_i}, g_{u_i}, x_i, y_i, \sigma_0, M_i) \\
& = \prod_i CRP(k_i | \alpha_0) \text{Gam}(\alpha_{k_i} | \theta_a) \text{Gam}(\beta_{k_i} | \theta_b) \text{Mul}(u_i | u_0) \text{Gam}(a_{i,u_i} | \alpha_{k_i}) \\
& \quad \text{Gam}(b_{i,u_i} | \beta_{k_i}) N(l_{i,u_i} | \mu_{i,u_i}, \sqrt{M_i} \sigma_0)
\end{aligned} \tag{3.10}$$

3.4 Inference: Gibbs Sampling for UtSeg model

Gibbs Sampling is a Markov Monte Carlo method (MCMC) [65, 69], which is widely used in the inference. In the UtSeg model, each customer is assigned to a utility function based on the multinomial prior and Gaussian likelihood for the loss function. The parameters θ_a and θ_b are randomly initialized and u_i from the last step is used. The possible sampling result can be any existing table or starting a new table. For each customer i , the posterior probability to select a

¹For a Gamma distribution, we simplify both actual parameters into one parameter.

²This can be determined by the loss used. We use the quadratic loss, but Gaussian distribution is used to approximate the Chi-square distribution when data volume is large.

table k_i is:

$$\begin{aligned}
& p(k_i = k | k_{i-}, u_{1\dots n}, \alpha_0, \theta_a, \theta_b, l_{1\dots n, 1\dots 3}, a_{1\dots n, 1\dots 3}, b_{1\dots n, 1\dots 3}) \\
& \propto CRP(k | k_{i-}, \alpha_0) Gam(\alpha_k | \theta_a) Gam(\beta_k | \theta_b) Gam(a_{i, u_i} | \alpha_k) \\
& Gam(b_{i, u_i} | \beta_k) N(l_{i, u_i} | \mu_{i, u_i}, \sqrt{M_i} \sigma_0).
\end{aligned} \tag{3.11}$$

where k_{i-} represents the current table assignments except for customer i . Similarly we sample the form of utility u_i by:

$$\begin{aligned}
& p(u_i = u | k_{1\dots n}, u_{i-}, \alpha_0, \theta_a, \theta_b, l_{1\dots n, 1\dots 3}, a_{1\dots n, 1\dots 3}, b_{1\dots n, 1\dots 3}) \\
& \propto Mul(u | u_0) Gam(\alpha_k | \theta_a) Gam(\beta_k | \theta_b) Gam(a_{i, u_i} | \alpha_k) \\
& Gam(b_{i, u_i} | \beta_k) N(l_{i, u_i} | \mu_{i, u_i}, \sqrt{M_i} \sigma_0).
\end{aligned} \tag{3.12}$$

By sampling all the k_i and u_i iteratively, we can get the utility function allocation for all customers.

3.5 Experiments

Our experiment compares the performance of UtSeg and other models that could be used for customer segmentation, using a synthetic data set and a real-world supermarket data set.

3.5.1 Experiment Setup

Baseline models:

- **UtSeg-(1-3)**: This is a simplified model from UtSeg, which is based on CRP and one utility function to describe customer purchase behavior. Each method corresponds to one of the utility functions.[68]
- **CRP-GM**: This baseline is CRP with Gaussian mixture component [65]. We use $x_{i,j}$ and $y_{i,j}$ to compute the likelihood of CRP. Because the number of observations for each customer i is different, we use $\mathcal{Y}_i = \frac{\sum_j y_{i,j}}{\sum_j x_{i,j}}$ as the

observation, which represents the average number of purchased products with unit price.

- **NHPP:** This model is based on [19], which assumes that the mixture component is NHPP with different types of intensity functions. However, this model describes the behavior over time. Here we convert the temporal information into price information. The price-purchase data can then be translated into the purchase matrix showing the occurrence of purchase events $y_{i,j}$ at each price interval.
- **Clustering:** This model is parametric segmentation, which includes classic clustering models **K-Means(KM)** and **Density Peak(DP)** [70]. The K-Means algorithm utilizes certain columns of the input table as features and clusters the original data into multiple categories based on the similarity calculation method specified by the user. DP is a clustering algorithm that aims to identify the density peaks in a dataset and assign each data point to its corresponding cluster based on its distance to the density peaks. In density-based clustering, each sample is represented as a point in an N-dimensional space, characterized by the number of neighbors within a specified radius and the minimum distance to other points with higher densities. The density is typically computed using a density formula based on the truncated distance:

$$p_i = \sum_{k=0}^n f(d_{ij} - d_c) \quad (3.13)$$

The symbol d_c refers to the specified radius size, while d_{ij} represents the distance between sample i and sample j .

In this experiment, the clustering is based on $a_{i,1}$ and $b_{i,1}$. We use linear regression to learn a and b from price and purchase quantities. The k value is set from 3 to 8, which is within the reasonable range [71].

Table 3.1: Evaluation results on synthetic data.

	ADIG	ADBG	Segmentation LL	CM accuracy
UtSeg	0.563	0.670	-1321500	0.383
UtSeg-1	0.626	0.616	-1523572	0.372
UtSeg-2	0.856	0.709	-2651849	0.301
UtSeg-3	1.077	0.816	-3247918	0.193
CRP-GM	0.687	0.532	-1650943	0.348
NHPP	0.572	0.730		0.427
K-means K=5	0.696	0.616		0.533
K-means K=7	0.580	0.690		0.394
Density peak k=5	0.670	0.689		0.405
Density peak k=7	0.665	0.765		0.376

Evaluation measurements:

- **Confusion matrix:** With ground truth data, we can use the confusion matrix (CM) to show the true and learned grouping. This evaluation is used on synthetic data as we often do not have grouping information for real data.
- **Clustering distance:** The average distance inside groups and the average distance between groups are used. The average distance inside groups (ADIG) refers to the average distance between sample points of the same group, *lower* is better. The average distance between groups (ADBG) refers to the distance between groups, which the *greater* distance means the larger difference between groups.
- **Segmentation Log-Likelihood(LL):** Segmentation LL can compare the fitness of different models. A *higher* log-likelihood value means the model fits better. To compare all models, we use the obtained group index. The likelihood is obtained by the Poisson distribution parameter on the average of the selected coefficients in each group. For CRP-GM, we double the likelihood as it only has one parameter [23].

3.5.2 Synthetic Data Set

We follow Eq.(3.9) to generate synthetic data for experiment. We generated pairs of purchased number and price for 100k customers. The evaluation results are shown in `tb: rsl-synthetic`. Firstly, UtSeg has the best result for both distances. Naturally, using the parametric method can obtain better results when the chosen parameters happen to be similar to the true parameter. However, without special design or knowledge, the parameters are unknown, which is the largest obstacle to use those parametric methods. On the contrary, our method can be generalized to unseen cases without such settings. Besides, our method can return the utility functions (as shown in `fig: realworld data`) from the function pool which other methods cannot except NHPP. The NHPP can return better results since it is the full model using stochastic process. It is more flexible but the inference is complex and slow.

In terms of computation time, the UtSeg model is compared with the optimized NHPP model as implemented in [19]. The result is shown in `subfig: pt`, with the same epoch in learning. UtSeg shows shorter computation time, and this is more significant when more data samples are used in the model. We also compare the convergence speed as shown in `subfig: ll`, which contains the normalised LL summarised over all data samples. The normalisation is to divide LL with the maximum LL for each method. UtSeg is about to converge after 150 epochs while NHPP does not show a sign of convergence until epoch 200.

3.5.3 Case Study

In this section, we present a case study on a real data set of 1,529,057 purchase transaction records collected by an Australian national supermarket chain in 2014 and we use 41,210 of them. For each transaction, there are customer id, item name and type, purchased quantity, and the corresponding price. Base on the variation of customer purchase quantities (normalized according to product volume) and purchase price(normalize as a discount rate, which is in the range of

[0,1]), we test the methods on four product types, including milk, chips, chocolate, and soft drink. We run the algorithm on each product type separately to see how the customers are segmented for their behavior. Our result can support the product providers to set promotion or stimulation. In the preprocessing, we further normalize quantity by purchased quantity per week subtracting the average amount throughout the year. This normalization can remove the influence of demand levels.

Table 3.2: Evaluation results on real data.

	Measure	UtSeg	UtSeg-1	UtSeg-2	UtSeg-3	NHPP
Milk	ADIG	0.61	0.74	0.86	1.08	0.65
	ADBG	0.90	0.88	0.86	0.67	0.83
	Segment LL	-1.3E+05	-1.5E+05	-1.9E+05	-2.5E+05	
Chips	ADIG	0.63	0.73	0.83	1.01	0.68
	ADBG	0.80	0.67	0.71	0.76	0.82
	Segment LL	-1.1E+05	-2.1E+05	-2.4E+05	-2.6E+05	
Chocolate	ADIG	0.55	0.59	0.71	0.75	0.58
	ADBG	0.83	0.79	0.74	0.77	0.79
	Segment LL	-4.5E+04	-1.8E+05	-2.0E+05	-2.4E+05	
Softdrinks	ADIG	0.56	0.69	0.74	0.83	0.54
	ADBG	0.67	0.62	0.57	0.52	0.59
	Segment LL	-6.9E+04	-1.4E+05	-1.9E+05	-2.1E+05	
	Measure	CRP-GM	KM-3	KM-7	DP-3	DP-7
Milk	ADIG	0.90	1.06	0.55	1.07	0.59
	ADBG	0.83	0.60	0.72	0.79	0.89
	Segment LL	-2.0E+05	—	—	—	—
Chips	ADIG	0.73	0.97	0.78	0.95	0.67
	ADBG	0.61	0.88	0.53	0.89	0.54
	Segment LL	-1.8E+05	—	—	—	—
Chocolate	ADIG	0.66	0.83	0.55	1.31	0.69
	ADBG	0.75	0.58	0.72	0.58	0.81
	Segment LL	-1.8E+05	—	—	—	—
Softdrinks	ADIG	0.79	0.93	0.49	1.34	0.70
	ADBG	0.53	0.56	0.77	0.53	0.90
	Segment LL	-1.6E+05	—	—	—	—

The results are given in 3.2. In the study, the UtSeg model provides better results in most cases, without setting functions and parameters using domain knowledge. The results are consistent with what we observed in the synthetic data. Also, UtSeg can provide us the utility functions for customers in each group. We listed 5 utility functions for each product type in fig: realworld data.

3.6 Conclusion

In this chapter, we propose a bayesian nonparametric (BNP) method for customer segmentation without prior knowledge of their purchase behavior utility functions. By utilizing the semi-parametric approach, we unify the parameters of different utility functions into the same representation, enabling them to be generated by the same distribution. This approach by using domain knowledge (utility functions) significantly reduces the effort required to design a special inference algorithm, allowing for efficient learning of the latent variables. Furthermore, this approach makes it easier to generalize our method to accommodate more utility functions. Our work involves incorporating rule-based domain knowledge into the model, which not only improves its performance but also enhances trust in the model due to the endorsement of expert' domain knowledge. If experts can get feedback information about the model results and understand it, they can provide more accurate domain knowledge based on their understanding, so that the model learning effect is better. After completing this work, we will conduct research on the interaction between human and models, promoting co evolution between human and machine learning. Please refer to the next chapter for specific content.

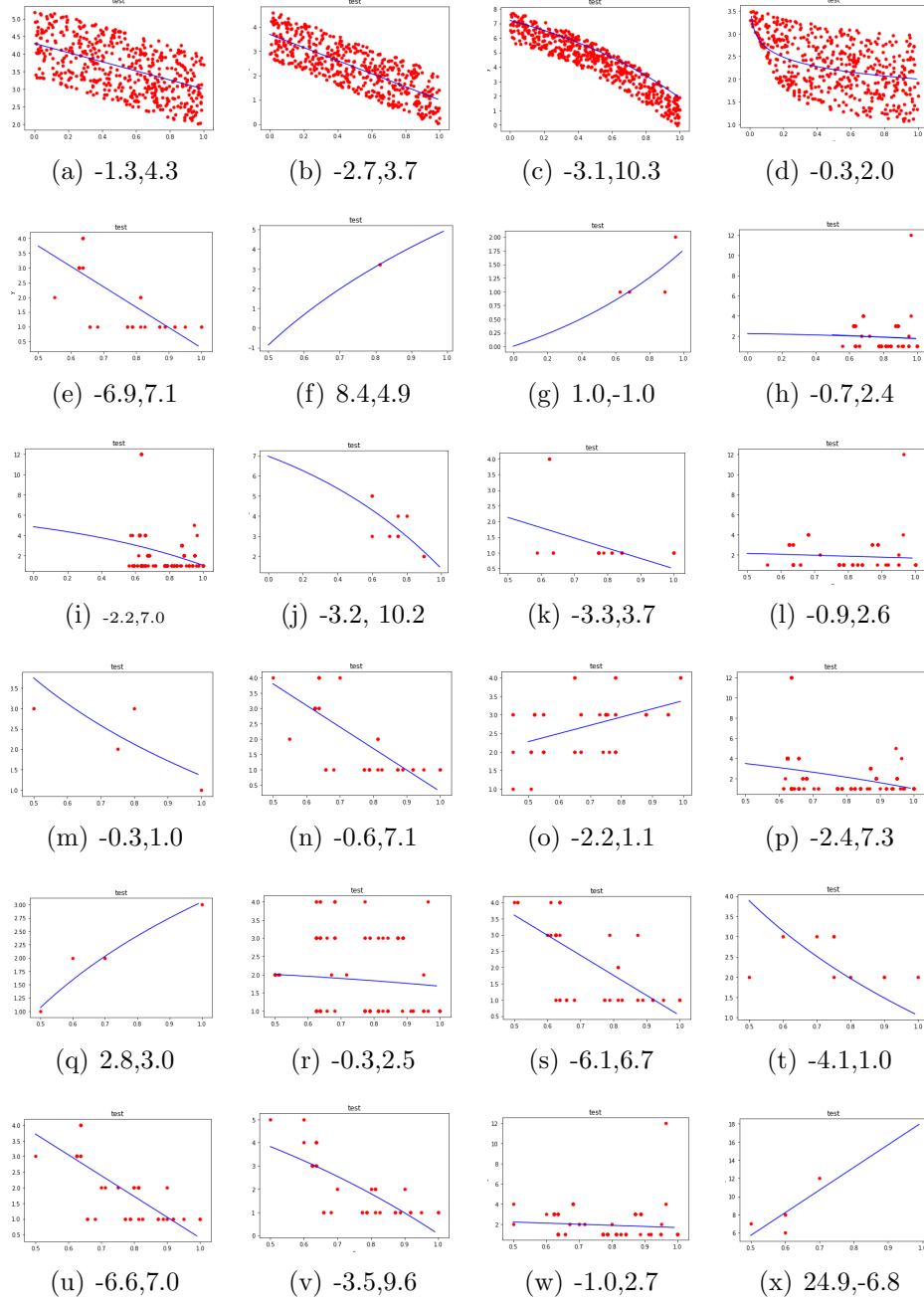
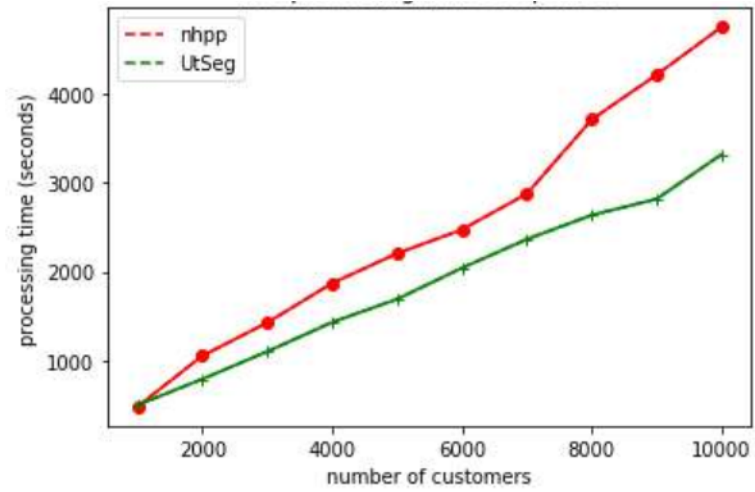
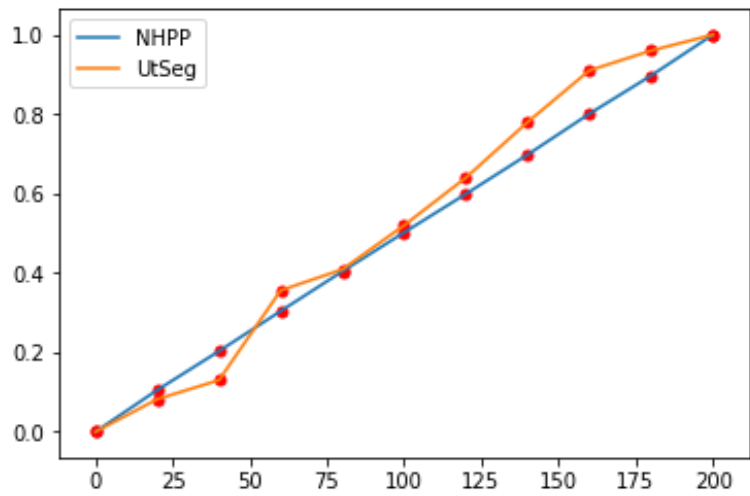


Figure 3.3: Utility functions for different data sets: (a)-(d) are for synthetic data, (e)-(x) are for four product types. 5 utility functions and observed points are shown for each type. The learned parameters α_k and β_k are under the plots.



(a) Computation time



(b) Normalized LL summation

Figure 3.4: Comparison results of model efficiency.

Chapter 4

Co-teaching accountable learning framework with distilled and domain knowledge

In Chapter 4, we provide a solution to help experts interact with models, which can improve model performance (adapt to biased data) and increase experts' confidence in the model due to the interaction between experts and models. In our work, we conducted relevant research on medical data and propose the co-teaching with dual-knowledge: accountable learning framework with distilled and domain knowledge as a solution. We have verified our solution using medical dataset, which demonstrates that our approach is effective.

4.1 Motivations

To avoid overfitting and gain robustness, high-capacity models, such as deep neural networks, require training with a large number of examples and substantial computing power. These machine learning models typically excel in perception tasks as they are designed to fit the given data. However, they often struggle with logical reasoning and are susceptible to the influence of noisy or biased data, which can significantly impact their learning process. On the other hand, humans have the ability to learn correct knowledge from a small number of samples through reasoning (Solso, 2005). Given the powerful perception abilities of modern machine learning models, is it possible to enhance them to include similar reasoning capabilities as humans?

Existing machine learning models have explored perception or reasoning individually (Santoro et al., 2017), but combining both processes is challenging due to its complexity. Perception-based neural networks (Santoro et al., 2017) have

attempted to incorporate reasoning, but full logical reasoning is still lacking (Dai et al., 2019). Conversely, statistical relational learning (SRL) methods (Koller Friedman, 2007) learn perception-driven reasoning but require computationally expensive processing.

In this chapter, we propose a new machine learning framework called co-teaching learning, which integrates dual knowledge from source data (perception) and domain knowledge (reasoning) into the target model during the learning process. By incorporating reasoning knowledge from domain experts, the learning process can be enhanced with critical and accurate information. It is important to note that our framework does not aim to design a specific model that learns both perception and reasoning simultaneously. Instead, the co-teaching framework allows experts to contribute reasoning knowledge to the model, enabling effective learning even when the available data is scarce. Co-teaching (Cook Friend, 1995) is originally an educational concept that involves assigning multiple educators with complementary strengths to instruct the same group of students. The core idea of co-teaching is to leverage each educator’s unique skills and perspectives. Similarly, our co-teaching framework leverages both a teacher model, which may contain inaccurate knowledge from a noisy or biased dataset, and domain experts, who can provide more reliable but limited information. These two sources of knowledge, referred to as perception and logical reasoning, guide the training of the student (targeted) models.

The implementation of the framework faces three significant challenges. The first challenge is the issue of knowledge disproportion, where the impact of supportive data and domain knowledge needs to be evenly considered. However, the loss-based learning mechanism tends to prioritize the knowledge from supportive data due to the overwhelming volume of data compared to the limited domain knowledge provided by experts. To address this issue, we use a combined loss function based on few-shot learning to magnify the importance of domain knowledge, allowing us to control the balance between both knowledge sources.

The second challenge is the issue of abstruseness, which hinders domain experts from understanding how well the model has been trained. Domain experts need to comprehend the inner workings of the model instead of dealing with a black-box model. This issue also gives rise to trust problems, as users may not trust the model’s outcomes even if they are empirically accurate. To tackle this challenge, we incorporate an explainable component into the framework, which assists experts in interpreting the model’s status during training.

The third challenge is the issue of accessibility, where the prevailing autonomous learning process does not provide an interface for experts to provide insightful feedback. Experts are limited to either passing knowledge to model designers or using simplified or one-time knowledge acquisition methods, such as priors in Bayesian models. Our proposed framework overcomes this challenge by enabling experts to easily access the training process and provide feedback, thereby enhancing accessibility.

As shown in 4.1, our co-teaching framework consists of three components: the teacher model, experts, and the student model. Specifically, our work makes the following contributions: **1.** We propose the framework to involve both the supportive data and the domain knowledge into the learning process. **2.** Our framework is the first solution to consider how experts impart the critical knowledge for reasoning with feasibility and convenience, without the requirement to specify the model to infer the reasoning. **3.** Base on our co-teaching framework, we investigated how perception and reasoning interactively affect each other during the learning process.

4.2 Problem Formulation

When the general aim of machine learning is to estimate the distribution $P(X, Y)$ based on the sampled data $D = (X, Y) \in \mathbb{R}^{n \times p} \times 0, 1^n$, the motivation behind our framework is that D is too small to empirically estimate a good $P(X, Y)$ (student model s) with certainty. A straightforward solution is to add more data

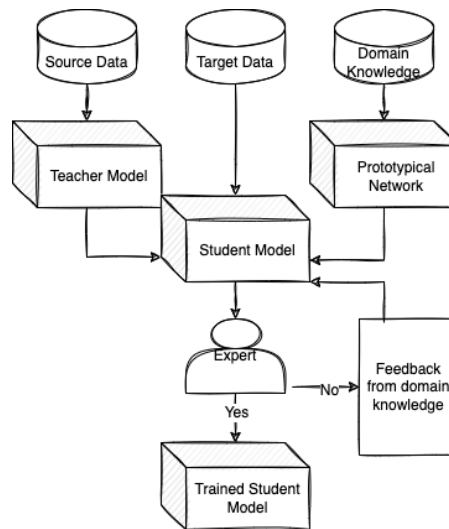


Figure 4.1: The general framework of our co-teaching

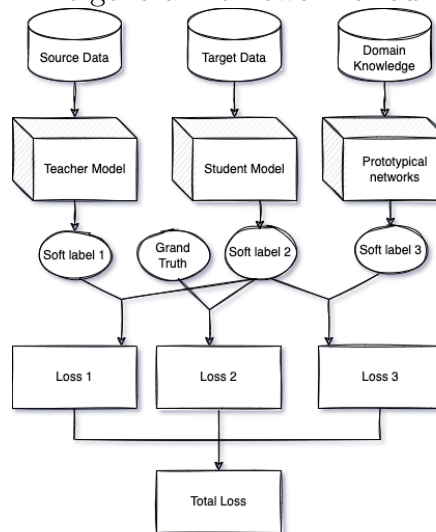


Figure 4.2: Loss function process

or seek help from domain experts. In order to incorporate knowledge from both sources during training, our problem is how to construct a co-teaching framework that engages domain experts in the learning process. This framework should provide accessible understanding from s , allow for feedback to s , and maintain a balance in learning s . The teacher model t is trained with a larger (but possibly shifted) dataset $\hat{D} \sim P(X, \hat{Y})$, and a transfer-learning-style method is employed to endorse s . Simultaneously, an interactive-style method is used to facilitate communication between domain experts and s .

4.3 Co-Teaching with Dual-Knowledge

Learning of s with data defects in our framework is not a one-off but a loop as shown in fig: High-level. We will start with an initial s learned on D , explain it to the expert, and use the expert’s feedback as D^* to provide logical reasoning to s by updating s with t . We will examine how t to s is designed, how experts provide logical reasoning to s , and how to learn s .

4.3.1 Perception from Teacher Model

We implement knowledge distillation [27] as the teacher model t . Given a sample $(\vec{x}_i, \hat{y}_i) \in \hat{D}$, knowledge distillation uses the difference between the prediction scores q_i^T and p_i^T as a soft loss, referred to as the *teacher loss*. Here, $q_i^T = \Psi(t^*|\vec{x}_i, \hat{y}_i)$ represents the score of the teacher model t , and $p_i^T = \Psi(s|\vec{x}_i, \hat{y}_i)$ represents the score of the student model s . The function $\Psi(\cdot|\vec{x}_i, \hat{y}_i)$ maps any model \cdot to a score within the range of $[0, 1]$. For example, the last layer of a neural network before the softmax function can be used as the score to obtain Ψ . The parameter T is a hyper-parameter used to soften the logits, effectively smoothing out the probability distribution [27].

In the simplest form of distillation, knowledge is transferred to the distilled model by training it on a transfer set, where a high-temperature softmax is used with the cumbersome model to generate soft target distributions for each case.

During the training of the distilled model, the same high temperature is employed initially, and later a temperature of 1 is used. Further improvement can be achieved when correct labels are available for some or all of the transfer set. One approach is to modify the soft targets using the correct labels, but we found that a better approach is to use a weighted average of two different objective functions.

The first objective function is the cross-entropy with the soft targets, computed using the high temperature softmax of the distilled model. The second objective function is the cross-entropy with the correct labels, computed using the same logits in the softmax of the distilled model but at a temperature of 1. Typically, the second objective function is given a considerably lower weight for optimal results. Since the gradients produced by the soft targets scale as $1/T^2$, it is crucial to multiply them by T^2 when using both hard and soft targets. The derivation process is as follows:

- Soft Target: L_{soft}

$$\begin{aligned} L_{soft} &= - \sum_j^N p_j^T \log(q_j^T) \\ &= - \sum_j^N \frac{z_j/T \times \exp(v_j/T)}{\sum_k^N \exp(v_k/T)} \left(\frac{1}{\sum_k^N \exp(z_k/T)} - \frac{\exp(z_j/T)}{\left(\sum_k^N \exp(z_k/T)\right)^2} \right) \\ &\approx - \frac{1}{T \sum_k^N \exp(v_k/T)} \left(\frac{\sum_j^N z_j \exp(v_j/T)}{\sum_k^N \exp(z_k/T)} - \frac{\sum_j^N z_j \exp(z_j/T) \exp(v_j/T)}{\left(\sum_k^N \exp(z_k/T)\right)^2} \right) \end{aligned}$$

- Hard Target: L_{hard}

$$L_{hard} = - \sum_j^N c_j \log(q_j^1) = - \left(\frac{\sum_j^N c_j z_j}{\sum_k^N \exp(z_k)} - \frac{\sum_j^N c_j z_j \exp(z_j)}{\left(\sum_k^N \exp(z_k)\right)^2} \right)$$

Since the magnitude of $\frac{\partial L_{soft}}{\partial z_i}$ is approximately $\frac{1}{T^2}$ times that of $\frac{\partial L_{hard}}{\partial z_i}$, when using both soft target and hard target simultaneously, it is necessary to multiply the soft target by a coefficient of T^2 beforehand to ensure that the gradient contributions

from the soft target and hard target are roughly consistent.

This ensures that the relative contributions of the hard and soft targets remain approximately unchanged, regardless of variations in the distillation temperature during experimentation with meta-parameters.

The *teacher loss* $\mathcal{L}_t(s)$ and *student loss* $\mathcal{L}_s(s)$, using the cross entropy, can be written into:

$$\begin{aligned}\mathcal{L}_t(s) &= - \sum_i p_i^T \log(q_i^T), \\ \mathcal{L}_s(s) &= - \sum_i y_i \log(p_i^T)\end{aligned}\tag{4.1}$$

An illustration of our framework using a simple synthetic learning problem is shown in fig:case. In this example, we show that t can be varied from the true distribution of s so that s learned using t is also varied from the true distribution.

4.3.2 From Explanation to Feedback

After each round of optimization of s , updated s will be explained to the expert, and the expert will provide some feedback. Model explanation and domain expert feedback are ambiguous terms that are inconvenient to re-productivity. For this purpose, we will formulate these two critical steps. The definitions will define the whole process, what an explanation should be and how good the expertise of feedback is. First, the whole explanation-feedback component can be taken as a stimuli-reaction mapping process (SRMP) which is defined as:

Stimuli-Reaction Mapping Process (SRMP) Given a model s trained by D , the stimuli-reaction mapping $b : e(s, D) \rightarrow D^*$ is a human-centered mapping, where e is a model explanation function, worked as the stimuli to explain s to experts. $D^* \sim P^*(X, Y)$ is the expert's feedback as the reacting, but in the instance space (same to (X, Y)). Then for any expertise agent A , we can represent A 's stimuli-reaction mapping process at round j as $D_j^* = b_A(e, j)$.

Here e could be a simpler model [44] or some meaningful rules. The model

explanation itself is for human users to understand the decisions made by the model. In the process, we hope the explanation function e is good. However, how to evaluate a model *explanation* is a yet-to-answer question. There are several optional targets, such as fairness, soundness and accuracy. Here we use accuracy as the target, which is to make sure the explanation correctly (accurately) reflects s . We proposed the ε -accurate method to measure the accuracy of the model explanation, which is defined as follows:

[ε -Accurate Model Explanation Function] Given a model $s \in \mathcal{H}$ trained by D , and a model explanation function $e \in \mathcal{E}$, with a mapping $v : e \rightarrow e_s \in \mathcal{H}$. If $\varepsilon = 1 - \frac{\ell(e_s(X), s(X))}{\ell(1-Y, Y)}$, the explanation function e is ε -accurate. Here ℓ is a loss function.

Here \mathcal{H} is the hypothesis space containing all possible models. In the definition, v exists in most explanations, for example, rules can be programmed, or simpler models can be used directly. Here ε is defined for empirical calculation. This definition can be used to evaluate most explanation methods, and it can be used to know if the explanation is good. In SRMP, ε can also be shown to an expert to build his/her confidence. When ε is low, we may not obtain good D^* from $b_a(e)$, which means the convergence may need more rounds. In our implementation, we define a tree function to explain the model. Here ε is proportional to the tree depth. It selects the most important features and indicates the most possible label values for certain intervals of the feature. Then more features or more detailed intervals are named after the previously named feature, and the most possible labels are given. More explanation methods, such as LIME [44], can also be employed here and verified by ε -Accurate.

Then the expert will give his/her feedback. The feedback can be anything, including rules or data, but we assume that they can all be projected into the instance space by SRMP. So what is the difference between experts' feedback and adding more randomly sampled data? And why do we say the feedback is expertise? Here we list two critical properties that the expertise feedback

must have, which are criticalness and correctness. Strictly speaking, these two properties are the necessary conditions for feedback expertise, but we can use them to evaluate the expertise in an unsophisticated manner.

[Criticalness of Feedback] The expert's feedback is limited in size but aims at the most critical part. Given X that is provided by experts, we have $P(P(s(X) = 1 - Y) - P(s(X) = Y) \geq q_0) \geq \gamma_s$, and $-\epsilon < q_0 < 100\%$, higher γ_s means X is more critical to s .

If s has not learned X , then $P(s(X) = Y)$ is either lower (wrong prediction with $P(s(X) = 1 - Y) - P(s(X) = Y) > 0$) or similar (random guess with $|P(s(X) = 1 - Y) - P(s(X) = Y)| < \epsilon, \epsilon \geq 0$, e.g. we let $\epsilon = 5\%$) with $P(s(X) = 1 - Y)$. In other words, critical feedback tends to give the information that has not been learned by s , so an expert can help s learn quicker than randomly sampled data. Because of the criticalness, the expertise feedback even needs fewer convergence rounds than in similar settings such as active learning. For the uncertainty case, when s is not continuous, active learning does not know where the discontinuous points are because they can only ask questions for the learned part using the smoothing assumption. But these points are critical for learning. Furthermore, wrong (or small) data could cause over-confidence in active learning to select the wrong data to update. Here we can use $\max(\gamma_s)$ as the quantification of how expertise the feedback is for criticalness, when $\gamma_s = 100\%$, the feedback can improve s if it is correct. Otherwise, an **ordinary person** will provide feedback with $\gamma_c = 50\%$, and an **ignoramus person** will have $\gamma_s < 50\%$, which means the feedback is redundant.

[Correctness of Feedback] Given the true distribution $P(X, Y)$, the feedback is $(X, Y^*) \sim P^*(X, Y)$. we have $P(Y|X) > 50\%$ and $P(P(Y^* = Y|X) > P(Y^* = 1 - Y|X)) \geq \gamma_{c_0}$. An expert's feedback will be correct given $\gamma_c = \max(\gamma_{c_0}) > 50\%$ ¹.

We can use γ_c to evaluate the correctness. An **ordinary person** will have

¹It is assumed the data is balanced between both classes.

$\gamma_c = 50\%$, and **ignoramus person** has $\gamma_c < 50\%$. So their feedback is not recognized as correct.

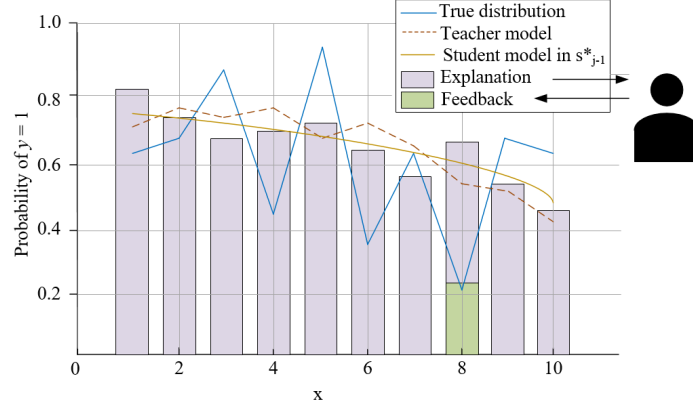


Figure 4.3: The example of using co-teaching framework on iteration j .

Algorithm 1: Co-teaching algorithm

Input: Teacher model t based on dataset \hat{D} ,
expert agent a with a feedback function b_a

Output: Student model s^*

```

1 Initial  $s$  with  $D$ ;
2 while  $e(s, D)$  is unaccepted by  $a$  do
3   Learn  $e$  with  $s$  and  $D$ ;
4   (optional) Test  $e$  with  $\varepsilon - Accuracy$ ;
5   Obtain  $D^*$  from  $b_a(e(s, D))$ ;
6   Update  $s$  by minimising:  $\mathcal{L}$ ;
7   Break if  $s$  does not change;
8 end
9  $s^* = s$ ;

```

Figure 4.4: Algorithm

In the illustration of fig:case, the explanation (e) is interpreted as discretizing the model for easier understanding ($e_s = e$), which are the purple bars. Using mean-absolute error as the loss as $\ell_e = \sum_{u \in U} |p(u) - e_s(u)|$, where U are the 10 middle points of discretised intervals on x . Then ε is $\frac{1}{10}\ell_e$. After reviewing the explanation, experts may give feedback of $P(Y = 0|X = 8) = 0.85$. In the ground truth, $P(Y = 0|8) > P(Y = 1|8)$, so $\max(\gamma_c)$ is 100% (1 out of 1 feedback is correct). $P(s(8) = 1) - P(s(8) = 0) \approx 0.2$, so $\max(\gamma_s = 100\%)$ (1 out of 1

feedback is critical). The two values show that the feedback could be expertised.

4.3.3 Knowledge Passing from Expert

The knowledge provided by experts is much less in data volume than that by the teacher model: $|D^*| \ll |\hat{D}|$. To boost the feedback, we use Prototypical Network [31] as a few-shot learning model. The model is to find the prototype center of each category from D^* in the label space. Given $(\vec{x}, y^*) \in D^*$, the centers \vec{c}_k for class labels k can be learned by

$$\vec{c}_k = \frac{1}{\sum_i \delta(y_i^* - k)} \sum_{y_i^* = k} f(\vec{x}_i) \quad (4.2)$$

Here f is optional for non-linear mapping. Then the distance function d between data (\vec{x}, y) used by s and \vec{c}_k can be used to obtain the loss as the *expertise loss* through softmax:

$$\mathcal{L}_e(s) = \sum_i (1 - P(s(\vec{x}_i) = y_i)) = \sum_i \left(1 - \frac{\exp(-d(f(\vec{x}_i), \vec{c}_{y_i}))}{\sum_{k'} \exp(-d(f(\vec{x}), \vec{c}_{k'}))} \right) \quad (4.3)$$

4.3.4 Optimization

To facilitate the co-teaching framework, we assemble the components with an additional classifier. Each component is represented by a loss. The overall loss function L is formulated as:

$$L = \lambda_t * \mathcal{L}_t(s) + \lambda_s * \mathcal{L}_s(s) + \lambda_e * \mathcal{L}_e(s) \quad (4.4)$$

For the training of co-teaching, Fig 4.2 shows the learning steps for overall loss function L . $\lambda_t, \lambda_s, \lambda_e$ can be updated as hyper-parameters. There are many strategies to set them. For example, increase or decrease them based on the convergence speed. In the testing, we use Adam [72] as an optimizer to solve the optimization problem.

One of the key components of the Adam algorithm is its use of exponential weighted moving averages to estimate the momentum and second moments of the gradients. In other words, it utilizes state variables to track and update these estimates:

$$\begin{aligned}\mathbf{v}_t &\leftarrow \beta_1 \mathbf{v}_{t-1} + (1 - \beta_1) \mathbf{g}_t, \\ \mathbf{s}_t &\leftarrow \beta_2 \mathbf{s}_{t-1} + (1 - \beta_2) \mathbf{g}_t^2.\end{aligned}$$

Here, β_1 and β_2 are non-negative weighting parameters. They are often set to $\beta_1 = 0.9$ and $\beta_2 = 0.999$, respectively. This means that the moving average estimate of the variance is updated much more slowly than the moving average estimate of the momentum. Note that if we initialize \mathbf{v}_0 , we will get a significant initial bias. This can be solved by using $\sum_i 0^i \beta^i = \frac{1-\beta^t}{1-\beta}$. The normalized state variables are obtained by the following equation:

$$\hat{\mathbf{v}}_t = \frac{\mathbf{v}_t}{1 - \beta_1^t} \text{ and } \hat{\mathbf{s}}_t = \frac{\mathbf{s}_t}{1 - \beta_2^t}$$

With the correct estimates, we can now write down the update equation. First, we rescale the gradients in a similar manner to the RMSProp algorithm to obtain the updated values.

$$\mathbf{g}'_t = \frac{\eta \hat{\mathbf{v}}_t}{\sqrt{\hat{\mathbf{s}}_t} + \epsilon}$$

gradients themselves. Additionally, there is a slight difference in the rescaling factor, where we use $\frac{1}{\sqrt{\hat{\mathbf{s}}_t} + \epsilon}$ instead of $\frac{1}{\sqrt{\hat{\mathbf{s}}_t}}$. The former has shown slightly better performance in practice, differentiating it from the RMSProp algorithm. Typically, we choose $\epsilon = 10^{-6}$ to strike a balance between numerical stability and fidelity. Finally, we can summarize the update as follows:

$$\mathbf{x}_t \leftarrow \mathbf{x}_{t-1} - \mathbf{g}'_t$$

Looking back at the Adam algorithm, its design inspiration is quite clear. Firstly, the momentum and scaling are clearly visible in the state variables. The unique definitions of these variables allow us to remove biases (which can be corrected by

slightly different initialization and update conditions). Secondly, the combination of the two terms in the RMSProp algorithm is straightforward. Lastly, the explicit learning rate η enables us to control the step size to address convergence issues.

4.4 Experiments

We conducted experiments to examine the performance of our framework. There are three parts. First, we will discuss the general performance and usefulness of each component. Then, we will examine where our framework can resolve the abstruseness issue. At last, we will assess whether our feedback is better than other methods that provide interaction during learning.

Datasets: Dataset 1(D1) [73] is a Heart disease dataset that contains 1026 patients and 14 attributes, including age, sex, resting blood pressure, etc. The label refers to the presence of heart disease in the patient. Dataset 2(D2) [74] is from the National Institute of Diabetes and Digestive and Kidney Diseases with 768 patients and 9 attributes. The objective is to predict whether a patient has diabetes based on diagnostic measurements. Dataset 3(D3) [75] is a stroke dataset according to the World Health Organization. This dataset is used to predict whether a patient is likely to get stroke based on the input parameters like gender, age, various diseases, and smoking status. Dataset 4(D4) [76]: The Music Emotion in 2015 database is a large music classification dataset. We extract tempo, mode, brightness, and loudness as features with MIR toolbox [77].

Settings: We randomly pick 10% data from training data for training student model, 80% data as training data for the teacher model, and 10% data as training data for experts. Then we manipulate the training data to create the learning scenarios with teacher models, student models, experts, and baseline models. We use below settings:

① +: Training data is the same as the original data; ②−: Training data shuffled 50% of label from the original data; ③ S : Student model which is a basic BP network, here S^+ represents a good student which trained by +data while

S^- represents a student cannot learn well which trained by $-$ data; ④ T : Teacher model is a basic BP network, here T^+ represents a teacher gives correct knowledge which trained by $+$ data while T^- will teach biased knowledge which trained by $-$ data; ⑤ T : Expert is a prototypical networks, here L^+ represents a expert gives correct domain knowledge which trained by $+$ data while L^- will teach biased domain knowledge which trained by $-$ data; We also apply levels of expertise by setting: ignoramus person(L^-): $\gamma_c = 40\%$ correct label; ordinary person(L^0): $\gamma_c = 60\%$ correct label; expert(L^+): $\gamma_c = 100\%$ correct label; ⑥ AL : Variational Adversarial Active Learning model (VVAL) [78]: VVAL is a pool-based active learning algorithm that learns the sampling mechanism in an adversarial manner; ⑦ DA : Domain-Adaptive Few-Shot Learning(DAPN) [79]: DAPN can explicitly strengthen source/target classification and alleviate the negative impact of domain alignment on FSL; ⑧ $Noisy$: DivideMix: Learning with Noisy Labels as Semi-supervised Learning [80]: DivideMix is a framework for noisy label learning which uses a mixture model to model the loss distribution of distinct clean samples and noisy samples. We use this method because we provide the $-$ dataset. We will see how a noise-free model can correct the results.

4.4.1 Performance of Co-teaching

This experiment is to study the general performance of the co-teaching framework under various settings in knowledge disproportion. We compare the accuracy, F1 score, recall, and precision of prediction in scenarios including different components of the co-teaching model and baseline methods.

The results are shown in Table 4.1 and Table 4.2. Our experiments are carried out in two main streams, with S^+ as a common case, and S^- as a biased case. The challenge of S^+ is from data sparsity, while S^- may be affected by wrong data. We have some observations: ① With S^+ , the scenario with T^+ and L^+ is proved to be significantly boosting the performance, this is because of the full knowledge of perception and reasoning. The $T^+S^+L^+$ provides the best

performance, better than S^+ , AL^+ , $Noisy^+$ and DA^+ in the most of the time. ②When we removed the perception (S^+L^+) or reasoning (T^+S^+), we can see the effects of each component. We observed that combining them will improve even more than the summarising of using each. ③With providing S^- , another scenario with biased data. In our test, we can see that putting T^+ and L^+ are the best case, comparing to any of these components provides biased knowledge, and all the baselines using $-$ data. ④We also see the effectiveness of each component, even though the data from S^- is biased. Here $T^-S^-L^+$ and $T^+S^-L^-$ are both shown, helping us further understand the interactive effects between perception and reasoning. The results show that even though all other components are biased, individual knowledge source is still reliable to improve performance.

4.4.2 Co-teaching for Abstruseness

The second experiment is designed to show whether co-teaching is helpful for abstruseness issues. The intuitive setting is to variate the explanation part to compare the results from useful explanation and useless explanation. However, explanation evaluation itself is an uncertain area under exploration [81]. Therefore, we will use our three-level experts to test the performance of solving abstruseness. Among these different level experts, different people will provide information with different accuracy. This scenario simulated knowledge from people to estimate how useful the explanation is.

Our results from L^+ can then represent the expert whose input is a useful explanation. While the L^- is neither critical nor correct. We show the convergence speed (rounds of interaction), accuracy, and explanation (the range of selected features) in 4.3.

The results show that the range of L^- is unstable and that the overlappings between consecutive rounds are smaller than others. The accuracy is also lower than others. Sometimes the accuracy is even lower in a new round. L^0 can provide a better solution for each round, however, the convergence is slow. Through the

comparison between L^0 and L^+ , we can evaluate the effectiveness of explanation that can connect experts into the model.

4.4.3 Co-teaching for Accessibility

The accessibility issue is that the prevailing autonomous learning process does not provide an interface for experts to provide insightful feedback. The experiment will show that if experts can provide feedback during the learning, it will accelerate the learning when data is scarce.

We set up a scenario with scarce data, using 50 data points as training data and 10 correct examples from an expert in each round. The feedback setting is used throughout all methods, including our baselines, but the updated examples are different. We have shown that our final performance is better in experiment 1. However, we use delta accuracy in the evaluation to emphasise the connection between feedback and convergence. The delta accuracy is the difference between the model accuracy (evaluated on an additional test dataset) of consecutive rounds. In Fig 4.5, the co-teaching model ($T^+S^+L^+$) has the best results since all delta accuracy are greater than 0, which means all feedback can help the model. However, active learning and domain adaption have negative delta accuracy, which will converge slower, and their feedback can impede learning. In the case of few data samples, expert feedback will help the rapid learning of the model.

4.5 Conclusion

In this chapter, we propose the co-teaching framework to enable the learning process to consult both the distilled knowledge and domain experts. The framework is feasible and convenient for experts to understand the status of learning so that they can provide their critical and correct knowledge to model. It is different from other learning regimes in that the provided knowledge guides the learning instead of modeling or data. We further compare it with other similar frameworks and observe how machine learning model can learn from experts. We further expect

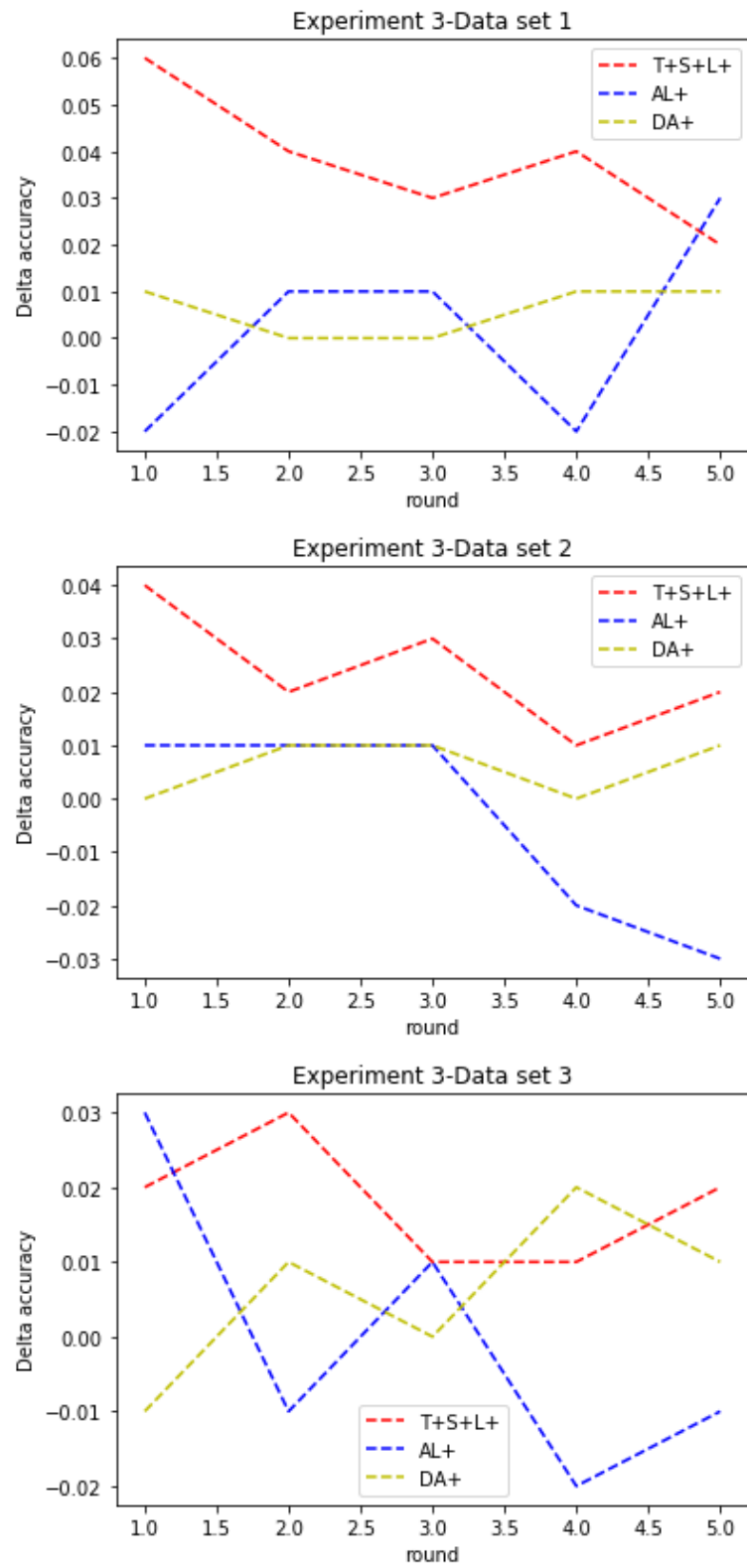


Figure 4.5: Co-teaching for Accessibility

that the framework can be a fundamental design for machine learning for application, which can bring closer collaboration between machine-learning researchers and domain experts.

This work has initially completed a framework for human and machine learning to progress together, but due to the complexity of machine learning feedback (model explanation), such co-evolution can only be limited to machine learning model and machine learning experts. In the next chapter, we aim to study the feedback from machine learning models to non-machine learning experts in coevolution.

Table 4.1: Performance of Co-teaching(different part of co-teaching)

Dataset	Metric	S^+	$T^+S^+L^+$	S^-	$T^-S^-L^-$	$T^+S^-L^-$	S^-L^-	$T^-S^-L^+$	T^+S^-	$T^+S^-L^+$
D1	acc	0.68	0.83	0.54	0.6	0.62	0.58	0.62	0.66	0.67
	f1	0.51	0.85	0.52	0.61	0.59	0.69	0.63	0.7	0.73
	recall	0.51	0.85	0.46	0.58	0.55	0.85	0.63	0.75	0.8
	precision	0.52	0.84	0.76	0.66	0.63	0.69	0.67	0.66	0.67
D2	acc	0.68	0.73	0.55	0.65	0.67	0.63	0.68	0.69	0.71
	f1	0.78	0.67	0.42	0.43	0.64	0.44	0.35	0.46	0.62
	recall	0.59	0.79	0.59	0.68	0.85	0.63	0.45	0.59	0.69
	precision	0.53	0.6	0.32	0.31	0.55	0.37	0.29	0.41	0.58
D3	acc	0.65	0.85	0.51	0.54	0.57	0.53	0.55	0.63	0.64
	f1	0.55	0.76	0.67	0.69	0.52	0.47	0.66	0.7	0.64
	recall	0.65	0.74	0.91	0.75	0.45	0.43	0.69	0.79	0.77
	precision	0.68	0.79	0.54	0.71	0.66	0.66	0.65	0.67	0.57
D4	acc	0.7	0.73	0.52	0.54	0.64	0.54	0.58	0.63	0.71
	f1	0.64	0.69	0.53	0.48	0.61	0.55	0.62	0.63	0.71
	recall	0.64	0.69	0.53	0.48	0.61	0.55	0.62	0.65	0.77
	precision	0.9	1	0.54	0.52	0.6	0.52	0.67	0.7	0.83

Table 4.2: Performance of Co-teaching(baseline models)

Dataset	Metric	AL^+	AL^-	$Noisy^+$	$Noisy^-$	DA^+	DA^-	Dataset	Metric	AL^+	AL^-	$Noisy^+$	$Noisy^-$	DA^+	DA^-
D1	Acc	0.7	0.65	0.69	0.63	0.75	0.49	D3	Acc	0.74	0.61	0.65	0.6	0.82	0.47
	F1	0.73	0.46	0.74	0.64	0.75	0.6		F1	0.41	0.26	0.33	0.27	0.36	0.29
	Recall	0.71	0.45	0.9	0.77	0.68	0.69		Recall	0.32	0.35	0.43	0.38	0.27	0.36
	Precision	0.76	0.47	0.62	0.57	0.87	0.57		Precision	0.6	0.23	0.3	0.25	0.62	0.26
D2	Acc	0.73	0.55	0.66	0.58	0.78	0.64	D4	Acc	0.73	0.66	0.7	0.65	0.72	0.52
	F1	0.81	0.56	0.39	0.48	0.64	0.22		F1	0.74	0.56	0.74	0.74	0.68	0.36
	Recall	0.93	0.5	0.3	0.37	0.69	0.25		Recall	0.63	0.46	0.65	0.58	0.57	0.26
	Precision	0.73	0.68	0.65	0.69	0.68	0.27		Precision	0.82	0.73	0.85	0.72	0.83	0.57

Table 4.3: Co-teaching for Abstruseness

accuracy	round 1	round 2	round 3	round 4
L^-	0.46/(95, 185)	0.44/(83, 145)	0.48/(125, 194)	0.45/(83, 172)
L^+	0.58/(100,200)	0.58/(103, 196)	0.61/(97, 189)	0.62/(105, 200)
L^0	0.55/(121, 198)	0.56/(117, 180)	0.58/(100, 181)	0.59/(118, 189)

Chapter 5

An explainable framework for multimodal information

In the previous chapter, we encountered many difficulties in providing feedback information from the model. In most cases, feedback from machine learning, including the model’s explanation, is designed for machine learning experts. For non-machine learning experts, these feedback explanations can seem obscure and difficult to understand. In Chapter 5, we propose a solution to help non-machine learning experts provide information to the model and use multimodal information to obtain more intuitive explanations of the model. This will enable non-machine learning experts to understand the model and choose to trust it. In this work, we conducted relevant research on emotion recognition for music and proposed an explainable multimodal privileged information solution. We verified our solution using a musical dataset, which demonstrated that our approach is effective. We also studied people’s trust in the model using a questionnaire.

5.1 Motivations

The field of science and technology has witnessed substantial advances, primarily driven by the central role of machine learning. However, it is often overlooked that humans play a crucial role in this process. If machine learning is to be deployed as a reliable tool, it must earn the trust of its users. Several methods have been proposed to bolster users’ confidence in machine learning [44]. Unfortunately, many of these approaches require a background in machine learning or considerable domain expertise to comprehend the model’s explanations. Traditional explanation methods include rephrasing the model or using data to clarify its impact (LIME utilizes rephrasing and data), or selecting features to explain.

Features can be close to the factors that need to be understood, but they often require conversion to a space that is familiar to the audience. For instance, even in image analysis, providing explanations using just pixels can be challenging to comprehend. Existing methods often extract meaningful parts of an image to explain.

In music, privileged information may include the main melody and five dimensions of the music. Experts can label the main melody, and the five-dimensional information can be extracted from the music. This privileged information is only used during the training process. A simple approach involves combining the privileged and music information in the same list and using a neural network to learn from them. However, a significant challenge is that privileged information is often unavailable during testing. Obtaining privileged information during testing is also constrained by conditions. We can build trust with the user during training by utilizing an interactive learning framework with privileged information as an explanation. This approach can help enhance the user’s understanding of the model, whereas traditional explanations usually occur at the prediction stage. The framework faces three major challenges during implementation:

The first challenge we face is establishing trust between machine learning models and musicians or music enthusiasts. Understanding which aspects of the input drive the model’s decisions is difficult, making it challenging to explain the predictions. In traditional machine learning, users can observe and correct the system’s predictions, but the predictions are not typically explained to them. As a result, users may struggle to trust the prediction if the explanation is unclear or inaccurate. To tackle this issue, interactive machine learning methods, such as those proposed by Teso et al. [82], enable users to interact with the model by querying experts who can relabel the data based on the prediction and explanations, if necessary.

However, explaining machine learning models in a way that is understandable by musicians poses a significant challenge. Most explanations are too complex

for non-machine learning experts. For example, SLIME [83] uses explanations on temporal, spectral, and time-frequency, which is not of interest to most musicians or music enthusiasts. Although some explanations are simple enough, they usually employ the same mode of input and explanation. For instance, using a picture to explain a picture or text to explain text. If users lack domain knowledge, same-mode explanations are difficult for them to trust the model. Therefore, multimodal explanations can help people understand models in unfamiliar areas. When people learn, they use different modes of knowledge to recognize new objects. For example, when we describe music using words, it uses language as an explanation to help people learn about music. Different modes of knowledge can serve as each other’s explanation, providing mutual authentication.

In our proposed framework, we will leverage the ability to provide intuitive comments, comparisons, and explanations from Learning Under Privileged Information (LUPI) and incorporate basic musical notation on a typical five-line staff. The musical notation represents the composer’s explanation of music, and it can help establish a bridge of trust between the machine learning model and musicians or music enthusiasts.

The second challenge is the diversity and uncertainty inherent in the art of music[13]. Musical diversity arises from the fact that people from different races, creeds, and backgrounds can express their own understanding of the world through music[14]. For instance, different musical styles can be applied to the same melody, and the same melody can evoke different emotions when played in different keys, such as C major and A minor. The uncertainty in music is particularly prominent in absolute music. Consider two pieces of music, one with a positive emotion and the other with a negative emotion. The artist can alternate between sad and happy music, and vice versa. Due to the diversity and uncertainty in the art of music, it is challenging for both humans and machine learning algorithms.

To address this challenge, we propose to annotate the melody (the main tune

of a song) of the music based on the original data set. Expert annotations of the main melody can help the model better understand the intended emotion of the music. However, this approach presents a new problem as expert-labeled data is expensive and difficult to obtain for testing purposes.

The third challenge is the lack of data, specifically a shortage of data that has been annotated by experts. Emotion recognition data encompasses both music sound data and music background information, so it is crucial to combine different types of data. Useful complementary information related to the task is provided by different modalities, allowing us to train more robust predictive models. However, it is impossible to have different modal information at all times. For example, it can be difficult to find related sheet music or provide the melody for some music.

Our solution involves implementing a teacher-student model, as described in Privileged Features Distillation[84], where the teacher model is trained on multi-modal data, utilizing privileged information, and the student model subsequently learns to distill the feature descriptors corresponding to the missing modality.

In this paper, we introduce a novel machine learning framework for emotion recognition called Learning with Explainable Privileged Information via Multi-modal Distillation (EPMD). The framework aims to enhance the explainability of the model by leveraging privileged information, which is typically available during the training phase but not during the testing phase. Our proposed framework utilizes multimodal distillation to transfer the knowledge learned from the privileged information to the regular data, improving the accuracy and explainability of the model. The framework’s ability to incorporate privileged information as an explanation for the model’s decision-making process makes it a promising approach for applications where transparency and explainability are crucial.

In this paper, our contributions are:

- (1) To propose a multimodal explanation framework by using privileged information;
- (2) To propose a fine-grained method in audio classification for iden-

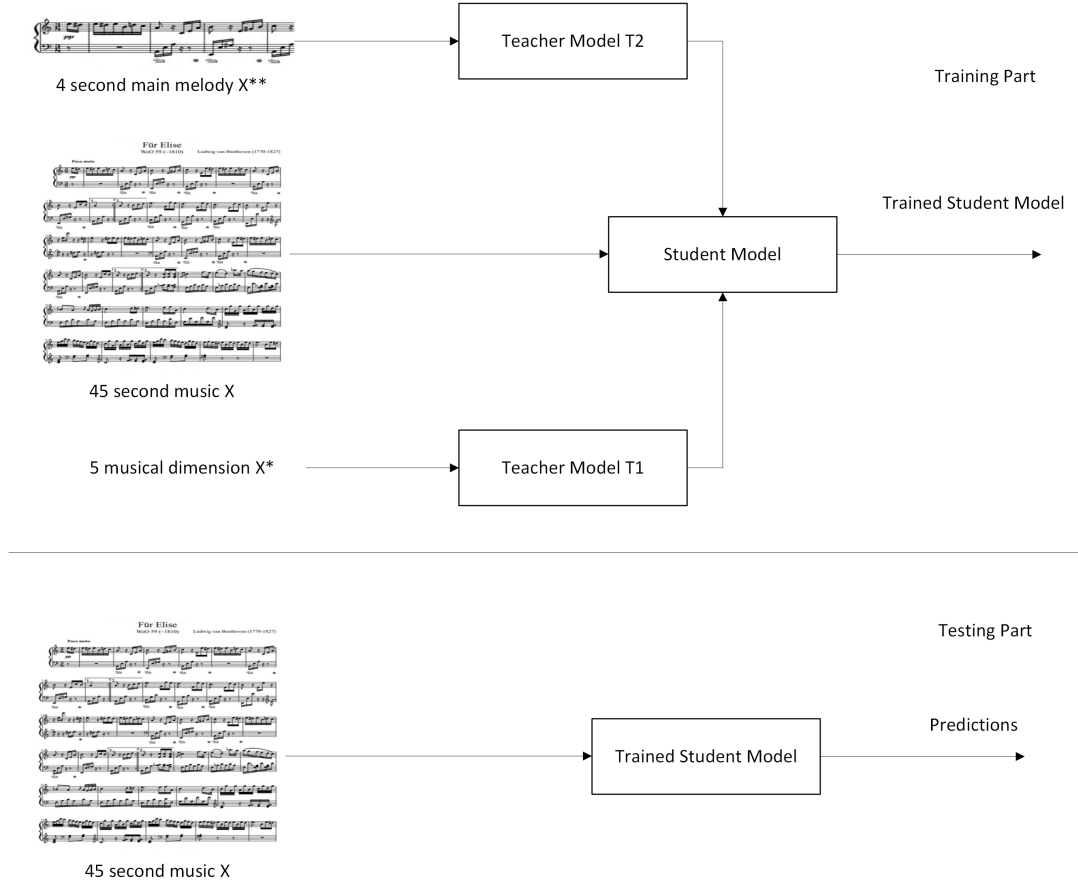


Figure 5.1: Model

tifying the main melody in music, which can help improve the learning efficiency of the model; (3) To demonstrate that by training the model with multimodal information, we can use the model with only music information during testing, which helps reduce the complexity of labeling for the model; (4) To conduct a questionnaire to comprehensively evaluate people's understanding of the model through their explanation. Furthermore, we compare the performance of different models and our model and explore how data marked with the main melody can assist the model.

5.2 Emotion Recognition and Musical Domain knowledge

5.2.1 Emotion Recognition

Emotion Recognition [85] with activity and valence refers to the process of identifying and categorizing emotional states of individuals based on their activities

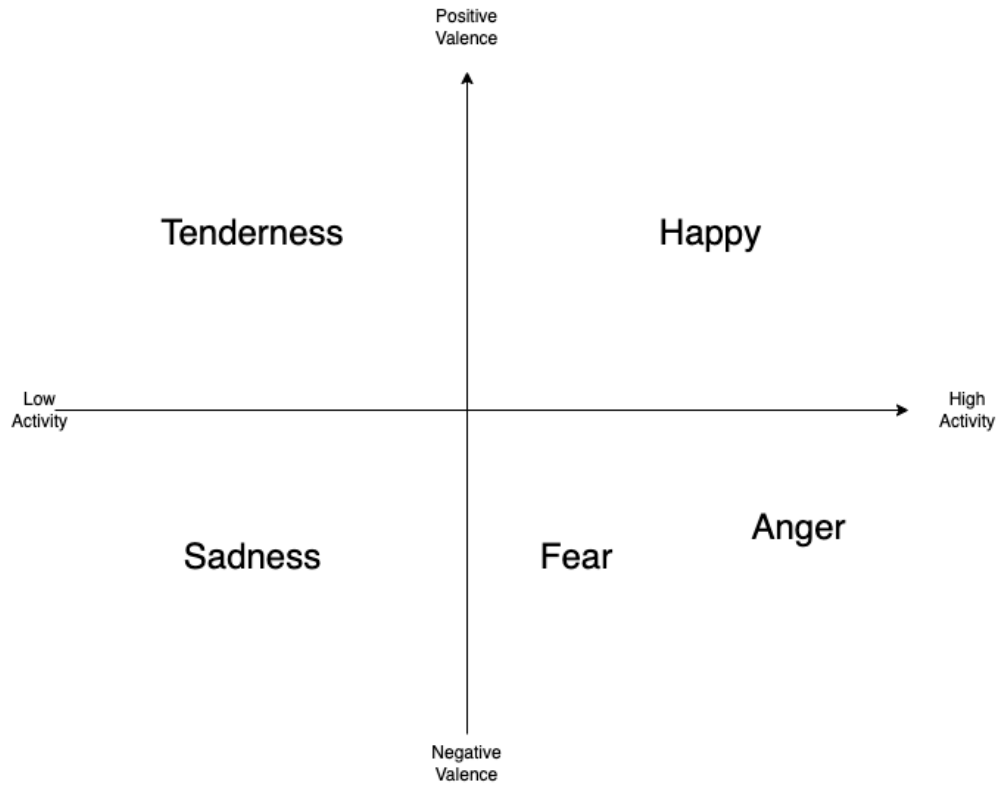


Figure 5.2: Emotional

and valence (the degree of positivity or negativity). This can be achieved through different modalities, including physiological signals (e.g. heart rate), behavioral observations, and self-reports. The resulting information can offer insights into an individual's emotional state and can be applied in mental health monitoring, human-computer interaction, and personalized communication systems.

The majority of research on music emotion recognition presently concentrates on constructing discriminative features and classifiers. However, this data-driven approach fails to effectively leverage the expertise in the emotional and musical domain knowledge, specifically the innate psychological connection between human emotion and music. This connection holds crucial information for accurately recognizing emotions in music.

5.2.2 Musical Domain knowledge:5 Musical Dimension

In this section, we introduce the dependencies between the 5 musical elements and emotions from the summarized music theory [86].

Music is a complex and multi-dimensional art form that involves various elements working together to create a unique and powerful emotional experience for the listener. Tempomode, brightness, loudness and pitch are five of the most important dimensions [87] of music shown in Fig 5.3.

Tempo

In music, tempo [88] refers to the pace at which a composition is played. It is measured in beats per minute (BPM) and can range from slow and relaxed to fast and energetic. Tempo can significantly influence the emotions conveyed by a piece of music and the moods it creates in the listener.

Different tempos are associated with different emotional qualities. For instance, slower tempos often convey more introspective or melancholic emotions, while faster tempos convey more energetic or joyful emotions. Moderate tempos, on the other hand, can convey a sense of balance or contentment. Musicians can often use tempo to influence the emotional perception in music by selecting specific tempo that best convey the intended emotional qualities of a piece. For instance, when aiming to evoke sadness or melancholy, they may opt for slower tempo. However, due to the inherent uncertainty in music art, there are occasions when musicians employ contrasting techniques, such as using fast-paced music to express sadness. A notable example is the beginning of Johann Sebastian Bach's St. John Passion, which features a rapid tempo and continuous sixteenth notes, yet it evokes a heavy and oppressive atmosphere rather than a pleasant one.

Mode

In music theory, a mode refers to a type of musical scale characterized by a specific arrangement of whole and half steps. Several modes exist, including major, natural minor, harmonic minor, and melodic minor. Each mode has its own tonal character and emotional quality [89].

In simple terms, to create a happy and positive mood, they might choose a

major key, while to create a sad or negative mood, they might choose a minor key. Of course, there are also a very small number of cases where exceptions may occur. One can still compose joyful songs in minor keys, as demonstrated by English composer Purcell's "Round O," which was composed in the key of D minor. Similarly, composers can create sorrowful compositions in major keys, such as Canadian poet and singer-songwriter Leonard Cohen's "Hallelujah."

Brightness

In music, brightness refers to the quality of sound that conveys a sense of higher frequency content, often described as a "sparkling" or "crisp" sound. This quality can be achieved by boosting or emphasizing the high-frequency range of a sound or musical instrument [90].

Brightness can significantly influence the emotions conveyed by a piece of music. Brighter sounds are often associated with positive emotions such as happiness and excitement, while darker or duller sounds may convey more negative emotions such as sadness or anxiety.

Loudness

Loudness in music refers to the perceived sound pressure level of a composition or a specific sound within it. In other words, loudness is how loud or soft a sound or piece of music is perceived by the listener.

Loudness can significantly influence the emotions conveyed by a piece of music. Research has shown that louder music is often associated with more energetic and intense emotions, while softer music is linked to more subdued and relaxing emotions [89]. However, Beethoven's Symphony No. 5 in C minor is a loud piece of music, but it is difficult to categorize it as happy or positive.

	High A	Low A	High V	Low V
Fast Tempo	√			
Slow Tempo		√		
Major Mode			√	
Minor Mode				√
High Brightness	√			
Low Brightness		√		
High Loudness	√			
Low Loudness		√		

Figure 5.3: Musical Dimensions

Pitch

Pitch refers to the perceived frequency of a musical sound or note. It determines how high or low a sound is perceived to be and is one of the fundamental aspects of music. Pitch is determined by the frequency of the sound wave, with higher frequencies producing higher pitches and lower frequencies producing lower pitches [91]. To put it more intuitively on piano, the higher the octave, the higher the pitch.

Pitch can significantly influence the emotions conveyed by a piece of music. Different pitches can convey different emotions, and their use in music can evoke different emotional responses in listeners. For instance, high pitches are often associated with excitement and happiness, while low pitches are linked to sadness and fear. However, there are also many special cases, such as St John Passion by Johann Sebastian Bach.

5.2.3 Main Meledy of Music

The main melody of music refers to the most prominent and easily recognizable melody line in a song or piece of music [92]. Usually, the main melody is composed of the melody played by the lead vocalist or the primary instrument in the song or piece, and it occupies a dominant position throughout the work, being the most easily remembered and perceived part. The characteristics of the main melody typically include a relatively simple melodic structure that is easy to attract the

listener’s attention, easy to remember and sing, and is also the core part of the song or piece that expresses the emotion and theme.

Based on my experience, the main melody of music is a four-bar musical phrase, so in this article, we have enlisted the help of experts to identify a 4-second excerpt of the main melody as privileged information X^{**} .

5.3 Preliminary Knowledge

The ensemble algorithm uses many models to make predictions together, so it has better performance, but using these base classifiers to predict each time will consume a lot of storage and prediction time. Knowledge distillation [27] refers to compressing a large model or many models into a small model while maintaining the performance of these models.

In classification tasks, the softmax operation returns a probability distribution over classes. Specifically, it normalizes the logit $z_{i,c}$ for each class in the network output to obtain the probability of the c class for the i samples:

$$q_i = \frac{\exp(z_{i,c}/T)}{\sum_j \exp(z_j/T)}$$

Here, C is the number of classes, and T is a hyperparameter called "temperature".

In general training, cross-entropy loss (CE) is used, which is defined as:

$$\ell_{ce} = - \sum_{x_i, y_i \sim p(x, y)} \sum_{c=1}^C \mathcal{I}\{y_i = c\} \log q_{i,c}$$

If we denote the true label distribution of a sample as p , and the predicted distribution as q , then:

$$KL(p||q) = \int p(y) \log \frac{p(y)}{q(y)} dy = \int p(y) \log p(y) dy - \int p(y) \log q(y) dy$$

It can be observed that cross-entropy loss actually minimizes the KL divergence between the true and predicted distributions.

The performance of the teacher model is reflected in the output of the final network, which can provide a probability distribution $\hat{p}_{i,c}$ over classes for any given sample. Using this distribution as a label, the following distillation loss can be defined:

$$\ell_{kd} = - \sum_{x_i, y_i \sim p(x, y)} \sum_{c=1}^C \hat{p}_{i,c} \log q_{i,c}$$

The learning strategy for a student model given a teacher model is:

$$\ell = \ell_{ce} + \lambda \ell_{kd}$$

5.4 Model

5.4.1 Problem Formulation

Our task is to create a Learning with explainable Privileged Information via Multimodal Distillation for Emotion Recognition (EPMD) framework that can train the model by incorporating related data and existing but standby domain knowledge. The overall goal of EPMD is to identify an explainable model by using limited data.

In our setting, we define the setting of EPMD as follows: First we let $D = (X, Y) \in \mathbb{R}$ be the samples from an unknown distribution $P(X, Y)$. We also evaluate a loss function L that compares a predicted result with ground truth label. Machine learning aims to estimate $P(X, Y)$ by D . In EPMD, we have supplementary information for each data point defined in space X^* and X^{**} , which is only accessible during the training phase. The X^* In other word, the samples from distribution $P(X, Y)$ is $x_i, x_i^* x_i^{**}, y_i \sim P(x_i, x_i^* x_i^{**}, y_i)$ during the training. However, in the test phase, we only have $x_i, y_i \sim P(x_i, y_i)$. Typically, after being given a function class $S(\cdot; \mathbf{w})$ parameterized by \mathbf{w} and data $\{x_i, y_i\}_{i \in [n]}$, our objective is to solve an optimization problem:

$$\min_{\mathbf{w}} E_{x, y \sim p(x, y)} [l(y, h(x; \mathbf{w}))] \quad (5.1)$$

When we supply additional multimodal information X , X^* and X^{**} in training and only use x in test. A parametric function class for privileged information $S^+ : \mathcal{X} \times \mathcal{X}^* \rightarrow \mathcal{Y}$ is needed. The training problem becomes:

$$\min_{\mathbf{w}} E_{x, x^*, x^{**}, y \sim p(x, x^*, x^{**}, y)} [l(y, S^+(x, x^*; \mathbf{w}))] \quad (5.2)$$

This corresponds to a classic supervised learning task with input samples defined in space x, x^*, x^{**} , which can be solved using an RNN model. To address the inference problem, we consider the following marginalization:

$$h(x; \mathbf{w}) \equiv E_{x^*, x^{**} \sim p(x^*, x^{**} | x)} [S^+(x, x^*, x^{**}; \mathbf{w})] \quad (5.3)$$

5.4.2 Learning with Explainable Privileged Information via Multimodal Distillation

The main issue with the formula above is that the expectation of $p(x^*, x^{**} | x)$ is difficult to calculate because it is unknown. The authors propose a way to restrict the set of functions that is easily computable in terms of expectations. Specifically, they introduce a family of parameters where the implicit information controls the variance and the main information x (data that is present both in training and testing) controls the mean. The specific form is as follows:

$$S^+(x, x^*, x^{**}; \mathbf{w}) = S^o(x; \mathbf{w}^o) \odot \mathcal{N}^*(\mathbf{1}, T_1(x^*; \mathbf{w}^*)) \odot \mathcal{N}^{**}(\mathbf{1}, T_2(x^{**}; \mathbf{w}^{**})) \quad (5.4)$$

\odot represents the Hadamard product of matrices (i.e., element-wise multiplication). The function $\mathcal{N}^*(\mathbf{1}, T_1(x^*; \mathbf{w}^*))$ and $\mathcal{N}^{**}(\mathbf{1}, T_2(x^{**}; \mathbf{w}^{**}))$ are normal random variable with mean determined by x^*, x^{**} . and w^*, w^{**} . Classify the output using the softmax function and we will introduce the training details in next part.

We use the Adam method to optimize and minimize the loss

$$h = \lambda_1 CE(So_1 GrandTruth) + \lambda_2 CE(So_1, So_2) + \lambda_3 CE(So_1, So_3). \quad (5.5)$$

Algorithm 1: Loss function process

```

1 Input:Dataset  $x, x^*, x^{**}$ , student model  $S$ , trained teacher model  $T_1, T_2$ 
2 Output: $\hat{S}$ 
3 Obtain Soft label 1( $So_1$ ), Soft label 2( $So_2$ ) and Soft label 3( $So_3$ ) from
  Model  $S, T_1$  and  $T_2$  ;
4 Learn loss function by minimising:
   $\lambda_1 CE(So_1, GrandTruth) + \lambda_2 CE(So_1, So_2) + \lambda_3 CE(So_1, So_3)$  ;
5 set  $t = 1$  ;
6 while  $E(S_{t-1}^*)$  is unaccepted do
7   Obtain Soft label 1( $So_1$ ), Soft label 2( $So_2$ ) and Soft label 3( $So_3$ ) from
    Model  $S, T_1$  and  $T_2$  ;
8   Learn  $S_t^*$  by minimising:
     $\lambda_1 CE(So_1, GrandTruth) + \lambda_2 CE(So_1, So_2) + \lambda_3 CE(So_1, So_3)$  ;
9 end
10  $\hat{S}^* = S_t^*$  ;
11
```

Figure 5.4: Learning Algorithm

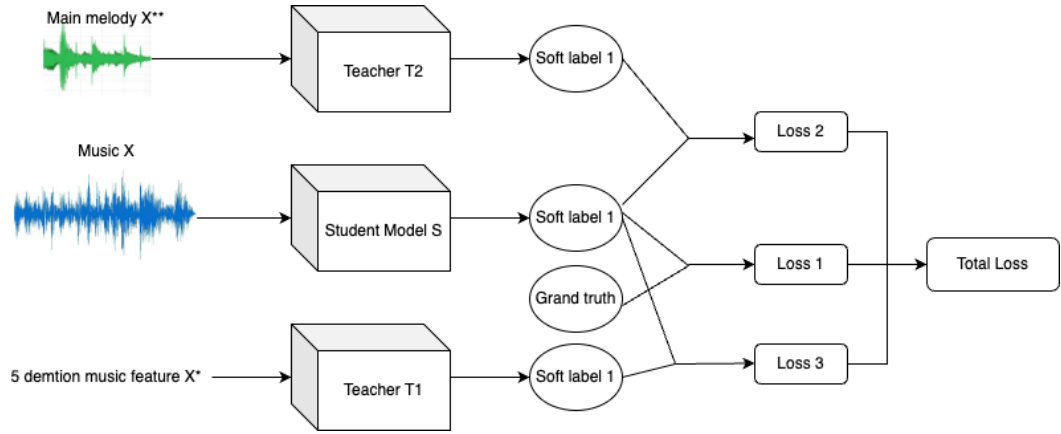


Figure 5.5: Loss function

The learning algorithm is shown in Fig5.4 .

5.5 Experiment

5.5.1 Experiment Setting

D1: "Music Emotion in 2015" is a data set focused on continuously assessing emotions in music. As the emotional tone of a piece of music can vary throughout, it poses a significant challenge to continuously evaluate its emotional character. The Valence-Arousal model will be utilized to analyze musical changes on two perpendicular axes. This is the third year the challenge is taking place, and

there is an abundant amount of data available. The emphasis this year is on annotation accuracy, with the best quality annotations from the previous two years (from 1744 existing songs) being chosen for the development set and an additional 250 annotated excerpts for the test set. The 45-second musical excerpts and annotations will be provided to participants, and all music is licensed under Creative Commons.

D2: "The AMG1608 dataset" is a high-quality emotion-annotated music emotion auto-recognition dataset for constructing accurate music emotion prediction models. This dataset publicly available to the research community, composed of 1608 30-sec music clips annotated by 665 subjects, with 46 subjects annotating over 150 songs, making it the largest of its kind to date.

$\widehat{D1}$ is a biased dataset compared to D1. We obtained the $\widehat{D1}$ dataset by modifying the labels of 50% of the samples in D1.

$\widehat{D2}$ is a biased dataset with D2. We obtained the $\widehat{D2}$ dataset by modifying the labels of 50% of D2.

Transformers(ACHFT) [93]: Wav2Vec 2.0 uses self-supervised contrastive learning to learn high-level speech representations instead of traditional low-level features. ACHFT trains the Wav2Vec 2.0 (base) model using the Hugging Face Transformers library and achieves state-of-the-art results on the Google Speech Commands Dataset.

Small convnet from scratch(Basic CNN) [94]: We convert music into spectrograms, which are images representing the frequency content of the audio over time. Then, we use a basic CNN for learning. This approach is based on the fact that some people classify sounds by converting them into images and using image classification methods for categorization.

Contrastive Learning of Musical Representations [95] (CLMR): CLMR is a self-supervised framework for raw waveform music representation learning using a large chain of audio data augmentations

ModelArts [96]: ModelArts is a cloud-based AI development platform pro-

vided by Huawei which provides pre-built AI models that can be used for audio classification.

S : Student model S is a basic RNN model which is used for learning target data (45-second music).

$S + T_1 + T_2$: Privileged information is employed in this study, where the student model S is combined with a simple backpropagation model T_1 using text information and a basic RNN model T_2 (using 4-second main melody to improve the accuracy of music classification).

$S + T_2$: Privileged information is employed in this study, where the student model S is combined with a basic RNN model T_2 (using 4-second main melody to improve the accuracy of music classification).

5.5.2 Data preprocessing

In this work we need to convert music into matrix information by using the Mel-scale Frequency Cepstral Coefficients method. Mel-scale Frequency Cepstral Coefficients (MFCC) [97] is the most commonly used speech feature in the field of Speech Recognition and Speaker Recognition. Extensive research on the auditory mechanism of the human ear has revealed that different frequencies of sound waves are perceived with varying sensitivities. Speech signals in the range of 200Hz to 5000Hz greatly affect speech clarity. When the human ear is exposed to two sounds with different loudness levels, the presence of higher frequency components in the louder sound can hinder the perception of lower frequency components, making them less noticeable. This phenomenon is known as the masking effect.

Due to the longer distance traveled by low-frequency sounds along the basilar membrane of the cochlea compared to high-frequency sounds, low-frequency sounds have a greater propensity to mask high-frequency sounds, while high-frequency sounds have a harder time masking low-frequency sounds. The critical bandwidth for sound masking is narrower at lower frequencies compared to higher

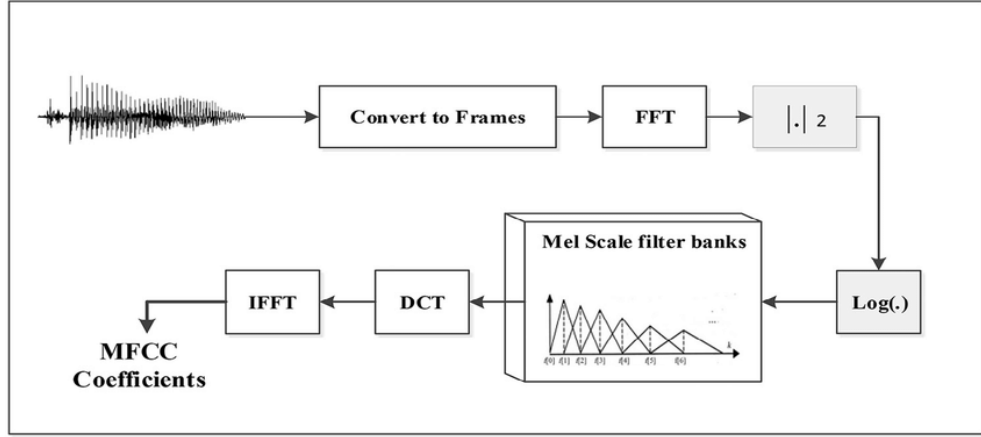


Figure 5.6: MFCC


frequencies. To address this, a series of bandpass filters is arranged in order of decreasing critical bandwidth within the frequency range from low to high. The input signal is filtered using these bandpass filters, and the energy of the signal output from each filter is utilized as a fundamental feature of the signal. Following further processing, this feature can serve as the input feature for speech recognition. Notably, this feature is independent of signal characteristics and does not impose any assumptions or restrictions on the input signal, making use of research insights from auditory models. As a result, it offers better robustness compared to LPCC based on vocal tract models and aligns more closely with the auditory characteristics of the human ear. Moreover, it maintains good recognition performance even in scenarios with low signal-to-noise ratios.

Mel-scale Frequency Cepstral Coefficients (MFCC) are cepstral parameters extracted within the Mel-scale frequency domain. The Mel scale captures the nonlinear properties of frequency perception in the human ear, and its relationship with frequency can be approximated by the following equation:

$$Mel(f) = 2595 * \log(1 + f/700) \quad (5.6)$$

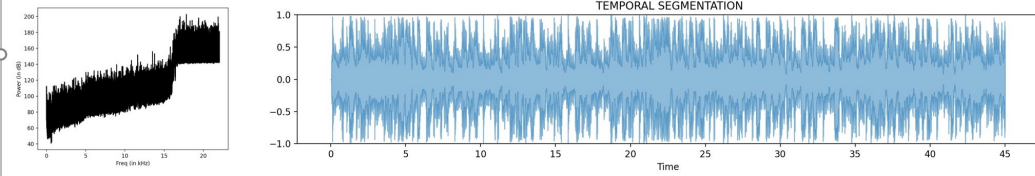
In the equation, f represents the frequency in Hz. The figure below illustrates the relationship between Mel frequency and linear frequency.

Multiple choice :

2.  If the emotion of this music is positive, which kind of explanation is easier to understand ? ()

A. The explanation from Slime

which are time frequency representation and temporal representation :



B. The explanation from music knowledge which are five elements of music : Tempo : $1.30 \cdot 10^0$; Brightness : $4.03 \cdot 10^{-1}$; Loudness : $1.41 \cdot 10^{-1}$; Pitch : $1.02 \cdot 10^{-4}$; Mode : $1.12 \cdot 10^{-1}$

C. The explanation from expert knowledge(main theme labeled by experts): 

Figure 5.7: Contents of the questionnaire

5.5.3 Experiment 1:The level of acceptance for different explanations

Experiment purpose: Most of the explanations are too complex for non-machine learning experts, especially when it comes to music. It is difficult to find specific explanations that are easy to understand. In order to gain the trust of human with no musical background or limited musical knowledge, we provided several explanations for the results of emotion recognition and allowed people to vote for the explanation they generally trust more through a questionnaire.

Experiment Setting: We start to set a questionnaire for potential machine learning model users. The questionnaire is focus on collecting data from people with different levels of musical knowledge. The contents of the questionnaire is multiple choice which let people choose more believable machine learning explanations. Firstly, we ask people to choose their level of mastery of music knowledge. Then we provide a forty-second piece of music and tell the user whether it is positive or negative. After that, the users The user will choose the following options from different explanations as shown in Fig 5.7: A:Explanations from slime are presented to the user as both time-domain and frequency-domain displays of the audio; B:Explanations from t1, using the five basic attributes of music to

Results of Questionnaire

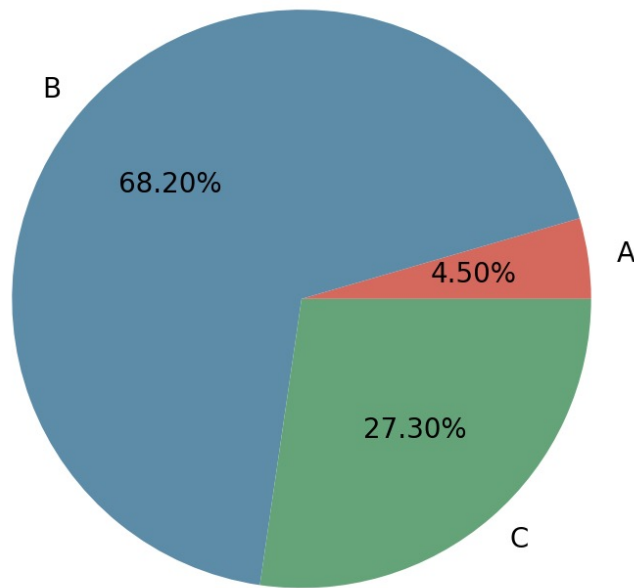


Figure 5.8: Result Of Questionnaire

explain; C:Explanations from t2, using the main melody to explain the entire piece of music.

Experiment Result: As results in Table 5.8 , we found that 68.2% of people believed that explaining music using the five basic attributes made them more confident, while less than a third thought that the main melody of the music could also serve as an explanation. The most professional option, option A, was only chosen by two people, both of whom were doctors, one of whom was a doctor who studies radar and relates sound to the time domain and frequency domain in his field of expertise. More than two-thirds of people chose to believe our explanation, which suggests that our explanation is more trustworthy in this specific field.

5.5.4 Experiment 2: Verify whether the melody labeling method is effective

Experiment purpose: To prove that labeling the main melody of music can improve the accuracy of model, we need to conduct a set of experiments for comparison.

Experiment Result: It is obvious in the table5.1 that the accuracy was 72.7% and 63.5% when only the S model was used. After adding the T_2 model that introduced the information of the main melody of the music, the accuracy of " ST_2 " was significantly improved to 78.1% and 78.3%. We used the 4-second melody instead of the 45-second music to learn from the model, and the obtained structure is better with the marked 4-second melody. This indicates that the information of the main melody of music can indeed help improve the accuracy of the model.

5.5.5 Experiment 3: Compare performance with other SOTA models

Experiment purpose: To further demonstrate the effectiveness of the proposed method, we compare it with the state of the art.

Experiment Result: We have observed in the table5.1 that our model has moderate accuracy among various models when the quality of the dataset is high and labels are correct. This is because our model, consisting of $S + T_1 + T_2$, is limited in terms of performance for this type of classification.

However, when the quality of the labels is low (as shown in $\widehat{D1}$ and $\widehat{D2}$), our accuracy of 68.2% and 67.5%, respectively, is among the best due to the incorporation of domain knowledge.

5.6 Conclusion

This chapter presents an explainable framework for multimodal information, making it easier for non-machine learning experts in machine learning to understand

	D1	D2	$\widehat{D1}$	$\widehat{D2}$
S	72.7	73.5		
ST_2	78.1	78.3		
ST_1T_2	79.6	80.4	68.2	67.5
ModelArts	82.8	81.5	58.9	52.7
ACHFT	79.3	80.7	55.6	52.5
CLMR	78.5	82.4	49.2	55.3
CNN	72.9	72.1	53.4	47.1
ModelArts(Main melody)	84.5	85.3	59.8	55.0
ACHFT(Main melody)	82.9	84.2	58.1	53.8
CLMR(Main melody)	80.6	81.7	53.9	52.6
CNN(Main melody)	81.5	79.4	55.3	51.7

Table 5.1: Results of Experiment 2 and Experiment 3

and trust the model. The framework is based on feature training, with text privileged information and privileged information on music expertise obtained from experts, to help train the student model. The model uses text information to explain music and has gained the trust of most non-machine experts. The framework was tested in the field of music emotion classification and achieved high performance.

Chapter 6

Conclusion and Future Work

In this chapter, we first conclude the entire thesis, and then show several interesting future directions.

6.1 Conclusion

With the growing popularity of machine learning, People’s dependence on artificial intelligence is increasing. Developing trustworthy and responsible AI and ensuring that AI models operate openly and transparently within a safe and controllable range has become a strong demand in various industries in society. The main focus of this thesis is on the coevolution of human and machine learning. To address this, theoretical and practical explorations have been conducted to tackle three main challenges:

1. How can domain knowledge be utilized to support a model.
2. How to use interaction between human and machine learning models to improve the performance of model output results.
3. What strategies can be used to understand and build trust in a model among non-machine learning experts.

Based on three questions, we proposed three solutions: a NHPP method for adding rule-based domain knowledge, a co-teaching accountable learning framework with distilled and domain knowledge and an explainable framework for multimodal information.

6.1.1 A BNP method for adding rule-based domain knowledge

In Chapter 3, we introduce a BNP method for customer segmentation that does not rely on knowledge of their utility functions for purchase behavior. Our semi-parametric approach unifies the parameters of different utility functions, allowing the parameters to be generated by the same distribution. This saves significant effort in designing specialized inference algorithms, allowing for efficient learning of latent variables. Moreover, this method is easily extensible to accommodate additional utility functions. Domain experts can also contribute to the model by incorporating their domain knowledge as utility function, leading to faster and more effective learning.

6.1.2 Co-teaching accountable learning framework with distilled and domain knowledge

In Chapter 4, we propose the Co-Teaching framework, which allows the learning process to leverage both distilled knowledge and domain knowledge. This framework is designed to be feasible and convenient for experts to understand the learning progress and provide critical and correct guidance. Unlike other learning approaches that rely on modeling or data, this framework emphasizes the role of provided knowledge in guiding the learning process. We also compare this framework with similar approaches and explore how machine learning models can learn from experts. This expectation is that this framework can serve as a fundamental design for machine learning applications, fostering collaboration between machine learning model and domain experts.

6.1.3 An explainable framework for multimodal information

In chapter 5, we proposes a novel multimodal explanation framework that leverages privileged information to enhance the learning process. Additionally, we

propose a fine-grained method for audio classification that focuses on identifying the main melody in music, thereby improving the model’s learning efficiency. By training the model with multimodal information, we demonstrate that the model can be used with only music information during testing, reducing the complexity of labeling for the model. We further evaluate the model’s effectiveness through a questionnaire that comprehensively evaluates people’s understanding of the model through their explanation. Our work also includes comparing the performance of different models, highlighting the benefits of our proposed framework. Finally, we explore how data marked with the main melody can assist the model in improving its accuracy and generalization capabilities. Overall, our work contributes to advancing the field of multimodal learning and audio classification, and we anticipate that our proposed framework will find practical applications in real-world settings.

6.2 Future Work

Due to limitations in time and technology, the work presented in the thesis has some restrictions: ① Only rule-based knowledge was utilized in the first work, but there are still many other forms of domain knowledge that need to be used; ② In the second work, the input data was selected by experts, rather than using a variety of styles of real-world knowledge; ③ A more general way of providing feedback from the model (model explanation) is needed in both the second and third works to make machine learning models more accessible to wider user base.

The driving force behind the research on coevolution for human and machine learning is twofold. On the one hand, it stems from the need to explore the limits of human cognition by investigating the intrinsic decision-making mechanisms of machine learning. On the other hand, it is driven by the demands of stakeholders in the field of artificial intelligence applications, as well as regulatory authorities. To satisfy these two major aspects, we can work on these research in the future:

1.Theoretical framework:The explainability of artificial intelligence in-

volves both the basic framework of cognitive theory and various complex technical methods. However, there is currently no complete explainable AI theory that can integrate all relevant content. The lack of a comprehensive theoretical guide also means that research on explainable AI theory is still in the experimental stage guided by trial and error. The first problem that needs to be addressed in explainable AI research is how to summarize and extract a theoretical system that is both in line with practical cognition and logically consistent from empirical experimental results.

2.Transfer domain knowledge: The models often lack a massive amount of data to learn from in many application domains. In such cases, we need to make the most of expert knowledge to help the model learn. Expert knowledge comes in different modalities, formats, and forms, so we need a unified approach to incorporate knowledge from different experts.

3.Machine learning in music: In my research, I found that the study of machine learning in the field of sound is significantly less than in images, especially in the field of music. This may be because humans are more sensitive to visual information, leading to a wider range of applications in computer vision, and also because visual data is easier to obtain and process. However, as more audio data and labels become available, and with the continuous development and improvement of machine learning algorithms, I believe that there will be more research and applications involving audio and music in the future. We can fill these gaps by further studying the relationship between machine learning and music.

4.Evaluation and certification of explainable machine learning: Model evaluation and certification are an integral part of the field of machine learning, just as important as model structure and parameters. They serve as important criteria for technology selection in critical industries such as banking, healthcare, and autonomous driving, and provide objectivity and standardization. Evaluation methods must be matched to the application scenario in order to better

address real-world problems. Currently, there is little research on the evaluation and certification of music classification using machine learning, and due to time constraints, this paper was unable to carry out related studies. From an application perspective, the explanation of artificial intelligence models needs to provide reliable and trustworthy usage basis for stakeholders in various industries and risk-controllable management mechanisms. This requires evaluation methods and corresponding technical indicators for interpreting results to test and determine whether different methods and forms of explanation can meet the needs of stakeholders in all aspects. At the same time, for risk control, it is necessary to translate systemic regulatory rules into executable technical solutions and obtain the endorsement and recognition of relevant experts and regulatory authorities. How to conduct reliable evaluation and certification of explainable AI to meet the above requirements is a technical challenge that must be addressed to promote the large number of application of artificial intelligence.

5.Feedback and guidance:When human discover potential risks in the model through the explanation results, these feedback information must be presented in an appropriate form to improve model performance and reduce potential risks. When the human is dissatisfied with the explanation results, the explanation should also be corrected to improve the model’s credibility. Technically, how to transform various human feedback into input information for machine learning and effectively integrate it into the model training and explanation generation process is an important issue that needs to be studied.

CERTIFICATE OF ORIGINAL AUTHORSHIP

I, Siqi Zhang, declare that this thesis is submitted in fulfilment of the requirements for the award of Doctor of Philosophy, in the Faculty of Engineering and information Technology at the University of Technology Sydney.

This thesis is wholly my own work unless otherwise referenced or acknowledged. In addition, I certify that all information sources and literature used are indicated in the thesis.

This document has not been submitted for qualifications at any other academic institution. *If applicable, the above statement must be replaced with the collaborative doctoral degree statement (see below).

*If applicable, the Indigenous Cultural and Intellectual Property (ICIP) statement must be added (see below).

This research is supported by the Australian Government Research Training Program.

Signature: Siqi Zhang

Date: 1/5/2023

REFERENCES

- [1] J. Pearl and D. Mackenzie, *The book of why: the new science of cause and effect*. Basic books, 2018.
- [2] J. Pearl *et al.*, “Models, reasoning and inference,” *Cambridge, UK: Cambridge University Press*, vol. 19, no. 2, 2000.
- [3] J. Angwin, J. Larson, S. Mattu, and L. Kirchner, “Machine bias,” in *Ethics of data and analytics*. Auerbach Publications, 2016, pp. 254–264.
- [4] N. A. Smuha, “The eu approach to ethics guidelines for trustworthy artificial intelligence,” *Computer Law Review International*, vol. 20, no. 4, pp. 97–106, 2019.
- [5] P. Hall, N. Gill, and B. Cox, *Responsible Machine Learning*. O’Reilly Media, Incorporated, 2020.
- [6] Y. N. Harari, *Homo Deus: A brief history of tomorrow*. random house, 2016.
- [7] J. N. Thompson, “Concepts of coevolution,” *Trends in Ecology & Evolution*, vol. 4, no. 6, pp. 179–183, 1989.
- [8] T. Mitchell, “Machine learning. macgraw-hill companies,” *Inc., Boston*, 1997.
- [9] T. G. Dietterich, “Steps toward robust artificial intelligence,” *Ai Magazine*, vol. 38, no. 3, pp. 3–24, 2017.
- [10] D. L. Silver, Q. Yang, and L. Li, “Lifelong machine learning systems: Beyond learning algorithms,” in *2013 AAAI spring symposium series*, 2013.
- [11] Z.-H. Zhou, “Learnware: on the future of machine learning.” *Frontiers Comput. Sci.*, vol. 10, no. 4, pp. 589–590, 2016.
- [12] R. L. Solso, M. K. MacLin, and O. H. MacLin, *Cognitive psychology*. Pearson Education New Zealand, 2005.
- [13] B. B. Smith, “Variability, change, and the learning of music,” *Ethnomusicology*, vol. 31, no. 2, pp. 201–220, 1987.
- [14] P. S. Campbell, *Music, education, and diversity: Bridging cultures and communities*. Teachers College Press, 2017.

- [15] S. J. Gershman and D. M. Blei, “A tutorial on bayesian nonparametric models,” *Journal of Mathematical Psychology*, vol. 56, no. 1, pp. 1–12, 2012.
- [16] W. Guo, S. Huang, Y. Tao, X. Xing, and L. Lin, “Explaining deep learning models—a bayesian non-parametric approach,” *Advances in neural information processing systems*, vol. 31, 2018.
- [17] K. Siegrist, “Probability, mathematical statistics, stochastic processes,” 2017.
- [18] H. Akaike, “Information theory and an extension of the maximum likelihood principle,” in *Selected papers of hirotugu akaike*. Springer, 1998, pp. 199–213.
- [19] L. Luo, B. Li, I. Koprinska, S. Berkovsky, and F. Chen, “Discovering temporal purchase patterns with different responses to promotions,” in *Proceedings of the 25th ACM international on conference on information and knowledge management*. ACM, 2016, pp. 2197–2202.
- [20] M. B. Sirvanci, “An empirical study of price thresholds and price sensitivity,” *Journal of Applied Business Research (JABR)*, vol. 9, no. 2, pp. 43–49, 1993.
- [21] L. Masiero and J. L. Nicolau, “Tourism market segmentation based on price sensitivity: Finding similar price preferences on tourism activities,” *Journal of Travel Research*, vol. 51, no. 4, pp. 426–435, 2012.
- [22] M. McDonald, M. Christopher, and M. Bass, “Market segmentation,” in *Marketing*. Springer, 2003, pp. 41–65.
- [23] P. Smyth, “Clustering sequences with hidden markov models,” in *Advances in neural information processing systems*, 1997, pp. 648–654.
- [24] A. G. Hawkes, “Spectra of some self-exciting and mutually exciting point processes,” *Biometrika*, vol. 58, no. 1, pp. 83–90, 1971.
- [25] P. J. Laub, T. Taimre, and P. K. Pollett, “Hawkes processes,” *arXiv preprint arXiv:1507.02822*, 2015.
- [26] K. Weiss, T. M. Khoshgoftaar, and D. Wang, “A survey of transfer learning,” *Journal of Big data*, vol. 3, no. 1, pp. 1–40, 2016.
- [27] G. Hinton, O. Vinyals, and J. Dean, “Distilling the knowledge in a neural network,” *arXiv preprint arXiv:1503.02531*, 2015.
- [28] V. Vapnik and A. Vashist, “A new learning paradigm: Learning using privileged information,” *Neural networks*, vol. 22, no. 5-6, pp. 544–557, 2009.
- [29] D. Lopez-Paz, L. Bottou, B. Schölkopf, and V. Vapnik, “Unifying distillation and privileged information,” *arXiv preprint arXiv:1511.03643*, 2015.

- [30] Y. Wang, Q. Yao, J. T. Kwok, and L. M. Ni, “Generalizing from a few examples: A survey on few-shot learning,” *ACM computing surveys (csur)*, vol. 53, no. 3, pp. 1–34, 2020.
- [31] J. Snell, K. Swersky, and R. S. Zemel, “Prototypical networks for few-shot learning,” *arXiv preprint arXiv:1703.05175*, 2017.
- [32] M. Gillies, R. Fiebrink, A. Tanaka, J. Garcia, F. Bevilacqua, A. Heloir, F. Nunnari, W. Mackay, S. Amershi, B. Lee *et al.*, “Human-centred machine learning,” in *Proceedings of the 2016 CHI conference extended abstracts on human factors in computing systems*, 2016, pp. 3558–3565.
- [33] A. Blum and T. Mitchell, “Combining labeled and unlabeled data with co-training,” in *Proceedings of the eleventh annual conference on Computational learning theory*, 1998, pp. 92–100.
- [34] B. Han, Q. Yao, X. Yu, G. Niu, M. Xu, W. Hu, I. Tsang, and M. Sugiyama, “Co-teaching: Robust training of deep neural networks with extremely noisy labels,” *arXiv preprint arXiv:1804.06872*, 2018.
- [35] B. Settles, “Active learning literature survey,” 2009.
- [36] Y. Gal, R. Islam, and Z. Ghahramani, “Deep bayesian active learning with image data,” in *International Conference on Machine Learning*. PMLR, 2017, pp. 1183–1192.
- [37] N. Kosmyna, F. Tarpin-Bernard, and B. Rivet, “Adding human learning in brain–computer interfaces (bcis) towards a practical control modality,” *ACM Transactions on Computer-Human Interaction (TOCHI)*, vol. 22, no. 3, pp. 1–37, 2015.
- [38] B. Hammer and M. Toussaint, “Special issue on autonomous learning,” 2015.
- [39] Y. Zhang, F. Zhou, Z. Li, Y. Wang, and F. Chen, “Bias-tolerant fair classification,” *arXiv preprint arXiv:2107.03207*, 2021.
- [40] D. Gunning, M. Stefik, J. Choi, T. Miller, S. Stumpf, and G.-Z. Yang, “Xai—explainable artificial intelligence,” *Science robotics*, vol. 4, no. 37, p. eaay7120, 2019.
- [41] A. Goldstein, A. Kapelner, J. Bleich, and E. Pitkin, “Peeking inside the black box: Visualizing statistical learning with plots of individual conditional expectation,” *journal of Computational and Graphical Statistics*, vol. 24, no. 1, pp. 44–65, 2015.
- [42] D. W. Apley and J. Zhu, “Visualizing the effects of predictor variables in black box supervised learning models,” *arXiv preprint arXiv:1612.08468*, 2016.

- [43] K. Cheng, Z. Lu, C. Ling, and S. Zhou, “Surrogate-assisted global sensitivity analysis: an overview,” *Structural and Multidisciplinary Optimization*, vol. 61, pp. 1187–1213, 2020.
- [44] M. T. Ribeiro, S. Singh, and C. Guestrin, “” why should i trust you?” explaining the predictions of any classifier,” in *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*, 2016, pp. 1135–1144.
- [45] A. Aamodt and E. Plaza, “Case-based reasoning: Foundational issues, methodological variations, and system approaches,” *AI communications*, vol. 7, no. 1, pp. 39–59, 1994.
- [46] B. Kim, R. Khanna, and O. O. Koyejo, “Examples are not enough, learn to criticize! criticism for interpretability,” *Advances in neural information processing systems*, vol. 29, 2016.
- [47] S. Wachter, B. Mittelstadt, and C. Russell, “Counterfactual explanations without opening the black box: Automated decisions and the gdpr,” *Harv. JL & Tech.*, vol. 31, p. 841, 2017.
- [48] P. W. W. Koh, K.-S. Ang, H. Teo, and P. S. Liang, “On the accuracy of influence functions for measuring group effects,” *Advances in neural information processing systems*, vol. 32, 2019.
- [49] P. W. Koh and P. Liang, “Understanding black-box predictions via influence functions,” in *International conference on machine learning*. PMLR, 2017, pp. 1885–1894.
- [50] K. Simonyan, A. Vedaldi, and A. Zisserman, “Deep inside convolutional networks: Visualising image classification models and saliency maps,” *arXiv preprint arXiv:1312.6034*, 2013.
- [51] G. Montavon, W. Samek, and K.-R. Müller, “Methods for interpreting and understanding deep neural networks,” *Digital signal processing*, vol. 73, pp. 1–15, 2018.
- [52] H. Yuan, J. Tang, X. Hu, and S. Ji, “Xgnn: Towards model-level explanations of graph neural networks,” in *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, 2020, pp. 430–438.
- [53] K. R. Varshney, “Trustworthy machine learning and artificial intelligence,” *XRDS: Crossroads, The ACM Magazine for Students*, vol. 25, no. 3, pp. 26–29, 2019.

- [54] E. Toreini, M. Aitken, K. Coopamootoo, K. Elliott, C. G. Zelaya, and A. Van Moorsel, “The relationship between trust in ai and trustworthy machine learning technologies,” in *Proceedings of the 2020 conference on fairness, accountability, and transparency*, 2020, pp. 272–283.
- [55] O. R. Wearn, R. Freeman, and D. M. Jacoby, “Responsible ai for conservation,” *Nature Machine Intelligence*, vol. 1, no. 2, pp. 72–73, 2019.
- [56] P. M. Winter, S. Eder, J. Weissenböck, C. Schwald, T. Doms, T. Vogt, S. Hochreiter, and B. Nessler, “Trusted artificial intelligence: Towards certification of machine learning applications,” *arXiv preprint arXiv:2103.16910*, 2021.
- [57] C. Zhang, Y. Xie, H. Bai, B. Yu, W. Li, and Y. Gao, “A survey on federated learning,” *Knowledge-Based Systems*, vol. 216, p. 106775, 2021.
- [58] G. A. Kaissis, M. R. Makowski, D. Rückert, and R. F. Braren, “Secure, privacy-preserving and federated machine learning in medical imaging,” *Nature Machine Intelligence*, vol. 2, no. 6, pp. 305–311, 2020.
- [59] X. Wu, L. Xiao, Y. Sun, J. Zhang, T. Ma, and L. He, “A survey of human-in-the-loop for machine learning,” *Future Generation Computer Systems*, 2022.
- [60] D. Xin, L. Ma, J. Liu, S. Macke, S. Song, and A. Parameswaran, “Accelerating human-in-the-loop machine learning: Challenges and opportunities,” in *Proceedings of the second workshop on data management for end-to-end machine learning*, 2018, pp. 1–4.
- [61] D. Honeycutt, M. Nourani, and E. Ragan, “Soliciting human-in-the-loop user feedback for interactive machine learning reduces user trust and impressions of model accuracy,” in *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing*, vol. 8, no. 1, 2020, pp. 63–72.
- [62] K. Arulkumaran, M. P. Deisenroth, M. Brundage, and A. A. Bharath, “Deep reinforcement learning: A brief survey,” *IEEE Signal Processing Magazine*, vol. 34, no. 6, pp. 26–38, 2017.
- [63] W. A. Kamakura and G. J. Russell, “A probabilistic choice model for market segmentation and elasticity structure,” *Journal of marketing research*, vol. 26, no. 4, pp. 379–390, 1989.
- [64] D. Yankelovich and D. Meer, “Rediscovering market segmentation,” *Harvard business review*, vol. 84, no. 2, p. 122, 2006.
- [65] T. L. Griffiths, M. I. Jordan, J. B. Tenenbaum, and D. M. Blei, “Hierarchical topic models and the nested Chinese restaurant process,” in *Advances in neural information processing systems*, 2004, pp. 17–24.

- [66] H. R. Varian, “Price discrimination and social welfare,” *The American Economic Review*, vol. 75, no. 4, pp. 870–875, 1985.
- [67] W. F. Samuelson and S. G. Marks, *Managerial economics*. John Wiley & Sons, 2008.
- [68] J. M. Kamen and R. J. Toman, “Psychophysics of prices,” *Journal of Marketing Research*, vol. 7, no. 1, pp. 27–35, 1970.
- [69] I. Porteous, D. Newman, A. Ihler, A. Asuncion, P. Smyth, and M. Welling, “Fast collapsed gibbs sampling for latent dirichlet allocation,” in *Proceedings of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM, 2008, pp. 569–577.
- [70] A. Rodriguez and A. Laio, “Clustering by fast search and find of density peaks,” *Science*, vol. 344, no. 6191, pp. 1492–1496, 2014.
- [71] I. S. Dhillon, Y. Guan, and B. Kulis, “Kernel k-means: spectral clustering and normalized cuts,” in *Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM, 2004, pp. 551–556.
- [72] D. P. Kingma and J. Ba, “Adam: A method for stochastic optimization,” *arXiv preprint arXiv:1412.6980*, 2014.
- [73] M. P. R. D. Andras Janosi, William Steinbrunn, “Heart disease uci,” Website, 1988, <https://archive.ics.uci.edu/ml/datasets/Heart+Disease>.
- [74] M. Akturk, “Diabetes dataset,” Website, 1999, <https://www.kaggle.com/mathchi/diabetes-data-set>.
- [75] Fedesoriano, “Stroke prediction dataset,” Website, 1999, <https://www.kaggle.com/datasets/fedesoriano/stroke-prediction-dataset>.
- [76] A. Aljanaki, Y.-H. Yang, and M. Soleymani, “Emotion in music task at mediaeval 2015,” in *In Working Notes Proceedings of the MediaEval 2015 Workshop*, 2015.
- [77] O. Lartillot and P. Toivainen, “A matlab toolbox for musical feature extraction from audio,” in *DAFx*, vol. 237. Bordeaux, 2007, p. 244.
- [78] S. Sinha, S. Ebrahimi, and T. Darrell, “Variational adversarial active learning,” in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2019, pp. 5972–5981.
- [79] A. Zhao, M. Ding, Z. Lu, T. Xiang, Y. Niu, J. Guan, and J.-R. Wen, “Domain-adaptive few-shot learning,” in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 2021, pp. 1390–1399.

- [80] J. Li, R. Socher, and S. C. Hoi, “Dividemix: Learning with noisy labels as semi-supervised learning,” *arXiv preprint arXiv:2002.07394*, 2020.
- [81] J. Zhou, A. H. Gandomi, F. Chen, and A. Holzinger, “Evaluating the quality of machine learning explanations: A survey on methods and metrics,” *Electronics*, vol. 10, no. 5, p. 593, 2021.
- [82] S. Teso and K. Kersting, “Explanatory interactive machine learning,” in *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society*, 2019, pp. 239–245.
- [83] S. Mishra, B. L. Sturm, and S. Dixon, “Local interpretable model-agnostic explanations for music content analysis.” in *ISMIR*, vol. 53, 2017, pp. 537–543.
- [84] C. Xu, Q. Li, J. Ge, J. Gao, X. Yang, C. Pei, F. Sun, J. Wu, H. Sun, and W. Ou, “Privileged features distillation at taobao recommendations,” in *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, 2020, pp. 2590–2598.
- [85] P. N. Juslin and J. Sloboda, *Handbook of music and emotion: Theory, research, applications*. Oxford University Press, 2011.
- [86] T. Eerola, “Finnish centre of excellence in interdisciplinary music research, finland.” *Psychomusicology: Music, Mind, and Brain*, vol. 22, no. 2, p. 180, 2012.
- [87] C. Schmidt-Jones, “Understanding basic music theory,” 2013.
- [88] A. Fernández-Sotos, A. Fernández-Caballero, and J. M. Latorre, “Influence of tempo and rhythmic unit in musical emotion regulation,” *Frontiers in computational neuroscience*, vol. 10, p. 80, 2016.
- [89] P. N. Juslin, “From everyday emotions to aesthetic emotions: Towards a unified theory of musical emotions,” *Physics of life reviews*, vol. 10, no. 3, pp. 235–266, 2013.
- [90] E. Brattico, V. Alluri, B. Bogert, T. Jacobsen, N. Vartiainen, S. Nieminen, and M. Tervaniemi, “A functional mri study of happy and sad emotions in music with and without lyrics,” *Frontiers in psychology*, vol. 2, p. 308, 2011.
- [91] P. N. Juslin and P. Laukka, “Communication of emotions in vocal expression and music performance: Different channels, same code?” *Psychological bulletin*, vol. 129, no. 5, p. 770, 2003.
- [92] W.-H. Tsai, H.-M. Yu, H.-M. Wang, and J.-T. Horng, “Using the similarity of main melodies to identify cover versions of popular songs for music document retrieval.” *J. Inf. Sci. Eng.*, vol. 24, no. 6, pp. 1669–1687, 2008.

- [93] S. Ghosh, S. Lepcha, S. Sakshi, and R. R. Shah, “Speech toxicity analysis: A new spoken language processing task,” *arXiv preprint arXiv:2110.07592*, 2021.
- [94] F. Chollet, *Deep learning with Python*. Simon and Schuster, 2021.
- [95] J. Spijkervet and J. A. Burgoyne, “Contrastive learning of musical representations,” *arXiv preprint arXiv:2103.09410*, 2021.
- [96] W. Chen, J. Tong, R. He, Y. Lin, P. Chen, Z. Chen, and X. Liu, “An easy method for identifying 315 categories of commonly-used chinese herbal medicines based on automated image recognition using automl platforms,” *Informatics in Medicine Unlocked*, vol. 25, p. 100607, 2021.
- [97] V. Tiwari, “Mfcc and its applications in speaker recognition,” *International journal on emerging technologies*, vol. 1, no. 1, pp. 19–22, 2010.

APPENDICES

A Appendix A Title

Detailed experimental procedures, data tables, computer programs, etc. may be placed in appendices. This may be particularly appropriate if the dissertation or thesis includes several published papers.

B Appendix B Title

C Papers Submitted and Under Preparation

- Siqi Zhang, Zhidong Li, Lin Luo and etc. Simultaneous customer segmentation and behavior discovery, *Submitted to International Conference on Neural Information Processing*, Published.
- Siqi Zhang, Zhidong Li, Feng Zhou and etc. The co-teaching with dual-knowledge: accountable learning framework with distilled and domain knowledge, *Submitted to The Conference on Information and Knowledge Management*, Submitted.
- Siqi Zhang, Zhidong Li, Feng Zhou and etc. An explainable framework for multimodal privileged information for Emotion recognition, *Submitted to The Conference on Information and Knowledge Management*, Submitted.