Predictive Learning from Real-World Medical Data: Overcoming Quality Challenges

ZEYUAN WANG

Supervisor: Dr Simon Poon
Associate Supervisor: Dr Josiah Poon

A thesis submitted in fulfilment of
the requirements for the degree of
Doctor of Philosophy

School of Computer Science
Faculty of Engineering
The University of Sydney
Australia

19 December 2023

Abstract

Randomized controlled trials (RCTs) are pivotal in medical research, notably as the gold standard, but face challenges, especially with specific groups like pregnant women and newborns. Real-world data (RWD), from sources like electronic medical records and insurance claims, complements RCTs in areas like disease risk prediction and diagnosis. However, RWD's retrospective nature leads to issues such as missing values and data imbalance, requiring intensive data preprocessing. To enhance RWD's quality for predictive modeling, this thesis introduces a suite of algorithms developed to automatically resolve RWD's low-quality issues for predictive modeling.

In this study, the AMI-Net method is first introduced, innovatively treating samples as bags with various feature-value pairs and unifying them in an embedding space using a multi-instance neural network. It excels in handling incomplete datasets, a frequent issue in real-world scenarios, and shows resilience to noise and class imbalances. AMI-Net's capability to discern informative instances minimizes the effects of low-quality data. The enhanced version, AMI-Net+, improves instance selection, boosting performance and generalization. However, AMI-Net series initially only processes binary input features, a constraint overcome by AMI-Net3, which supports binary, nominal, ordinal, and continuous features. Despite advancements, challenges like missing values, data inconsistencies, and labeling errors persist in real-world data. The AMI-Net series also shows promise for regression and multi-task learning, potentially mitigating low-quality data issues. Tested on various hospital datasets, these methods prove effective, though risks of overfitting and bias remain, necessitating further research. Overall, while promising for clinical studies and other applications, ensuring data quality and reliability is crucial for these methods' success.

Author Attribution

The work contained in the body of this thesis, except otherwise acknowledged, is the result of my own investigations.

Chapter 3 is published in International joint conference on neural networks (IJCNN). Zeyuan Wang (ZW) designed the study and conducted the experiment with help from Josiah Poon (JP) and Simon Poon (SP). Shiding Sun (SS) assisted with the collection of samples. The manuscript was written by ZW but all co-authors contributed intellectually to the drafts.

Chapter 4 is published in Australian Journal of Intelligent Information Processing Systems. ZW designed the study and conducted the experiment. The co-authors of this study include JP and SP. The manuscript was written by ZW. All co-authors provided important feedback to manuscript drafts.

Chapter 5 is published in Artificial Intelligence in Medicine. The co-authors of this study are ZW, JP, Shuze Wang (SW), SS and SP. ZW designed the study and conducted the experiment with the help of SW and SS. All co-authors have reviewed drafts for the manuscript in preparation for submission.

Chapter 6 is published in Hawaii International Conference on System Sciences (HICSS). ZW designed the study and conducted the experiment with help from JP and SP. The manuscript was written by ZW but all co-authors contributed intellectually to the drafts.

In addition to the statements above, in cases where I am not the corresponding author of a published item, permission to include the published material has been granted by the corresponding author.

Zeyuan Wang, 16 July 2023

As supervisor for the candidature upon which this thesis is based, I can confirm that the authorship attribution statements above are correct.

Josiah Poon, 16 July 2023

Thesis Statement of Originality

I, Zeyuan Wang, hereby declare that this graduation thesis, titled "Predictive Learning from Real-World Medical Data: Overcoming Quality Challenges," is my original work, conducted under the supervision of Dr Simon Poon and Dr Josiah Poon, as a requirement for the completion of my PhD degree in Computer Science.

I further declare that this thesis has not been submitted, in part or in whole, for the award of any other degree or diploma at any university or educational institution. The data and results presented in this thesis are authentic and have not been falsified or manipulated.

I acknowledge that the responsibility for the content and interpretation of this thesis rests solely with me. I understand that any unauthorized use or plagiarism of the work presented in this thesis may have serious academic and legal consequences.

Zeyuan Wang

22 May, 2023

Acknowledgements

I wish to express my deepest gratitude to my esteemed supervisors, Dr. Simon Poon and Dr. Josiah Poon, whose guidance has been pivotal in shaping me from a technologist with limited knowledge to a researcher with a strong foundation in critical thinking and problem-solving. Their unwavering commitment to my growth has enabled me to broaden my horizons, explore innovative research avenues, and make significant contributions to the field. Their keen insight, patience, and encouragement have inspired me to surpass my own limitations and pursue excellence in my work.

Under the mentorship of Dr. Simon Poon and Dr. Josiah Poon, I have gained invaluable knowledge about the intricacies of research problems and the methodologies necessary for effective problem-solving. They have furnished me with essential research tools and techniques while fostering the development of my unique perspective and voice. Their guidance has not only shaped my academic development but also instilled a mindset that will continue to impact all aspects of my life.

Their dedication to nurturing a collaborative and inclusive research environment has significantly contributed to my development as a scholar. The opportunities they have provided for engaging with fellow researchers, presenting my work at conferences, and collaborating on projects have enriched my academic experience and fostered a sense of community and camaraderie essential to my success.

I am truly privileged to have learned from mentors of their caliber and will always treasure the lessons they have imparted. The immeasurable impact Dr. Simon Poon and Dr. Josiah Poon have had on my life and career will forever be appreciated, and I am deeply grateful for the honor of being their student.

My heartfelt appreciation goes to my parents, whose unwavering support and encouragement have formed the foundation of my academic journey. Their silent concern for my well-being and steadfast belief in my capabilities have empowered me to persevere through adversity. They have provided a solid base for my growth, enabling me to focus on my research and immerse myself in the quest for knowledge and intellectual fulfillment.

Their love, sacrifice, and commitment to my success have been the pillars sustaining me during the most challenging times. They have consistently offered advice, encouragement, and a listening ear when needed most. Their unyielding faith in my potential has been a constant source of motivation, inspiring me to overcome obstacles and strive for excellence in all my endeavors.

I am profoundly grateful to my loving wife and our little one, whose warmth and happiness have brought light into my life. Your unwavering care and support have been a beacon during the darkest hours, reminding me of the importance of family and the joy of being surrounded by loved ones. You have both taught me to cherish simple moments and find balance between academic pursuits and personal well-being. Your presence has been a source of comfort, motivation, and inspiration, encouraging me to tackle every challenge with resilience and emerge stronger.

Together, we have created a home filled with love, laughter, and hope. Your patience, understanding, and resilience have been instrumental in helping me navigate the highs and lows of this journey, and your unwavering belief in me has fueled my determination to succeed. I am forever indebted to you for your selflessness, compassion, and love.

Lastly, I am thankful for the opportunity to have lived and studied in Sydney. My journey from China to the United States, and ultimately to Australia, has been one of tremendous personal growth and cultural enrichment. Along the way, I have encountered cultural shocks that have expanded my horizons and deepened my understanding of the world and its diverse inhabitants.

This unique experience has been invaluable in molding my worldview and fostering a deep appreciation for the richness of human experiences across various cultures. The lessons I have learned in Sydney have challenged my preconceived notions and assumptions, forcing me to confront my own biases and prejudices and encouraging me to adopt a more inclusive and empathetic approach to life.

I have been fortunate to meet and interact with people from all walks of life, whose stories and experiences have left an indelible mark on my heart and mind. Their kindness, generosity, and resilience have inspired me to be a better person, a more compassionate researcher, and a more engaged global citizen. I am truly grateful for these encounters and the profound impact they have had on both my personal and professional growth.

Contents

Abstract ii

Author Attribution iv

Thesis Statement of Originality vi

Acknowledgements vii

Contents x

List of Figures xv

Chapter 1 Introduction 1

1.1 Real-World Data in Medicine 2

1.1.1 Predictive Learning on RWD 3

1.1.2 Low Quality Challenges 3

1.2 Strategies for Low Quality Issues 4

1.3 Contributions 6

Chapter 2 Background 9

2.1 Data Pre-processing for Low Quality Challenges 13

2.1.1 Feature Selection Approaches 13

2.1.1.1 Filter Methods 13

2.1.1.2 Wrapper Methods 15

2.1.1.3 Embedded Methods 18

2.1.2 Resampling Approaches for Imbalanced Data 20

2.1.3 Data Imputation for Missing Values 25

2.1.3.1 Case Deletion 27

2.1.3.2 Univariate Imputation 28

2.1.3.3 Machine Learning-Based Imputation Techniques 32

2.2 Weakly Supervised Learning 37

2.3 Multi-Instance Learning 39

2.3.1 Classical MIL Methods 44

2.3.2 Multi-Instance Neural Network 46

2.3.2.1 Multi-Instance Pooling 47

Chapter 3 AMI-Net 49

3.1 Methodology 50

3.1.1 Multi-Instance Pooling (MIL Pooling) 52

3.1.2 Attention-based MIL Pooling 53

3.1.3 Multi-Head Attention 55

3.2 Experiments 58

3.2.1 Datasets 58

3.2.2 Experimental Setup 60

3.2.3 Comparison with Baseline Models 61

3.3 Results and Analysis 64

3.3.1 Comparison with Different Models 64

3.3.2 Comparison of Different Number of Heads 66

3.3.3 Comparison of Different MIL Pooling Methods 67

3.3.4 Influence of Data Noise 68

3.3.5 Influence of Incomplete Data 70

3.3.6 Visualization of Attention 71

3.4 Summary 72

Chapter 4 AMI-Net+ 74

4.1 Model Architecture 76

4.1.1 Self-Adaptive Multi-Instance Pooling 78

4.1.2 Focal Loss for Imbalanced Data 80

4.2 Experiments 81

4.2.1 Data Description 81

4.2.2 Experimental Setup 82

4.3 Results and Analysis 83

4.3.1 Comparison with Baseline Models 83

4.3.2 Comparison of AMI-Net+ with Different Number of Heads 84

4.3.3 Comparison of AMI-Net+ with Different Multi-Instance Pooling Methods 86

4.3.4 Evaluation of Focal Loss 88

4.4 Summary 89

Chapter 5 AMI-Net3 90

5.1 Proposed Method 91

5.1.1 Feature Transformation and Standardization 93

5.1.2 Feature Embedding 96

5.1.3 Gated Attention-based MIL Pooling 99

5.1.4 Model Training 100

5.1.4.1 Auxiliary Supervision 101

5.1.4.2 Focal Loss 102

5.2 Experiments 103

5.2.1 Data Description 103

5.2.2 Experimental Setup 104

5.2.3 Baseline Methods 105

5.3 Results and Analysis 106

5.3.1 Performance on Clinical Risk Prediction 106

5.3.2 Evaluation of Integrated Modules in MINN 109

5.3.3 Scenario Analysis: Employing BF Exclusively 110

5.3.4 Assessing the Influence of MIL Pooling 112

5.3.5 Comparison of Different MIL Pooling Techniques 113

5.4 Limitations 114

5.5 Summary 115

Chapter 6 AMI-Net for Regression 117

6.1 Background of Warfarin Dose Prediction 118

6.1.1 Methodology 122

6.1.2 Multi-Head Attention 122

6.1.3 Feed-Forward Network 125

6.1.4 Multi-Instance Pooling 125

6.2 Experiments 126

6.2.1 Data Description 126

6.2.2 Experimental Settings 126

6.2.3 Baseline Models 127

6.3 Results and Analysis 128

6.3.1 Performance Comparisons 128

6.3.2 Impact of Multi-Head Attention 129

6.3.3 Dose Subgroup Analysis 130

6.3.4 Different Feature Sets 132

6.4 Summary 133

Chapter 7 Conclusions 134

7.1 Future Work 136

Bibliography 141

Appendix List of Ph.D. Publications 173

List of Figures

1.1 Four Different Versions of AMI-Net 8

2.1 The process of forward section, backward selection, and stepwise selection. 17

2.2 Differences between undersampling and oversampling. 21

2.3 The illustration of missing value types. 26

2.4 The process of MICE method. 30

2.5 The process of GAN method. 35

2.6 Demonstration of bags and instances. A bag receives a positive label if it contains at least one positive instance within it. Conversely, if all instances within a bag are negative, the bag itself is labeled as negative. The process of training and learning in MIL equips the model with the ability to predict the class of an unlabeled bag effectively. 40

2.7 Difference between instance-level approach and bag-level approach in MIL. 42

3.1 The overview of AMI-Net. 52

3.2 The architecture of multi-head attention. 56

3.3 Comparison of different number of heads. 0 denotes the model without multi-head attention. 67

3.4 Comparison of different MIL pooling methods on the TCM dataset. 68

3.5 Comparison of different MIL pooling methods on the WM dataset. 69

3.6 Test for the influence of feature and label noise 70

3.7 Test for the influence of incomplete data 71

3.8 An example of informative instances selection 72

4.1 The overall architecture of AMI-Net+ 77

4.2 Comparison of AMI-Net+ with different number of heads in the multi-head attention. 86

4.3 Comparison of AMI-Net+ with different multi-instance pooling methods on instance level. 87

5.1 Continuous and ordinal features are referred to as CF, while binary features or those obtained through one-hot decoding from nominal features, such as headache, diagnosis, and medication history, are denoted as BF. The proposed feature embedding technique enables the method to effectively convey a comprehensive patient narrative in an informative embedding space using only observed data. Subsequently, the MINN algorithm analyzes the available information and generates the final output. The entire process is trained concurrently with both main and auxiliary supervision to ensure optimal performance. 92

5.2 The comprehensive architecture of AMI-Net3 comprises two components for model training: auxiliary supervision and main supervision. Auxiliary supervision involves two shallow neural networks specifically designed for bf and cf respectively. On the other hand, the main supervision part consists of the primary computational modules, where 2*Conv1D represents the application of two-layer convolutions. Both components utilize the focal loss function to optimize the performance of AMI-Net3. 93

5.3 To illustrate the process of feature transformation and standardization for both binary features (BF) and continuous/ordinal features (CF), a concrete example is provided. It is worth noting that, in this analysis, only BF entries that have received positive responses are considered for further investigation. 94

5.4 The objective of the feature embedding module is to assign a unique parameter vector to each input instance. These parameter vectors represent the instance attributes and are subsequently subjected to multi-head attention for further processing. The aim is to capture the specific characteristics of each instance and enable effective attention-based operations to extract meaningful information from the input data. 98

5.5 We run experiments to evaluate the value of the computational modules we integrate in MINN. (95% CI) 111

5.6 The experiment is to evaluate the effectiveness of the proposed feature embedding. (95% CI) 112

6.1 **(a)** An example of common data processing way. **(b)** An example of how transforming and processing the original data from each individual patient. 121

6.2 **(a)** The figure outlines the overall structure of the proposed methodology, which consists of two primary levels. The first level incorporates a feature embedding module designed to convert the observed information into an embedded space. The second level comprises a sophisticated neural network armed with multi-head attention, a feed-forward network, and multi-instance pooling, all intended to decipher and unravel hidden patterns among features for an accurate final prediction. Notably, these two levels are trained in a unified, end-to-end manner. **(b)** The methodology employs multi-head attention to reveal correlations among embedded vectors, i.e., observed features. Importantly, multi-head attention enables the detection of relationships between G_{rep} and other embedded vectors. G_{rep} acts as a representative of overall patient information, facilitating further processing and analysis. 123

6.3 The Performance Comparisons of Different Number of Heads 130

6.4 The IPPs Comparison in Different Dose Subgroups 131

CHAPTER 1

Introduction

Medical evidence is crucial for the advancement of healthcare, as it provides the foundation upon which clinicians base their decisions, guidelines are developed, and policies are established [14, 211]. The generation of high-quality medical evidence is paramount to ensure the safety and effectiveness of interventions, leading to improved patient care and better health outcomes [93, 100, 136]. Over the years, researchers have developed various methods to generate medical evidence, with randomized controlled trials (RCTs) being considered the "gold standard" due to their ability to minimize biases and establish causal relationships between interventions and outcomes [51, 58, 71, 216].

However, conducting RCTs can be challenging, especially when studying special populations such as pregnant women and newborns [162, 227]. Moreover, RCTs are often expensive, time-consuming, and may have limited applicability to real-world settings. As a result, researchers and clinicians have increasingly turned to alternative sources of medical evidence, such as real-world data (RWD), to complement RCT findings and provide a more comprehensive understanding of the safety and effectiveness of interventions in real-world settings [69, 88, 219].

1.1 Real-World Data in Medicine

Real-world data (RWD) in medicine refers to the vast and diverse array of data collected from sources beyond the controlled environment of traditional clinical trials. These sources include electronic health records (EHRs), medical claims and billing data, patient registries, wearable devices, social media platforms, and even patient-generated health data from various health apps [22, 219]. Unlike the highly structured, rigorously controlled data obtained from clinical trials, RWD reflects the everyday experiences of patients in real-life healthcare settings, capturing the complexity and heterogeneity of patient populations, their conditions, and the treatments they receive [25].

The growing interest in real-world data is fueled by the increasing need for evidence-based medicine, which seeks to optimize clinical decision-making by integrating individual clinical expertise with the best available external evidence from systematic research [192]. RWD has the potential to enhance understanding of disease patterns, treatment effectiveness, safety profiles, and patient outcomes in a way that is more representative of the broader patient population [163]. This is particularly relevant given that clinical trial participants are often a highly selected group that may not adequately represent the full spectrum of individuals living with a particular condition [207].

Furthermore, real-world data provides a unique opportunity to study the long-term effects of medical interventions, identify rare adverse events, and monitor the performance of healthcare systems [43]. It can also be used to support comparative effectiveness research, which is crucial for determining the most appropriate therapeutic interventions for specific patient groups [231]. Moreover, the rich information contained in RWD can be harnessed to develop advanced predictive models, enabling personalized medicine and more efficient resource allocation in healthcare systems [183].

1.1.1 Predictive Learning on RWD

Predictive learning based on RWD represents an emerging and transformative approach in medical informatics, leveraging the power of artificial intelligence (AI) and machine learning (ML) techniques to derive actionable insights from large-scale, observational datasets. By extracting patterns and relationships from diverse and complex data sources, predictive learning can provide valuable information on disease progression, treatment effectiveness, patient outcomes, and resource utilization [18].

The application of predictive learning in medicine has the potential to revolutionize healthcare by enabling more accurate prognostication, early detection of diseases, identification of optimal treatment strategies, and personalized care tailored to individual patients' needs and characteristics [245]. Furthermore, predictive learning can facilitate the development of clinical decision support systems, allowing clinicians to make more informed, data-driven decisions in real-time, ultimately leading to improved patient outcomes and more efficient healthcare systems [221].

1.1.2 Low Quality Challenges

Despite its promise, the implementation of predictive learning based on RWD also presents challenges. One of the most significant concerns with RWD is the potential for low data quality, which can introduce biases and ultimately compromise the validity and reliability of findings derived from these sources. There are several factors contribute to low quality data including missing values, redundant and highly correlated features, insufficient sample sizes, label noise, and imbalanced data [260, 262]. Each of these issues can impact the performance and interpretability of predictive learning models, leading to erroneous conclusions and potentially compromising patient care [37].

Missing values often arise due to incomplete records or inconsistencies in data collection practices, reducing the statistical power and potentially introducing biases in the analysis [15]. Redundant features and feature correlations can result from duplicated or highly correlated variables, leading to multicollinearity and overfitting in predictive models [61, 117]. Insufficient sample sizes can limit the generalizability of findings and hinder the identification of significant associations between variables, leading to type II errors [246]. Label noise, which refers to errors in the assignment of outcome variables, can lead to misclassification and negatively impact model performance [121]. Imbalanced data, where the distribution of outcome variables is highly skewed, can result in biased models that favor the majority class, thereby undermining the predictive accuracy for minority classes [21].

Addressing these low-quality data issues is crucial for ensuring the reliability and validity of predictive learning models based on real-world data.

1.2 Strategies for Low Quality Issues

Previously, numerous strategies have been proposed to address these problems, primarily focusing on the data pre-processing part with three directions: feature selection to eliminate redundant or highly correlated features, resampling techniques for imbalanced data, data imputation for handling missing values [30, 39, 54, 60, 167, 189, 212]. These approaches also have collectively contributed to enhancing model robustness and mitigating the adverse effects of noisy data and small datasets.

Despite the substantial contributions of these approaches, the need to handle the increasingly complex nature of data has highlighted significant limitations [263]. The reductionist nature of feature selection can result in the loss of essential features or complex interactions, limiting its overall efficiency and effectiveness [264]. Resampling techniques have their drawbacks as they might lead to the over-generalization of the minority class or cause overfitting issues. Similarly, data imputation methods, while inherently valuable, can introduce bias and distortion if not carefully executed [116].

In light of these limitations and in response to the swift advancement of machine learning and deep learning technologies, the paradigm of learning directly from low-quality data for predictive purposes has emerged as a focal point in contemporary research. A pertinent example is seen in the healthcare sector where patient data often fails to provide a comprehensive narrative. Due to patients not undergoing all possible tests and variations in information recording standards, complete data documentation is often a formidable task [137]. Addressing this issue invites the development of probabilistic models predicated on certain assumptions regarding the missing data mechanism, allowing for decision functions to be derived solely from observed data. This innovative approach aims to harness the inherent structure within available data and seeks to provide more accurate predictions and deeper insights, despite the challenges presented by low-quality or incomplete datasets [78, 142, 220, 224, 225, 228].

In the face of different challenges of low quality data issues, a multitude of strategies have been developed to identify valid data and informative features, facilitating direct learning without the necessity for prior data pre-processing, such as EM-DD, mi-SVM, mi-Graph, miFV, and Unicon [126, 285, 296]. However, these methods typically fail to address a majority of low-quality issues simultaneously, instead only resolving a portion of the problems at a time. As a result, the robustness and suitability for clinical application of these methods remain inadequate.

1.3 Contributions

The majority of my PhD study probes into the uncharted territory of merging diverse strategies for direct learning from medical data that encompasses most potential low quality issues. The objective is to form an amalgamated, robust predictive learning mechanism, specifically tailored for the unique challenges and requirements of the medical field.

From the standpoint of machine learning, this scenario is typically categorized as weakly supervised learning (WSL), with multi-instance learning (MIL) being a representative example [294]. Based on the assumptions of MIL, neural networks are proposed to be integrated to perform predictive learning directly on low-quality medical data without any data pre-processing strategies.

Moreover, this work propose a concept that treats each patient as a 'bag of instances', namely symptoms, and projects them into an embedding space. In this space, instances correlate with each other across different embedding dimensions, each representing a specific bodily condition. Following this, a Multi-Instance Learning (MIL) Neural Network is performed on it to identify informative instances and obtain the bag score for the final prediction.

Based on this assumption:

- This thesis presents one of the pioneering implementations of a Multi-Instance Learning (MIL) neural network, harmoniously integrated with an embedding method, meticulously crafted for predictive learning within the medical domain.
- This approach eliminates the need for intentional data collection or manual data screening, instead automatically processing the available data.

- It is capable of autonomously identifying any critical feature or sign amidst a large volume of low-quality data.
- Notably, this strategy's application is not limited to the domain of Western Medicine (WM), but it is also applicable in the field of Traditional Chinese Medicine.
- This method exhibits high scalability and can be seamlessly applied to a diverse range of predictive tasks, including classification and regression, requiring minimal adjustments.

At the first attempt in this field, the foundational architecture model on binary data for the classification task is proposed, named AMI-Net (Chapter 3). This model can effectively learn directly from both WM and TCM datasets, despite their high-dimensional feature space and a significant amount of missing values. However, AMI-Net's capability to select informative instances is constrained by its pooling method, and it struggles to handle imbalanced data effectively. To overcome these limitations, AMI-Net+ (Chapter 4) is introduced, which incorporates an innovative self-adaptive MIL pooling method to improve key instance identification and utilizes a focal loss approach to address the imbalanced data issue. It is important to note that both AMI-Net and AMI-Net+ primarily model binary data sets. To expand the model's applicability within the medical field, AMI-Netv3 is further proposed(Chapter 5). In this version, an innovative feature embedding method is designed that maps any type of features to a unified embedding space for modeling. Additionally, a novel supervision strategy is introduced, termed auxiliary supervision, to enhance the model's stability and predictive performance. While AMI-Net, AMI-Net+, and AMI-Net3 all offer significant advantages for classification tasks, they currently lack the flexibility necessary for regression tasks. In the field of medicine, regression tasks are quite common and vital, with applications such as drug dosage prediction [117] and length of stay prediction [12]. To address this, AMI-Net3 is further enhanced to accommodate regression tasks (Chapter 6). The model's efficacy has been confirmed through its successful application to the warfarin prediction task. The figure 1.1 demonstrates the 4 different versions of AMI-Net.

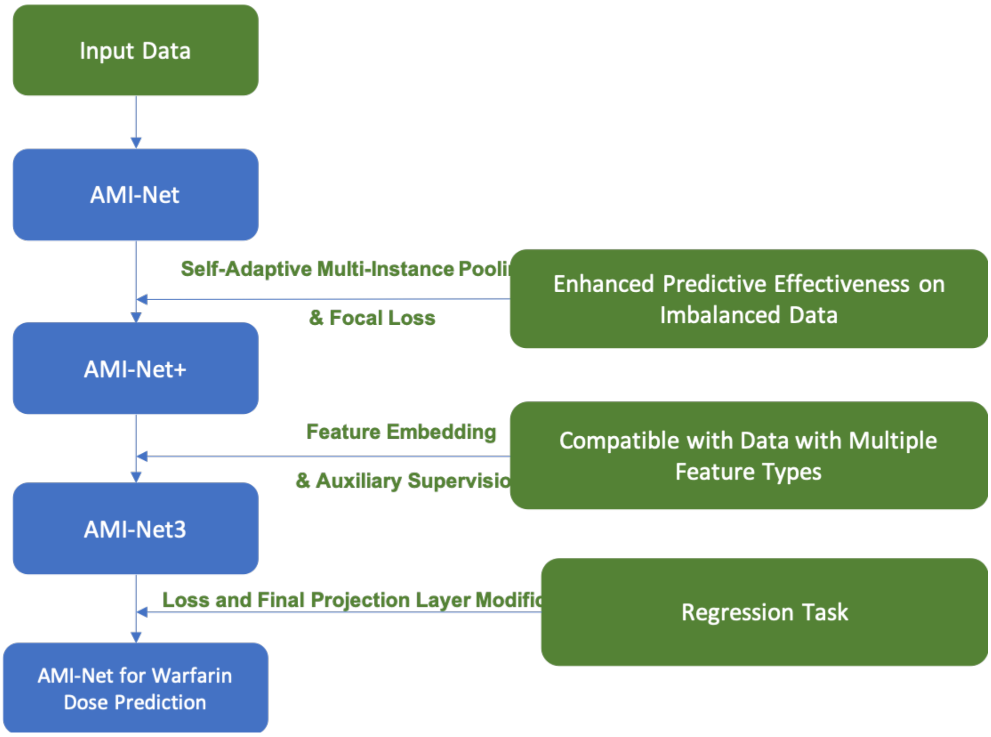

Input Data
AMI-Net
Self-Adaptive Multi-Instance Pooli
& Focal Loss
Enhanced Predictive Effectiveness on Imbalanced Data
AMI-Net+
Feature Embedding
& Auxiliary Supervisio
Compatible with Data with Multiple Feature Types
AMI-Net3
Loss and Final Projection Layer Modific
Regression Task
AMI-Net for Warfarin Dose Prediction

FIGURE 1.1: Four Different Versions of AMI-Net

To summarize, the key contributions of this thesis can be outlined as follows:

- In AMI-Net, a unified framework is proposed that combines the embedding method with the MIL neural network to directly learn from low-quality data.
- In AMI-Net+, it introduces a novel MIL pooling method designed for the detection of informative instances, and incorporate focal loss to address imbalanced prediction issues
- In AMI-Net3, it introduces a novel feature embedding method applicable to all types of features, and a unique supervision strategy aimed at enhancing classification performance.
- The functionality of AMI-Net3 is further expanded to include regression tasks.

CHAPTER 2

Background

The advent and rapid expansion of big data have sparked profound changes across various sectors, with healthcare standing at the forefront of this transformation. The exponential increase in diverse, complex, and swiftly growing data—emanating from a range of sources such as electronic health records (EHRs) [56], genomic sequencing [165], and wearable health devices has paved the way for innovative insights and bespoke treatments [158].

This surge in healthcare data has given rise to the growing importance of predictive models in medicine, a shift propelled by the evolution towards personalized healthcare [91, 105, 117, 261, 264]. These models leverage mathematical algorithms and computational techniques to anticipate possible outcomes or events in patients. They do so by examining a wide array of data including historical health records, demographics, and clinical information [80]. Consequently, predictive models have shifted the medical paradigm from a traditional symptom-based approach to a more sophisticated data-driven methodology [268]. This shift empowers healthcare professionals to deliver care that is not only more precise and individualized but also proactively anticipates patient needs [16].

The applications of predictive models in medicine are diverse and far-reaching, extending from early disease diagnosis and risk stratification [65, 141, 169, 205] to personalized treatment planning [239, 278, 280, 292] and efficient health resource management [115, 230, 250]. These models have led to the development of robust tools that can predict, for example, the risk of readmission for heart failure patients [173, 191, 217], the likelihood of diabetes onset [119, 181, 182], the progression of cancer [134, 247], and the individual response to specific pharmaceutical treatments [91, 105, 117, 261], among others.

Furthermore, numerous specialized models have been designed, each serving crucial roles from risk assessment to treatment planning and resource management [200]. Risk Stratification Models, for instance, lay the groundwork for preventive healthcare, categorizing patients based on their potential of developing specific diseases [235]. A quintessential example is the Framingham Risk Score, a broadly implemented model that anticipates a patient's 10-year risk of succumbing to cardiovascular disease [270]. By examining a range of variables, including age, sex, blood pressure, cholesterol levels, and smoking status, this model fabricates a comprehensive risk profile [55]. The output produced is instrumental in steering physicians towards identifying suitable preventive measures for patients across various risk levels [67]. This proactive approach has immense potential to manage public health by preemptively identifying at-risk individuals, thereby significantly reducing the associated healthcare costs and burden of disease.

On another front, Diagnostic Models stand as a cornerstone in disease detection and diagnosis, significantly enhancing the precision and speed of these critical processes [183]. Machine learning algorithms, notably convolutional neural networks (CNNs), have profoundly transformed radiology, enabling the detection of subtle anomalies that might escape the human eye. For instance, CNNs can efficiently identify lung nodules in CT scans, a capability crucial for early lung cancer diagnosis. This advanced diagnostic approach not only improves patient prognosis through timely intervention but also alleviates the emotional and economic stress associated with late-stage diagnoses [9, 70, 147, 243, 266].

Further, Prognostic Models provide a lens into the future health trajectory of patients by predicting disease progression or patient outcomes based on current health status and history [235]. The Acute Physiology and Chronic Health Evaluation II (APACHE II) score serves as a testament to the utility of such models [132]. Leveraged to measure the severity of disease in patients admitted to ICUs, the APACHE II score takes into account a plethora of parameters, including age, history of severe organ insufficiency, and various physiological measurements [303]. These prognostic models guide healthcare providers in forming personalized care plans, facilitating decision-making on the level of intervention needed, thereby improving patient outcomes and reducing the likelihood of overtreatment [20, 29, 111, 223].

In the same vein, Treatment Response Models represent another category of predictive models, devised to forecast a patient's response to particular treatments or medications [236]. Oncotype DX, a genomic test, exemplifies the practicality of these models in clinical settings [185]. This test predicts the recurrence of breast cancer and the likelihood of a patient benefiting from chemotherapy [232]. By doing so, it aids clinicians and patients in making informed decisions about the necessity and potential efficacy of chemotherapy, thus avoiding unnecessary treatment and its associated side effects in cases where it might be of limited benefit [2, 233].

Finally, Resource Utilization Models occupy a crucial role in healthcare management, forecasting variables such as patient flow, hospital readmissions, or ICU bed occupancy [16]. These models contribute significantly to the efficient allocation of healthcare resources, ensuring optimal utilization of medical facilities and minimizing wastage. The LACE index, which calculates the risk of unplanned readmission or death within

30 days after hospital discharge, is one such example [254]. By predicting readmissions, this model allows healthcare institutions to anticipate and manage patient loads better, contributing to improved patient care and more efficient use of hospital resources [125, 293].

The various types of predictive models, when viewed collectively, highlight the transformative potential of data-driven decision-making in healthcare. From prevention and diagnosis to prognosis, treatment, and resource management, these models are setting the stage for an unprecedented era of personalized, efficient, and high-quality healthcare services. It must be noted, however, that the success of predictive learning is deeply dependent on the quality of the underlying data, a common challenge in healthcare settings.

Indeed, healthcare data are notoriously marred by quality issues, as detailed by [210]. Complications like noise, missing values, and redundancies are commonly encountered obstacles in the quest for meaningful and reliable predictive insights [262]. Failing to address these problems can drastically undermine the accuracy and generalizability of the derived predictive models [263].

As such, data pre-processing emerges as a critical step, transforming raw, disorganized healthcare data into a format ready for sophisticated analysis. By rectifying the inherent quality challenges in medical datasets, pre-processing techniques lay a solid foundation for successful predictive modeling [282]. There is a broad spectrum of these techniques available, each crafted to overcome a specific data challenge. Notably, feature selection, data imputation, and resampling methods have demonstrated their effectiveness in resolving the key quality issues found in healthcare data: redundancy, missing values, and imbalanced data, respectively [94, 101, 157].

2.1 Data Pre-processing for Low Quality Challenges

In this section, a few state-of-art data pre-processing techniques to be introduced detailedly here. We also focus more on the algorithms that used ensemble learning and deep learning since such algorithms caught more attention in recent years.

2.1.1 Feature Selection Approaches

Feature selection is a crucial step in the development of predictive models, as it aims to identify the most informative and relevant features while reducing noise, overfitting, multicollinearity, and computational complexity. Feature selection methods can be broadly categorized into three main groups: filter methods, wrapper methods, and embedded methods. Each of these categories encompasses several techniques designed to handle different data properties and balance the trade-off between model complexity and predictive performance.

2.1.1.1 Filter Methods

Filter methods are a category of feature selection techniques that evaluate the importance of features independently of the learning algorithm [167]. They rely on statistical measures, such as correlation, mutual information, or statistical tests, to assess the relationship between individual features and the target variable. Filter methods are computationally efficient, as they do not involve the training of predictive models, but they may not always capture feature interactions or dependencies on the specific model used. Some commonly used filter methods include:

- **Pearson's Correlation Coefficient [189]:** Pearson's correlation coefficient, a number ranging between -1 and 1, is commonly used to measure the strength and direction of the linear relationship between two variables. It is calculated as the ratio of the covariance between the two variables to the product of their standard deviations. Given paired data $\{(x_1, y_1), \ldots, (x_n, y_n)\}$, the Pearson's correlation coefficient r is defined as

$$r = \frac{\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n} (x_i - \bar{x})^2} \sqrt{\sum_{i=1}^{n} (y_i - \bar{y})^2}} \tag{2.1}$$

where n is sample size, x_i, y_i are the samples indexed with i, $\bar{x}$ and $\bar{x}$ are the sample means.

- **Mutual Information (MI) [54]:** MI quantifies the extent of mutual dependence between two variables, thereby encapsulating the volume of information obtained about one variable through the observation of the other. It is applicable to both continuous and categorical features as well as target variables. Features with elevated MI scores are perceived to possess more shared information with the target variable, thereby marking them as pivotal for model construction. Let (X, Y) denote a pair of random variables with values spanning over the space $\mathcal{X} \times \mathcal{Y}$. Assuming their joint distribution to be $P_{(X,Y)}$ and the marginal distributions as P_X and P_Y, the mutual information is given by

$$I(X;Y) = D_{\mathrm{KL}}\left(P_{(X,Y)} | P_X \otimes P_Y\right) \tag{2.2}$$

where D_{KL} is the Kullback–Leibler divergence [123].

- **Information Gain [198]:** IG is a measure that quantifies the reduction in entropy, or uncertainty, of the target variable upon gaining knowledge of a specific feature. Rooted in the concept of entropy from information theory, IG is usually employed with categorical features and target variables. Features with high IG scores are considered to contribute more information about the target variable, hence their importance in model construction. Let's define the information entropy H from a preceding state that accounts for certain information:

$$IG(V, a) = \mathrm{H}(V) - \mathrm{H}(V \mid a) \tag{2.3}$$

where $\mathrm{H}(V \mid a)$ symbolizes the conditional entropy [161] of V given the value of attribute a. This interpretation is intuitively justified when viewing entropy H as a measure of the uncertainty of a random variable V: by acquiring (or assuming) a about V, the uncertainty regarding V diminishes (i.e. $IG(V, a)$ is positive), unless V is independent of a, in which case $\mathrm{H}(V, a) = \mathrm{H}(V)$, and thus, $IG(V, a) = 0$.

2.1.1.2 Wrapper Methods

Wrapper methods for feature selection are techniques where the feature selection process is guided by the accuracy performance of a predetermined predictive model. [133] The term "wrapper" is derived from the fact that these methods "wrap" themselves around the predictive model to evaluate the usefulness of subsets of features based on the model performance.

The procedure typically involves creating various subsets of features and using the predictive model to assess their performance. Common strategies to form these subsets include forward selection, backward elimination, and recursive feature elimination. [30, 39, 212].

- **Forward Selection and Backward Elimination:** These methods aim to find the optimal subset of features that maximizes the performance of a specific machine learning algorithm for a given problem.

Forward Selection: Forward selection starts with an empty model and adds predictors to the model, one-at-a-time, until no other predictors can be added to improve the model to a statistically significant extent. At each step, the predictor that gives the greatest additional improvement to the model is included. The process starts with a simple model and gets increasingly complex. It stops when adding any of the remaining variables would not improve the performance of the model by a statistically significant amount.

Backward Elimination: In contrast to forward selection, backward elimination starts with a full model that includes all potential predictors, and it removes predictors one-at-a-time. At each step, the predictor that is the least significant (or that detracts the most from the performance of the model) is removed. The process starts with a complex model and simplifies it step by step. It stops when removing any more of the variables would significantly worsen the performance of the model.

There is also a combination of the two, known as bidirectional elimination or stepwise selection, which both adds and removes predictors as part of the model building process (Shown as Figure 2.1). These methods should be used as part of an exploratory analysis, rather than for confirming specific hypotheses.

- **Recursive Feature Elimination (RFE) [60]:** RFE aims to identify the subset of features that contributes most to the prediction of the target variable. It is a wrapper-type feature selection algorithm, which means that it uses a machine learning algorithm and its performance as a measure to evaluate the importance of features. The steps of RFE are as follows:

  Step 1: RFE begins with a machine learning model trained on the initial set of features.

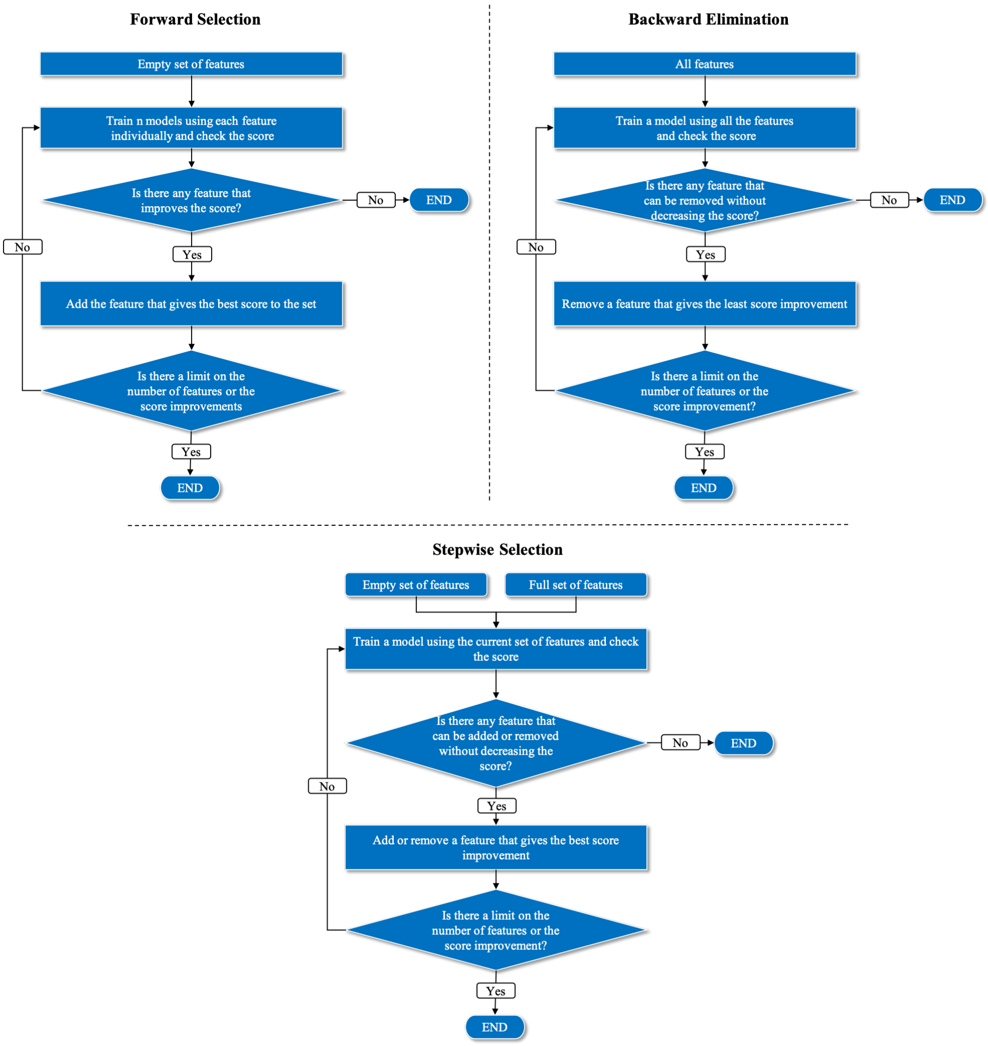

Forward Selection
Empty set of features
Train n models using each feature individually and check the score
Is there any feature that improves the score?
No
END
Yes
Add the feature that gives the best score to the set
Is there a limit on the number of features or the score improvements
No
Yes
END
Backward Elimination
All features
Train a model using all the features and check the score
Is there any feature that can be removed without decreasing the score?
No
END
Yes
Remove a feature that gives the least score improvement
Is there a limit on the number of features or the score improvement?
No
Yes
END
Stepwise Selection
Empty set of features
Full set of features
Train a model using the current set of features and check the score
Is there any feature that can be added or removed without decreasing the score?
No
END
Yes
Add or remove a feature that gives the best score improvement
Is there a limit on the number of features or the score improvement?
No
Yes
END

FIGURE 2.1: The process of forward section, backward selection, and stepwise selection.

Step 2: The importance of each feature is obtained either directly from the model (for those models that provide a way to rank feature importance, such as decision trees or linear models with coefficients), or by training the model multiple times, each time leaving out one of the features and measuring the decrease in performance.

Step 3: The least important features are pruned from the current set of features. This step involves discarding a specified number of least important features.

Step 4: The model is retrained on this pruned subset of features.

Step 5: Steps 2 to 4 are recursively repeated on the pruned set until the desired number of features is eventually reached.

RFE can be computationally expensive due to the need to repeatedly train models, but it has the advantage of taking into account the interactions between features and providing a ranking of features according to their importance [153].

2.1.1.3 Embedded Methods

Embedded methods integrate the process of feature selection within the construction of the machine learning model itself. They are more computationally efficient than wrapper methods and can capture complex feature interactions that filter methods might overlook [206]. The primary embedded methods for feature selection encompass strategies grounded in linear regression, decision tree algorithms, and neural network models.

- **Linear Regression Approaches:** These are statistical methods used for predicting a response variable. They include methods like Ridge Regression, Lasso, and Elastic Net:

  Lasso (Least Absolute Shrinkage and Selection Operator) [203]: Lasso performs L1 regularization, which adds a penalty equal to the absolute value of the magnitude of coefficients. This type of regularization can result in sparse models with few coefficients; Some coefficients can become zero and eliminated from the model. Larger penalties result in coefficient values closer to zero, which is the ideal for producing simpler models.

  Ridge Regression [168]: Ridge performs L2 regularization, i.e., it adds a penalty equivalent to the square of the magnitude of coefficients. This penalty term in the loss function forces the learning algorithm to not only fit the data but also keep the model weights as small as possible. It does not eliminate coefficients (like Lasso) but it will shrink them.

Elastic Net [304]: Elastic Net is a hybrid of Lasso and Ridge Regression techniques. It is trained with L1 and L2 prior as regularizer. Elastic-net is useful when there are multiple features that are correlated. Lasso is likely to pick one of these at random, while elastic-net is likely to pick both.

- **Decision Tree-based Approaches:** Decision trees inherently perform feature selection as they choose a subset of input features to split on at each node. They choose the feature that provides the most informative split according to some mathematical criteria like information gain or gini impurity, such as bagging strategy based Random Forest [196] and boosting strategy based Gradient Boosting tree [47, 127, 180]. Random Forest consists of a collection of decision trees. It provides an indication of the importance of features, computed as the total reduction in the criterion brought by that feature. Gradient Boosting Tree can also provide feature importances similar to Random Forests. It builds multiple weak learners [251] in a sequential manner where each subsequent model learns from the errors of its predecessor. The feature importances are calculated based on the number and performance of the splits across all trees.
- **Neural Network-based Approaches:** Traditional neural networks don't perform explicit feature selection but can be regularized to induce sparsity, which effectively results in feature selection.

  Regularization Techniques (L1, L2, Dropout) [190, 267, 271]: Regularization techniques, including L1, L2, and Dropout, serve as essential mechanisms to prevent overfitting in neural networks. Notably, L1 regularization can induce sparsity in the model, thereby performing a form of feature selection by prioritizing essential features and reducing the impact of less significant ones.

Autoencoders [193]: Autoencoders, a specific category of artificial neural networks, are engineered for learning efficient encodings of input data. Even though they're not traditionally associated with feature selection, their bottleneck layer - which captures a compressed representation of the input data - can be effectively utilized for this purpose, hence providing a condensed, yet rich, set of features.

Attention Mechanisms [10, 218]: Although not strictly a feature selection method, attention mechanisms in neural networks serve an essential role in providing insights into which parts of the input data the model deems significant during the learning process. As such, they indirectly highlight features of importance and contribute to the overall understanding of feature relevance in the model.

2.1.2 Resampling Approaches for Imbalanced Data

Imbalanced data is a common problem in machine learning where the classes in the target variable are not represented equally [101]. In such cases, standard machine learning models often show a bias towards the majority class, leading to inaccurate and unreliable predictions for the minority class [122]. Resampling techniques are frequently employed to counteract the imbalance in datasets, striving to achieve a balanced class distribution. These techniques can be broadly categorized into undersampling and oversampling methods. As depicted in Figure 2.2, the distinction between oversampling and undersampling techniques becomes evident, marking their unique contribution to balancing the dataset and hence improving model performance [171].

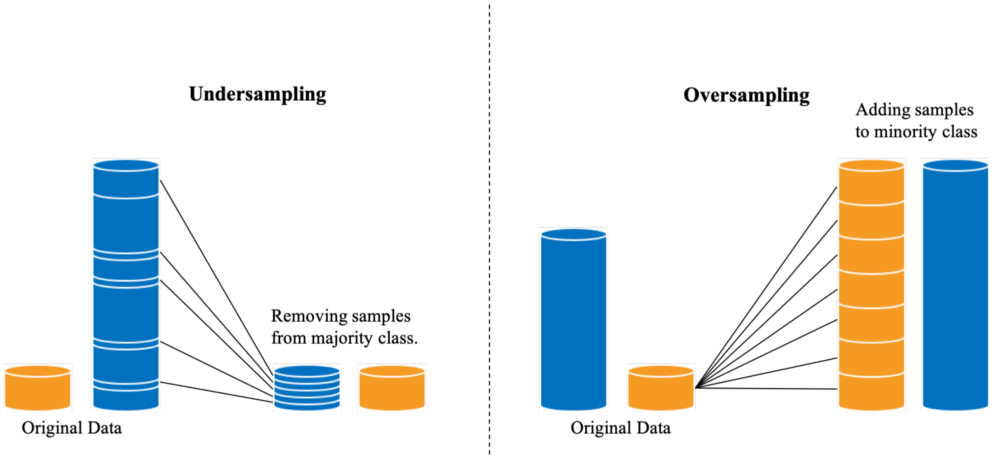

Undersampling
Removing samples from majority class.
Original Data
Oversampling
Adding samples to minority class
Original Data

FIGURE 2.2: Differences between undersampling and oversampling.

- **Undersampling Approaches:** Undersampling involves reducing the quantity of instances or samples from the majority class. This method effectively diminishes the dominance of the majority class, allowing the model to pay more attention to the minority class during the learning process. Here are some commonly used techniques:

  Random Undersampling [155]: This is the simplest form of undersampling and involves randomly removing instances from the majority class. Although it's straightforward to implement, it may lead to loss of information if instances that are potentially important to the decision function of a classifier are removed.

  Evolutionary-based Techniques [74, 289]: They aim to solve a binary optimization problem to determine which samples should be removed, but their application is restricted to small datasets due to computational constraints.

  Neighborhood-based Methods [84]: Neighborhood-based methods leverage the principle of the k-nearest neighbor rule, while clustering-based techniques segment data samples into distinct clusters with the aim of discarding noisy and less informative samples. These include classic algorithms like the Condensed Nearest Neighbor (CNN) algorithm [97], the Tomek-link (TL) and all k-nearest neighbors (AllKNN) algorithms [244], and the Edited Nearest Neighbors (ENN) algorithm [269]. Some combinations of these methods, such as the One-Sided Selection (OSS) method (which combines TL and CNN) [135] and the Neighborhood Cleaning Rule (NCR) (which combines CNN and ENN) [138], have been developed to improve performance. The Near Miss (NM) algorithm selects samples from the majority class based on the distance between the majority and minority classes [164], while the Instance Hardness Threshold (IHT) [226] method focuses on removing "hard samples" that are likely to be misclassified.

Clustering-based Approaches [288]: Clustering-based approaches partition data samples into clusters, aiming to eliminate noisy and less informative samples. The core idea behind clustering-based undersampling is to identify clusters within the majority class and then either downsample within each cluster or represent the cluster by its centroid or some other representative instance. This strategy maintains the general structure and distribution of the majority class while reducing its size. Different algorithms can be used for the clustering process. One commonly used algorithm is the K-means clustering algorithm, which partitions data into K distinct, non-overlapping subsets or clusters. The centroid of a cluster is used as a representative of that cluster. Other clustering methods like DBSCAN or hierarchical clustering can also be applied depending on the nature of the data [145]. After the clustering process, instances from each cluster can be randomly removed until the desired class balance is achieved. Alternatively, instances in a cluster can be replaced by the cluster centroid or another representative instance, effectively reducing the number of instances in the majority class.

- **Oversampling Approaches [99]:** Oversampling seeks to augment the quantity of instances or samples from the minority class, either by duplicating existing instances or generating new ones. This method bolsters the presence of the minority class in the dataset, facilitating its recognition by the model. In addition to random oversampling techniques, that instances from the minority class are randomly duplicated to balance the class distribution [171], the common used methods are:

Synthetic Minority Over-Sampling Technique (SMOTE) [42]: SMOTE and its derivative models [28, 95, 102, 160], standing for Synthetic Minority Over-Sampling Technique, is an oversampling procedure that synthesizes new instances from the minority class. The ultimate goal is to achieve a balanced or near-balanced training set, contributing to a more robust classifier training process. SMOTE constructs synthetic samples as linear interpolations of two instances (x and x_{random}) belonging to the minority class, described mathematically as:

$$s = x + u \cdot (x_{random} - x) \tag{2.4}$$

where, $0 \leq u \leq 1$, and x_{random} is a randomly selected instance from the 5 nearest neighbors of x within the minority class.

Importantly, SMOTE does not induce additional correlation amongst different variables. However, the synthetic samples created via SMOTE maintain a strong positive correlation with the original minority instances used to generate them (x and x_{random}), as well as with other synthetic instances derived from the same original instances. Despite the widespread use of SMOTE, it does come with certain drawbacks: (1) It has a tendency to oversample samples with low informative value [229]; (2) It also risks oversampling noisy data, potentially distorting the underlying data distribution; and (3) The process of determining the number of nearest neighbors for synthetic sample creation is challenging, and the selection of these neighbors lacks a clear direction, leading to potential inconsistencies. In this context, Generative Adversarial Networks (GANs) have been proposed.

Generative Adversarial Networks (GANs) [82, 92, 154, 184, 301]: GANs have been celebrated for their remarkable ability to generate realistic, diverse synthetic data. A typical GAN framework comprises two components: a Generator that creates synthetic data and a Discriminator that evaluates the authenticity of the generated data. These two components are trained simultaneously in a competitive setting, with the generator progressively improving its ability to generate synthetic instances that the discriminator cannot distinguish from real instances.

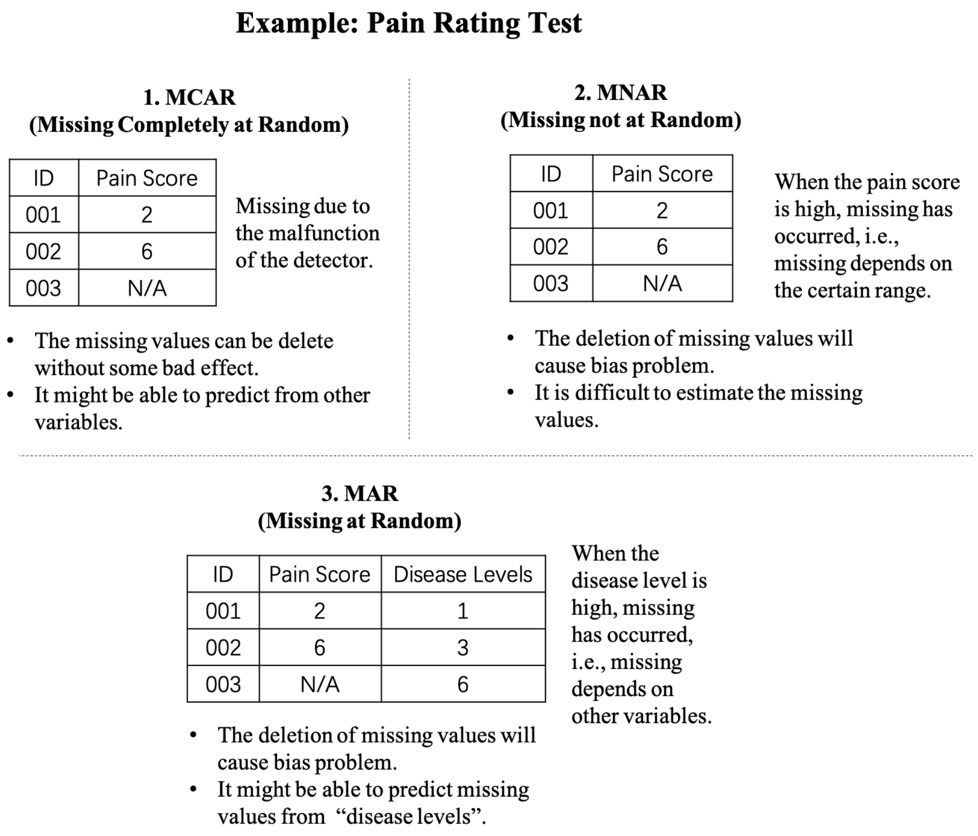
In the realm of image generation, as demonstrated in works such as [152, 199], Generative Adversarial Networks (GANs) have exhibited outstanding performance. GANs have also been explored in the context of generating adversarial examples, with various approaches being developed. For instance, [41, 146] employ GANs to produce malicious network traffic data, while [113] leverage similar models to synthesize malware samples.

In the domain of modeling tabular data, Generative Adversarial Networks (GANs) have proven to outperform classical methods for synthetic data generation, as evidenced by algorithms such as TGAN [272] and CTGAN [273]. Several GAN-based models have been proposed for the task of tabular data synthesis. For instance, CTAB-GAN [291] is a conditional table GAN that can effectively model diverse data types with complex distributions. In [175], a Cramer GAN, categorical feature embedding, and a Cross-Net architecture are utilized to synthesize Passenger Name Records (PNRs) data. GANs have also been used to generate continuous time series on Electronic Health Records (EHR) data [276], and to synthesize high-dimensional discrete variables from EHR data using the MedGAN model [49], which combines an autoencoder with a GAN. Additionally, TableGAN [187] utilizes a convolutional Discriminator and a de-convolutional Generator and a Classifier to ensure the semantic consistency of the synthetic data. Lastly, CopulaGAN [33], a variant of the CTGAN model, employs Cumulative Distribution Function (CDF)-based transformation to facilitate the training of the CTGAN model.

2.1.3 Data Imputation for Missing Values

In the realm of data analysis and machine learning, dealing with incomplete data presents a persistent challenge, particularly in healthcare where the nature of patient examinations and treatments often results in sparsely populated datasets. Patients, by virtue of their unique health profiles, do not undergo identical examinations, leading to varied and often missing data points [260, 262, 263]. Moreover, there are different types of missing values [66] (Shown in Figure 2.3).

- **Missing Completely at Random (MCAR) [66]:** MCAR is a mechanism in which the probability of an observation being missing is unrelated to any other observed or unobserved data. In other words, the fact that the data is missing is independent of any known or unknown variables. If data are MCAR, then excluding cases with missing data does not bias the results, although it may lead to loss of efficiency.
- **Missing at Random (MAR) [215]:** MAR is a mechanism in which the probability of an observation being missing is related to the observed data but not the missing data. If data are MAR, ignoring the missingness when conducting analyses could lead to biased results. However, missing data imputation methods can produce unbiased results under MAR.



Example: Pain Rating Test

1. MCAR
(Missing Completely at Random)

ID	Pain Score
001	2
002	6
003	N/A

Missing due to the malfunction of the detector.

- The missing values can be delete without some bad effect.
- It might be able to predict from other variables.

2. MNAR
(Missing not at Random)

ID	Pain Score
001	2
002	6
003	N/A

When the pain score is high, missing has occurred, i.e., missing depends on the certain range.

- The deletion of missing values will cause bias problem.
- It is difficult to estimate the missing values.

3. MAR
(Missing at Random)

ID	Pain Score	Disease Levels
001	2	1
002	6	3
003	N/A	6

When the disease level is high, missing has occurred, i.e., missing depends on other variables.

- The deletion of missing values will cause bias problem.
- It might be able to predict missing values from "disease levels".

FIGURE 2.3: The illustration of missing value types.

- **Missing Not at Random (MNAR) [204]:** MNAR is a mechanism in which the probability of an observation being missing is related to the missing data, even after controlling for the observed data. In this case, ignoring the missingness when conducting analyses will generally lead to biased results. MNAR data require more sophisticated techniques to handle, such as sensitivity analyses or methods that explicitly model the missing data mechanism.

Missing data is a pervasive issue that can adversely affect the interpretation and generalizability of research findings. Various methods have been developed to address this problem, some of which are more traditional and straightforward than others. This work first delve into these traditional approaches, their implementation, and their inherent limitations.

2.1.3.1 Case Deletion

Case deletion, also known as complete case analysis, is one of the most basic ways of handling missing data. There are two primary types of case deletion: listwise and pairwise [124].

- **Listwise Deletion:** In listwise deletion, an entire observation is excluded from analysis if any single value is missing. This method is simple and easy to implement but can lead to significant loss of data, especially if the missingness is widespread across variables. It also assumes that data are Missing Completely at Random (MCAR), an assumption that is rarely met in real-world datasets [4].
- **Pairwise Deletion:** Unlike listwise deletion, pairwise deletion (or available case analysis) uses all available data for each analysis. When calculating statistics, it includes every case that has valid data for that particular calculation. This method can lead to a more efficient use of data compared to listwise deletion, but it can also result in different analyses being based on different subsets of data, making comparisons challenging.

While the case deletion method provides a straightforward solution for data imputation, the potential loss of valuable information and the risk of skewing the data distribution are significant drawbacks that cannot be ignored. Given these considerations, the univariate imputation methods are proposed. Unlike case deletion, which entirely removes instances with missing values, univariate imputation focuses on replacing these missing values with plausible estimates.

2.1.3.2 Univariate Imputation

Univariate imputation replaces missing values in a variable by estimating the missing values based solely on the values of that same variable [208]. It assumes that the missing values are missing at random, which may not always be the case in practice. Additionally, univariate imputation may introduce bias into the analysis if the missing values are not actually missing at random [204].

There are several techniques for univariate imputation, including mean imputation, median imputation, mode imputation, and regression imputation [148]. Mean imputation replaces missing values with the mean of the non-missing values in the same variable. Median imputation replaces missing values with the median of the non-missing values, and mode imputation replaces missing values with the mode (i.e., the most common value) of the non-missing values.

Moreover, regression imputation [290] is a more sophisticated technique of univarite imputation that involves using a regression model to estimate the missing values based on the relationships between the variable with missing values and other variables in the dataset. Relevant variables may predict the missing data pattern. For example, suppose that men are more prone than women to skip certain questions; in this case, gender becomes a predicting factor for missing data. It's important that the variables in question show a moderate or stronger correlation with the variable that has missing values [178].

Traditional methods for dealing with missing data are simple to execute but come with significant limitations. For example, listwise deletion can result in a considerable amount of data loss, leading to diminished statistical power. When data is not missing completely at random (MCAR), both listwise deletion and mean imputation can introduce bias into the resulting estimates. Mean imputation and regression imputation are both susceptible to underestimating variances and covariances, as they fail to take into account the inherent uncertainty about the imputed values [3, 4]. Given these considerable drawbacks, the field of statistics has responded to this challenge by developing advanced imputation techniques that provide more robust and accurate solutions, such as multivariate imputation by chained equations [252], Bayesian multiple imputation [177], and nearest neighbors imputation [24].

- **Iterative Imputation [156, 176]:** Multiple imputation is to generate multiple imputations for missing data, as opposed to filling in a single value. And iterative imputation is a type of multiple imputation that uses a series of regression models, where each missing value is modeled conditionally upon the other variables in the data. The process is repeated multiple times, resulting in several completed datasets. It's a flexible method that can handle different variable types (e.g., continuous, binary, ordinal) and patterns of missing data.

  MICE [252]: MICE also known as Fully Conditional Specification (FCS), is a specific type of iterative imputation. It is an iterative method that imputes missing values by running a series of regression models, one for each variable with missing data. The process of MICE involves several distinct steps. Initially, a simple imputation, such as mean imputation, fills in each missing value, forming a fully populated but provisional dataset. Following this, an iterative process begins. For each variable with missing data, the preliminary imputations are reset to missing, and a regression model is established using the other variables as predictors. The missing values are then replaced using this model, leaving the observed values unaltered. This cycle is repeated for all variables, establishing a chain of imputations. This iterative step is then performed multiple times, allowing the distribution of the imputed values to gradually mirror the joint distribution of the variables. As a result, several complete datasets, each containing different imputations for the missing values, are produced. Each of these datasets can be analyzed using standard methods for complete data. Finally, the separate analysis results are consolidated into a single output, accounting for variability both within and between imputations due to sampling error and imputation uncertainty, respectively. The process is depited in Figure 2.4.

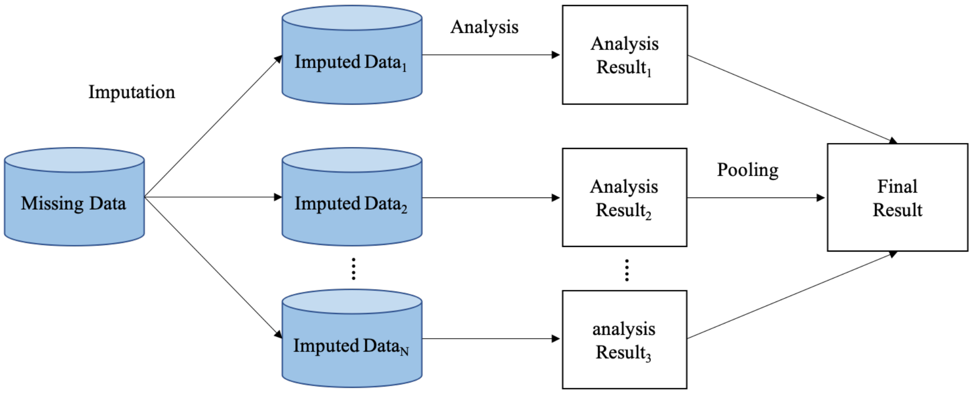


FIGURE 2.4: The process of MICE method.

- **Bayesian multiple imputation [148]:** Bayesian Multiple Imputation (BMI) presents a probabilistic approach for handling missing data by leveraging Bayesian statistical principles to substitute plausible values in place of absent ones, leading to the generation of multiple completed datasets instead of a single one. This technique builds a comprehensive joint probability model for both the observed and missing data, employing Markov Chain Monte Carlo (MCMC) methods [35] or comparable sampling procedures to draw from this model [209, 214].

  BMI encompasses a systematic procedure commencing with the development of a Bayesian model, incorporating both observed and missing data. This model encapsulates the ’analysis model’—the model designed for the completed data—and the 'missing data model'—the model depicting the probability mechanism responsible for the missing data [148, 209]. The subsequent step involves sampling from the posterior distribution of the missing data, given the observed data. These samples, rooted in the Bayesian model, replace the missing values with plausible counterparts, thereby generating multiple datasets, each complete and offering a slightly different imputation of the missing values [77, 214].

Every completed dataset is subsequently analyzed as if it were entirely complete, devoid of any missing values. The final phase amalgamates the results derived from these individual analyses into a singular outcome. This procedure accounts for the variability both within and between imputations, thereby offering a robust and comprehensive solution to the missing data challenge [209, 214].

- **Nearest Neighbors Imputation [63]:** Nearest Neighbors Imputation (NNI) is a non-parametric method that estimates missing values by finding similar observations (i.e., "neighbors") in the dataset based on available information [63]. The methodology is centered around identifying the 'nearest' instances in the multidimensional dataset space and employing their values to compensate for the missing data [5]. There exist numerous variants of NNI, exhibiting divergence primarily in their definition and computation of 'distance' or 'similarity', as well as in their aggregation of the nearest neighbors' values, such as:

  Weighted Nearest Neighbors Imputation [248]: This variant assigns weights to the nearest neighbors based on their 'distance' from the instance with the missing value. Neighbors that are closer have higher weights. The imputed value is the weighted average of the values of the nearest neighbors.

Locally Linear Embedding (LLE) Imputation [249]: LLE uses a manifold learning technique called Locally Linear Embedding to calculate the similarity between instances. It assumes that each instance lies on a locally linear manifold and imputes missing values based on this assumption.

Radius-Based Nearest Neighbors Imputation [31, 44]: In this variant, instead of specifying the number of neighbors, a 'radius' is specified. All instances within this radius are considered neighbors. The imputed value is the mean or median of the values of these neighbors.

While traditional techniques for addressing missing data are straightforward to apply, they frequently come with significant drawbacks, including data distortion, bias, and loss of crucial information [148, 215]. In light of these limitations, more advanced strategies, such as univariate imputation, are typically suggested to better manage missing data.

Despite approaches, ranging from iterative imputation, BMI, to NNI, offer ability to deal with more complex missing data patterns and relationships among variables. However, they also come with their set of limitations. For instance, iterative imputation relies relies heavily on the assumption that the missing data are MAR. If this assumption is violated, the imputed values could be biased. BMI require the specification of a joint model for the observed and missing data, which can be challenging especially in high-dimensional datasets. NNI, on the other hand, can be computationally intensive with large datasets and its performance can be sensitive to the choice of parameters. Given these limitations, machine learning based methods are proposed for imputation.

2.1.3.3 Machine Learning-Based Imputation Techniques

Machine learning-based imputation techniques have emerged as effective strategies to deal with missing data. These methods are capable of capturing complex, non-linear relationships and interactions between variables, and can thus provide a more accurate imputation of missing values compared to traditional statistical methods [120]. The mechanisms of these techniques share a common principle: the missing value of a variable is imputed by considering the observed values of other variables. Specifically, a variable presenting missing data is designated as the target variable, while the remaining variables function as input features. In this section, I will delve into three primary machine learning-based imputation techniques: Decision Trees-Based Imputation, Support Vector Machine (SVM)-Based Imputation, and Neural Networks for Data Imputation.

- **Decision Trees-Based Imputation [116, 240, 287]:** Decision tree-based imputation methods, such as Random Forests [240], XGBoost [287], and LightGBM [116], has been highlighted for its ability to capture non-linear relationships and interactions between variables, thereby providing an effective imputation strategy. The main advantage is its ability to handle both continuous and categorical data, as well as its robustness to outliers. However, the method may be computationally expensive for large datasets with high-dimensional feature space.
- **Support Vector Machine (SVM)-Based Imputation:** SVMs function [90] by mapping the input space (the features) into a higher-dimensional space where a hyperplane can be used to perform classification or regression tasks. Therefore, SVMs for imputation is able to solve high-dimensional data space and different data types of missing values [27].
- **Neural Network-Based Imputation [34]:** The effectiveness of decision trees and SVMs hinges significantly on the manner in which data is represented to them [98]. Nevertheless, the creation of such feature sets necessitates meticulous feature engineering and, crucially, extensive domain expertise [94]. An

innovative solution to circumvent this challenge involves deploying machine learning models themselves to identify and distil high-level, abstract features directly from unprocessed data [23]. Deep learning contributes substantially to this solution by generating a multi-tiered representation framework, which progressively transforms the data from one level of abstraction to a higher one [139]. This approach towards abstract data representation holds considerable promise in reconstituting meaningful features, thereby potentially mitigating the need for extensive domain knowledge and feature engineering [81]. There are several types of deep learning models that can be used for imputation:

Autoencoders [19]: Autoencoders are a type of neural network that are trained to reproduce their input. They consist of an encoder, which compresses the input into a lower-dimensional representation, and a decoder, which reconstructs the input from this representation. In the context of imputation, an autoencoder can be trained on the observed data, and then used to fill in missing values based on the learned representations.

Generative Adversarial Networks (GANs) [130]: GANs consist of two networks: a generator, which produces synthetic data, and a discriminator, which tries to distinguish between real and synthetic data. For imputation, a GAN can be trained to generate plausible values for missing data based on the observed data (Shown in Figure 2.5).

Recurrent Neural Networks (RNNs) [129]: RNNs are particularly effective for sequential data, as they can capture temporal dependencies. For time-series data with missing values, an RNN can be used to predict missing values based on the observed temporal patterns.

Deep learning-based imputation methods can provide highly accurate results, particularly for complex, high-dimensional data. However, they can also be computationally intensive and may require careful tuning of model

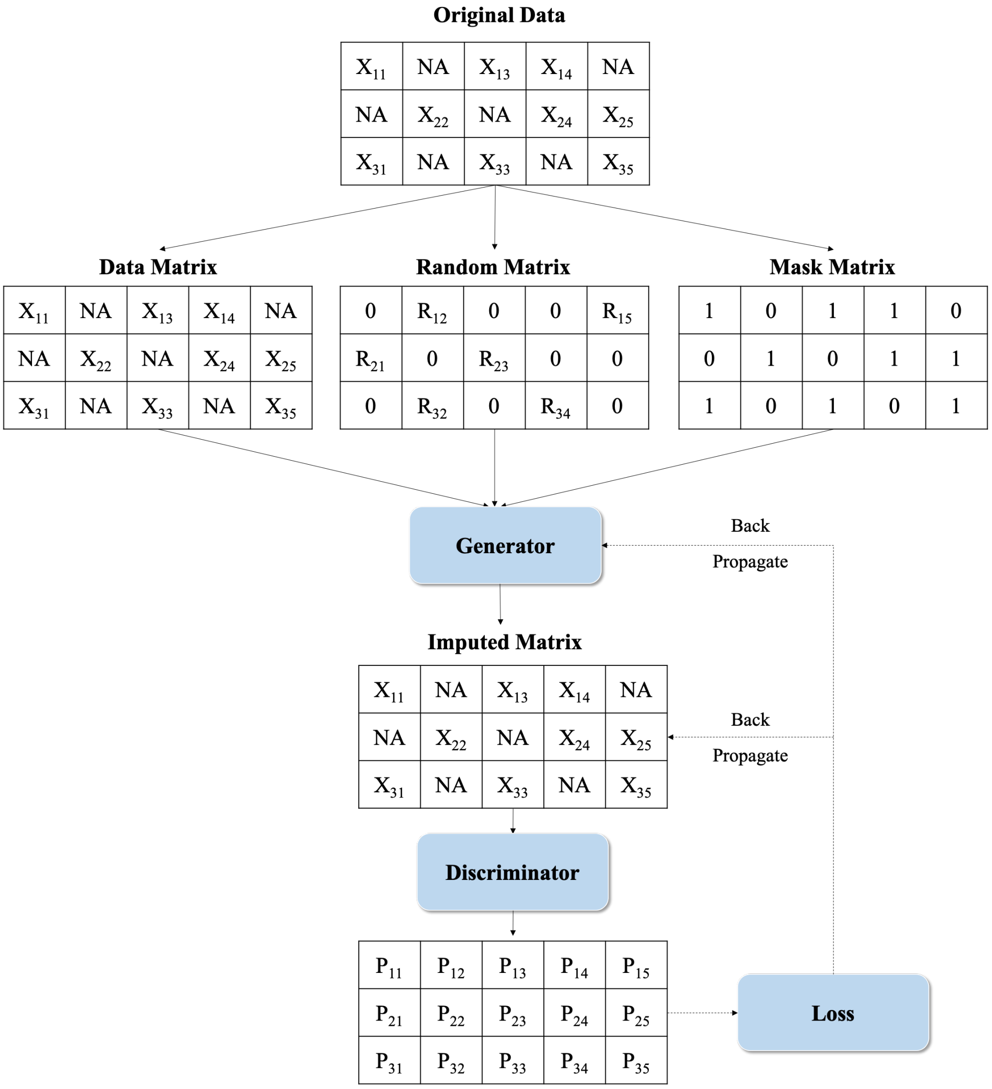

Original Data
X_{11} NA X_{13} X_{14} NA
NA X_{22} NA X_{24} X_{25}
X_{31} NA X_{33} NA X_{35}
Data Matrix
X_{11} NA X_{13} X_{14} NA
NA X_{22} NA X_{24} X_{25}
X_{31} NA X_{33} NA X_{35}
Random Matrix
0 R_{12} 0 0 R_{15}
R_{21} 0 R_{23} 0 0
0 R_{32} 0 R_{34} 0
Mask Matrix
1 0 1 1 0
0 1 0 1 1
1 0 1 0 1
Generator
Back
Propagate
Imputed Matrix
X_{11} NA X_{13} X_{14} NA
NA X_{22} NA X_{24} X_{25}
X_{31} NA X_{33} NA X_{35}
Back
Propagate
Discriminator
P_{11} P_{12} P_{13} P_{14} P_{15}
P_{21} P_{22} P_{23} P_{24} P_{25}
P_{31} P_{32} P_{33} P_{34} P_{35}
Loss

FIGURE 2.5: The process of GAN method.

parameters. Additionally, they typically require larger amounts of data to train effectively compared to traditional imputation methods.

Currently, a majority of prevalent methods necessitate the implementation of data assumptions during the pre-processing stage, followed by the execution of predictive learning. This protocol introduces an inherent bias and precludes the possibility of joint training, as the pre-processing component remains unoptimized by the final predictive loss [260]. A further limitation of these current methodologies is their predominant focus on feature-level information rather than instance-level attributes [118]. This approach tends to overlook vital information. For instance, indicators of Intensive Care Unit (ICU) admission play a significant role in mortality prediction. However, due to the relatively low incidence of such patient data, this critical information is often neglected during data pre-processing, resulting in the omission of potentially significant indicators [262]. Thus, a method that enables direct predictive learning without the need for extensive data pre-processing can have a profound impact [96]. In the medical field specifically, the ability to identify critical information while making predictions is crucial. An ideal predictive learning method should possess the capability to automatically extract and focus on the critical features within the data [38]. Moreover, the ability to make decisions based on subtle features is also important in the medical domain. Some medical conditions or diseases may exhibit subtle signs or patterns that are not easily noticeable [174]. A robust predictive learning method should be able to capture these subtle features and incorporate them into the decision-making process [149]. This would enhance the accuracy and effectiveness of predictions, potentially leading to better diagnosis, treatment, and patient outcomes. Thus, the development of computational models capable of extracting key information from extensive, often incomplete or low-quality data, in order to deliver reliable diagnostic results, has become a subject of widespread interest [263]. This thesis is undertaken within the expanding domain of this research field, with a commitment to developing a methodology capable of making predictions based solely on existing data, eliminating the necessity for pre-processing. Furthermore, it is designed to simultaneously tackle a range of challenges intrinsic to low-quality data, encompassing missing data, label noise, imbalanced data, and an excessive number of features.

Crucially, the proposed approach can capture individual-level information that significantly impacts final decision-making, thereby assisting physicians in achieving precision medicine while implementing personalized predictive modeling [263]. In this context, my research attention turns towards weakly supervised learning (WSL), and more specifically, the representative method of multi-instance learning [68]. When integrated with neural networks, the excellent scalability, compatibility, and flexibility of multi-instance learning offer distinct advantages for predictive learning, especially when dealing with real-world medical data [262].

2.2 Weakly Supervised Learning

Weakly Supervised Learning (WSL) is an emerging field that aims to strike a balance between fully supervised and unsupervised learning paradigms by utilizing less precise, indirect, or noisy labels to train machine learning models. This form of learning addresses situations where acquiring vast amounts of accurately labeled data is practically challenging or impossible [294]. It is typically categorized into three distinct types based on the nature of the weak supervision: incomplete supervision, inexact supervision, and inaccurate supervision [179, 302].

Incomplete Supervision: Incomplete supervision involves a learning scenario where only a subset of instances are labeled. Semi-supervised learning (SSL) and Positive-Unlabeled (PU) learning are the two commonly recognized techniques within this category. SSL employs a blend of a few labeled instances with a significant volume of unlabeled data during the training process [40]. SSL holds considerable promise and has substantial applications within the realm of medicine, particularly in the context of Electronic Health Records (EHR) data analysis. EHRs frequently consist of unstructured free-text clinical notes that necessitate meticulous analysis and interpretation. However, labeling these patient records is a labor-intensive task, often requiring the expertise of clinicians, leading to a limited pool of labeled data. SSL can be used to augment the small amount of labeled data (annotated by clinicians) with a large amount of unlabeled data, helping to develop models for predicting patient outcomes or identifying disease patterns [253]. On the other hand, PU learning contends with positive and unlabeled data, inherently assuming that the unlabeled data consists of both positive and negative instances [64]. PU learning can be particularly helpful in the context of rare diseases, where only a few positive examples might be available. In these cases, PU learning can be used to learn about the disease's characteristics and aid in its diagnosis [172].

Inaccurate Supervision: Inaccurate supervision involves situations where the labels assigned to instances could be incorrect or noisy. This category includes learning with noisy labels and learning with label noise. Both these approaches involve training the model with datasets containing mislabeled instances, and the task is to develop a model robust enough to handle the noise [179]. In the medical field, inaccurate supervision holds substantial relevance and offers unique advantages. For example, clinical datasets often contain erroneous or conflicting labels due to human error, ambiguous symptoms, or the subjective nature of certain diagnoses. As an example, consider the case of mental health disorders, where diagnoses often rely on subjective interpretation of symptoms. Here, learning with noisy labels can help in developing models that are more robust to such inconsistencies [237].

Inexact Supervision: This subclass focuses on scenarios where labels exist at different levels of abstraction rather than corresponding precisely to the instance. MIL is included in inexact supervision and learning from label proportions. In MIL, a label is attached to a bag (group) of instances rather than to each specific instance, demanding the model to discover the informative instances within the bag. Meanwhile, learning from label proportions deals with scenarios where only the proportion of each class within a group of instances is available.

2.3 Multi-Instance Learning

Multi-instance learning (MIL) represents a unique approach to machine learning, wherein the fundamental unit of training data is not a single instance, but rather a set, or "bag," of instances. Instead of individual instance labels, the entire bag is assigned a supervisory label. This form of weak supervision holds particular promise for scenarios where labeling individual instances may be impractical, costly, or impossible [294].

The fundamental assumption in MIL is that a bag is labeled positive if and only if at least one of its constituent instances is positive, and is otherwise labeled negative [59]. For instance, in a medical imaging context, a bag might represent a set of image patches derived from a patient's scan, and the bag would be labeled positive if it contains an image patch indicative of disease [275]. When a new bag of unlabeled instances is encountered, a trained MIL model is capable of predicting the bag's label based on the learned patterns from the training phase [38]. Figure 2.6 presents an illustrative example of Multi-instance Learning (MIL). The corresponding process can be formulated as follows.

Let X denote a feature space and Y a set of binary labels. In classical supervised learning, the goal is to learn a function $\Phi : \mathrm{X} \to \mathrm{Y}$ from the given data set $\{(\chi_1, \gamma_1), (\chi_2, \gamma_2), \ldots, (\chi_m, \gamma_m)\}$, where each $\chi_i \in \mathrm{X}$ and $\gamma_i \in \mathrm{Y}$ ($i = 1, 2, \ldots, m$). Each χ_i represents an instance labeled by a known class $\gamma_i = \{0, 1\}$.

Conversely, in the MIL problem, each $\chi_i \in \mathrm{X}$ is considered a 'bag' of instances $\chi_i = \{x_{i,1}, x_{i,2}, \ldots, x_{i,n}\}$ where each $x_{i,j} \in x_i$ ($j = 1, 2, \ldots, n$). In this case, only the bag label γ_i is provided. It should be noted that the number of instances in different bags may vary [299].

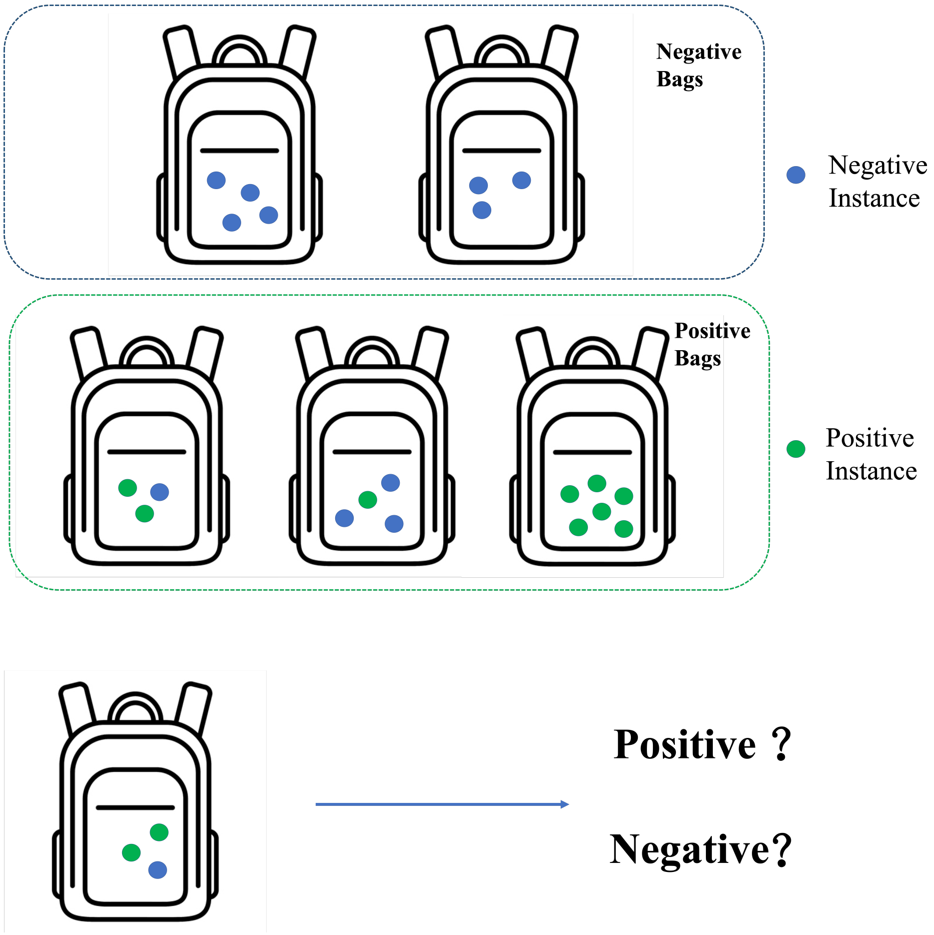


FIGURE 2.6: Demonstration of bags and instances. A bag receives a positive label if it contains at least one positive instance within it. Conversely, if all instances within a bag are negative, the bag itself is labeled as negative. The process of training and learning in MIL equips the model with the ability to predict the class of an unlabeled bag effectively.

The standard assumption made in the MIL problem is that a bag is labeled positive if and only if at least one instance contained within it is predicted as positive. This assumption can be mathematically formulated as follows:

$$\gamma_i = \begin{cases} 1, & \text{if } \exists x_{i,j} \in \chi_i : \Phi\left(x_{i,j}\right) = 1 \\ 0, & \text{otherwise} \end{cases} \tag{2.5}$$

Early MIL methods are fundamentally rooted in the aforementioned assumption [8, 59, 286]. These methods posit that the bag label is determined by one or several distinctive instances. However, these approaches do not take into account the correlations among instances and the distribution of instances within the bag.

To address this limitation, a more generalized assumption has been proposed [260]:

$$\gamma_i = \psi\left(\theta_{x_{i,j} \in \chi_i} \sigma\left(x_{i,j}\right)\right) \tag{2.6}$$

where σ denotes a transformation applied to the attributes of the instances. The MIL methods can be partitioned into two primary categories based on the distinct strategies for choosing θ and ψ [118] (Shown in Figure 2.7). One approach, instance-level strategy, attempts to infer instance labels indirectly and then use traditional supervised learning techniques for analysis [8]. Another approach, bag-level strategy, modifies traditional learning algorithms to work directly with bags, bypassing the need for instance labels entirely [297].

- **Instance-Level Strategies:** θ is a scoring function to obtain the positive probabilities of each containing instance. ψ is MIL pooling over instances to return the bag probability. These methods focus on the individual instances within each bag. They operate by attempting to assign labels to the instances within a bag, effectively transforming the multi-instance problem into a single-instance (or traditional) supervised learning problem [7]. An instance-level strategy often involves two steps: first, assigning pseudo-labels to the instances based on the bag label, and second, learning a traditional classifier based on these pseudo-labeled instances. Popular instance-level methods include the

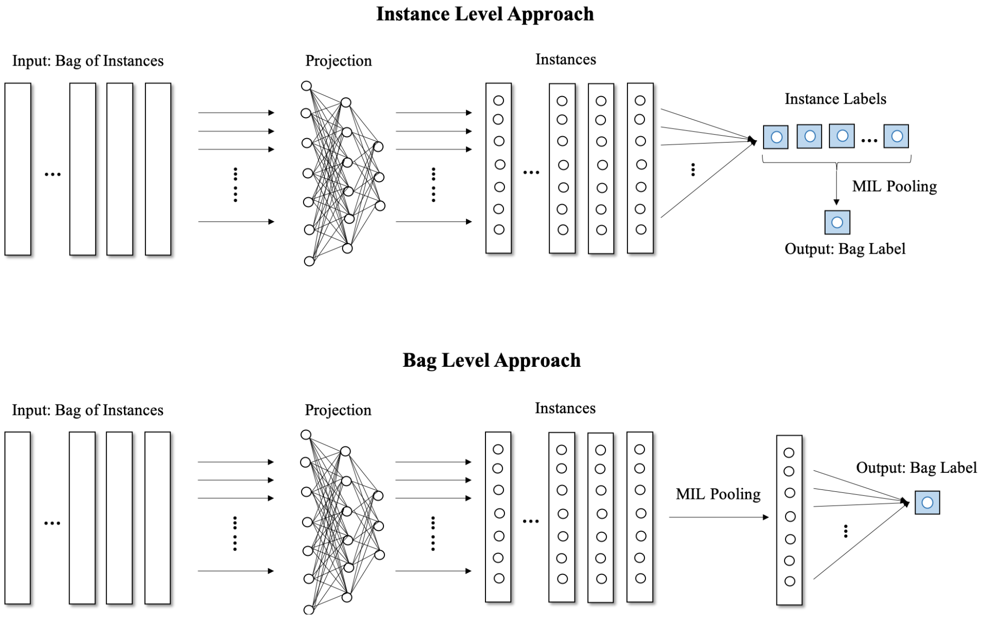


FIGURE 2.7: Difference between instance-level approach and bag-level approach in MIL.

diverse density (DD) approach [166] and the Expectation-Maximization (EM) algorithm [285].

- **Bag-Level Strategies:** θ is MIL pooling over instance attributes to obtain the low-dimensional bag embedding, that is further processed by a scoring function ψ to return the bag probability. These methods, conversely, consider the bag as a whole, rather than focusing on individual instances. They aim to learn a bag-level classifier that directly predicts the bag's label without assigning labels to individual instances [38]. This type of approach often treats a bag as an ordered or unordered set and learns from the set distribution. Methods in this category include the kernel-based methods that employ set kernels to compute the similarity between bags [75], and the neural network-based methods that use architectures like convolutional neural networks (CNNs) or recurrent neural networks (RNNs) to capture the bag's characteristics [118].

In the realm of MIL, both bag-level and instance-level approaches have their unique strengths and weaknesses. Bag-level strategies, as demonstrated by Wang et al. [259], tend to outperform in prediction accuracy and running time. This is primarily due to the fact that they consider the collective information of instances within a bag, thereby providing a more comprehensive view of the data. However, this approach may overlook the individual characteristics of instances, which could be crucial in certain applications.

On the other hand, instance-level approaches, while they may suffer from insufficient training due to unknown instance labels and introduce additional error, offer the advantage of efficiency and flexibility. They are capable of identifying key instances within a bag, which could be pivotal in determining the overall label of the bag. This is particularly useful in scenarios where a single instance can significantly influence the bag's label.

In this study, I aim to harness the strengths of both these approaches. A MIL pooling module over instances and their attributes is proposed to integrate, which allows to capture both the collective and individual characteristics of instances. This fusion of bag-level and instance-level strategies enables us to fully uncover the intricate instance-to-bag relationship. As a result, the method is able to achieve performance enhancement, as corroborated by Yan et al. [277]. This hybrid approach provides a more robust and comprehensive framework for MIL, thereby improving the effectiveness and reliability of the learning process.

Moreover, while Multi-Instance Learning (MIL) is predominantly utilized for assigning a single label to an instance or bag, its potential extends beyond this application. The concept of assigning multiple labels to bags is particularly pertinent as bags can encompass instances representing diverse concepts. This multi-label MIL notion has been the subject of several studies [106, 283, 299].

In addition to classification, MIL can also be employed for regression problems by substituting the bag-level classifier with a regressor [6, 264, 265]. Some methodologies have proposed ranking bags instead of assigning a class tag or score. This task differs from regression as the objective is not to achieve accurate real-value labels, but to compare predicted scores for sorting purposes. Ranking can be executed at either the bag-level [26] or the instance-level [114].

MIL can also be adapted for clustering tasks, which include searching for clusters or structures within a set of unlabeled bags. In certain scenarios, clustering is performed in bag spaces using conventional algorithms and set-based distance measures. For instance, the algorithm in [284] identified the most pertinent instances of each bag and executed maximum margin clustering on those instances. Alternatively, clustering can be carried out at the instance-level. For example, Wang et al. [258] performed instance clustering for dictionary learning, and Tang et al. [241] utilized instance clustering to stabilize the process of weakly-supervised object detection. The proposed instance embedding regularization method focuses on feature learning and could prove beneficial for multi-label MIL, multi-instance regression, multi-instance ranking, and multi-instance clustering.

2.3.1 Classical MIL Methods

MIL has been the subject of extensive research, with scholars making significant strides in the field. Numerous machine learning models have been developed to address the MIL problem. For instance, linear Support Vector Machines (SVM) [195] have been employed in MI-SVM and mi-SVM for bag-level and instance-level classification, respectively. The Citation-kNN model [256] adopts a lazy approach to the MIL problem, utilizing a k-nearest neighbor (kNN) classifier and various distance metrics.

Fretcit-kNN [295] applies the minimal Hausdorff distance between frequent term sets and uses both the references and citers of an unseen bag to determine its label in web recommendation tasks. G3P-MI [281] tackles MIL from a unique perspective, leveraging grammar-guided genetic programming. MI-Kernel [76] integrates kernel methods into MIL, while EM-DD combines the expectation-maximization algorithm with the diverse density (DD) algorithm for MIL [285].

The Multi-instance Fisher Vector (miFV) [213] is a representative algorithm for solving MIL problems from the perspective of embedded space. It maps instance features into a high-dimensional space through a pretrained Gaussian model and Fisher Vector coding. In addition to Fisher Vector coding, the Vector of Locally Aggregated Descriptors (VLAD) is also used in [213] for MIL, in a method known as miVLAD. The Multi-instance Dissimilarity (MInD) method [48] employs bag similarities for bag classification.

Zhou et al. [296] introduced a pioneering work in Multi-Instance Learning (MIL), proposing a non-independent and identically distributed method to handle instances from the bag and capitalize on the relationships between instances. Their mi-Graph algorithm offers improved MIL performance. Inspired by this work, the intention in my research is to enhance the quality of instance embedding by leveraging these instance relationships. In contrast to mi-Graph, which directly compares bags without considering instance embedding, instance embedding is proposed here as the backbone of MIL, further distinguishing between positive and negative instances within the bag. Moreover, while mi-Graph employs a non-deep learning method, the approaches proposed in this thesis are grounded in deep neural networks.

2.3.2 Multi-Instance Neural Network

Deep neural networks have emerged as powerful tools for tackling a wide array of machine learning challenges. Deep Belief Networks (DBN) [107] employ unsupervised pre-training and utilize a fixed-length vector for feature learning and classification. Deep Convolutional Neural Networks (CNN) [79, 104, 197, 238] and Vision Transformers (ViT) [128, 188, 202] process images as input and have become the standard for image recognition tasks. Deep Recurrent Neural Networks (RNN) [87], Long Short Term Memory (LSTM) networks [86] and transformers [255] handle sequential data such as text and speech, excelling in sequential prediction tasks.

In classical MIL problems, it's generally presumed that instances are represented by features that require no additional processing. However, for tasks such as image or text analysis, the necessity for further feature extraction steps becomes apparent. As a result, the idea of utilizing neural networks to parameterize all computational transformations becomes highly appealing. This methodology provides significant flexibility and allows for end-to-end training via backpropagation [118].

Nonetheless, the training of these deep networks requires a substantial amount of fully labeled data, implying that each instance must have a label. In the MIL context, only bag labels are accessible. Furthermore, MI data exhibits a complex structure, consisting of a set of instances with varying instance counts across different bags. These intricacies pose challenges when attempting to tackle the MIL problem using conventional neural networks [259].

Before the raising of deep learning, numerous research initiatives sought to address the MIL problem through the use of neural networks. Ramon and Raedt [201] pioneered the concept of a Multi-Instance Neural Network (MINN), which estimates instance probabilities prior to the final layer and computes bag probability using a convex max operator, also known as log-sum-exp. This network can be trained via back-propagation. In a similar vein, Zhang and Zhou [298] proposed a multi-instance network that calculates bag probability by directly taking the maximum of instance probabilities.

The underlying function of MIL provides flexibility, allowing us to model any transformation and score function, provided they adhere to the permutation-invariant property. Consequently, a class of transformations is parameterized through the neural network. Let X represent a bag of M instances. The transformer φ_τ, where τ are parameters, transforms instances to the embedding space with K dimensions, such that $v_{m,K} = \varphi_\tau\left(x_m\right)$ where $m \in M$. Then the bag probability of x_m is then determined by the transformation $\theta_\omega : \eta_{\phi_{k \in K}}\left(v_{m,k}\right) \rightarrow [0, 1]$.

In the case of employing the bag-level MIL pooling approach, θ_ω is an injective function, or alternatively, it is parameterized by the neural networks with parameters ω. f the trainable MIL pooling methods are utilized, ϕ also become parameters. This approach allows us to fully exploit the flexibility of neural networks, enabling us to model complex transformations and score functions that can capture the intricate structure of MIL data. This, in turn, can lead to improved performance in a wide range of MIL tasks.

2.3.2.1 Multi-Instance Pooling

MINN is proposed to endow the MIL methods with more flexibility, that it parameterizes all computational process and is trained end-to-end by back-propagation [201]. The key step in MINN is MIL pooling that it is used to aggregate the information contained within a bag into a single representation. This is necessary because traditional machine learning algorithms are designed to operate on individual instances, not bags of instances. The pooling operation allows these algorithms to be applied to MIL problems.

There exists a diverse array of multi-instance pooling methods, which can primarily be categorized into two distinct groups: trainable and non-trainable [262].

Trainable pooling methods are designed with parameters that can be optimized during the learning process. These methods are adaptive in nature, adjusting their behavior based on the data they encounter [277]. This adaptability can lead to improved performance, especially in complex tasks where the optimal pooling strategy may not be obvious or static [277]. Examples of trainable pooling methods include attention-based pooling [13] and learned pooling, where the pooling operation is guided by a secondary learning algorithm [81].

On the other hand, non-trainable pooling methods operate with fixed rules and do not adjust their behavior based on the data [139]. These methods are simpler and more computationally efficient than their trainable counterparts, making them suitable for tasks where computational resources are limited or the optimal pooling strategy is known a priori [32]. Examples of non-trainable pooling methods include the aforementioned max pooling, mean pooling, and log-sum-exp pooling [32].

Each category of multi-instance pooling methods has its own strengths and weaknesses, and the choice between them should be guided by the specific requirements of the task at hand [259]. In the following sections, I will delve deeper into the specifics of both trainable and non-trainable pooling methods, providing a comprehensive understanding of their underlying mechanisms, applications, and performance characteristics [81].

CHAPTER 3

AMI-Net

Effective and efficient analysis of clinical records is a crucial task in the medical field. These records, which comprise varying numbers of symptoms per patient, can be conceptualized as a ’bag’ of instances. The challenge lies in identifying informative symptoms (instances) and associating them with one or more diseases for accurate medical diagnosis. Conventional approaches often represent patients as vectors in a feature space and apply classifiers to generate diagnostic results. However, this method often grapples with issues arising from low-quality data, largely due to factors like data consistency, integrity, completeness, and accuracy.

To address these challenges, a novel method named the Attention-Based Multi-Instance Neural Network (AMI-Net) is first proposed. The model classifies single diseases based solely on valid information extracted from real-world outpatient records, taking an end-to-end approach. It inputs a bag of instances and directly outputs a bag label. An embedding layer maps instances into an embedding space, representing individual patient conditions. The model harnesses the power of a multi-head attention transformer, instance-level multi-instance pooling, and bag-level multi-instance pooling to capture instance correlations and their significance in the final classification [255].

The proposed approach is distinctive in its ability to integrate these components into a multi-instance neural network. The principal tasks of medical diagnosis from incomplete and low-quality data are addressed by mapping input instances in an embedding space, capturing instance correlations in different embedding subspaces, learning bag embedding, and selecting informative instances via an attention mechanism to obtain the bag score. This design leverages the Multi-Instance Learning (MIL) neural network for parameterization, imbuing the architecture with flexibility and simplicity [81]. Importantly, this approach does not necessitate manual data collection or screening. Instead, it automatically handles data, efficiently extracting useful information from a vast amount of low-quality data to support the final medical diagnosis.

The proposed method has been tested on two incomplete and highly imbalanced datasets, one in the Traditional Chinese Medicine (TCM) domain and the other in the Western Medicine (WM) domain. The experimental results have shown that AMI-Net significantly outperforms all baseline results [32].

3.1 Methodology

The architecture of the proposed AMI-Net is made up of multiple computational module for prediction. Here, we delve into a detailed explanation of each layer, elaborating on their functionality and contribution to the overall system.

Firstly, an embedding layer is included to map instances into a high-dimensional embedding space, a transformation that aids in the precise representation of individual patient conditions. The subsequent process of classification and disease prediction thus leverages these patient-specific embeddings.

Following the embedding layer is a multi-head attention transformer equipped with a residual connection [103]. Borrowed from transformer architectures [255], this module is designed to capture complex interactions among the instances. The residual connection assists in bypassing the transformation function, thus promoting the ease of training and mitigating issues related to vanishing gradients. The transformer allows the model to focus on different aspects of the input instances simultaneously, offering a more nuanced interpretation of the underlying medical data.

After the multi-head attention transformer, a series of instance-wise fully connected layers is applied. These layers serve to extract more intricate features from each instance, further refining the representations of the symptoms.

The architecture then incorporates an instance-level Multi-Instance Learning (MIL) pooling layer. This layer functions to aggregate the representations of the instances within each bag (i.e., patient), thereby synthesizing the individual symptom information into a unified representation.

Subsequently, a bag-level MIL pooling layer is used. This layer extends the pooling operation to bags, consolidating the bag-level representations and further emphasizing the most informative instances within each bag.

Finally, a sigmoid function is applied, serving as an activation function to produce the final output of the model. It ensures the output values lie between 0 and 1, corresponding to the probability of the presence of a particular disease.

Figure 3.1 provides a comprehensive visual overview of the AMI-Net, depicting the intricate interplay among its various components and highlighting the end-to-end nature of the architecture.

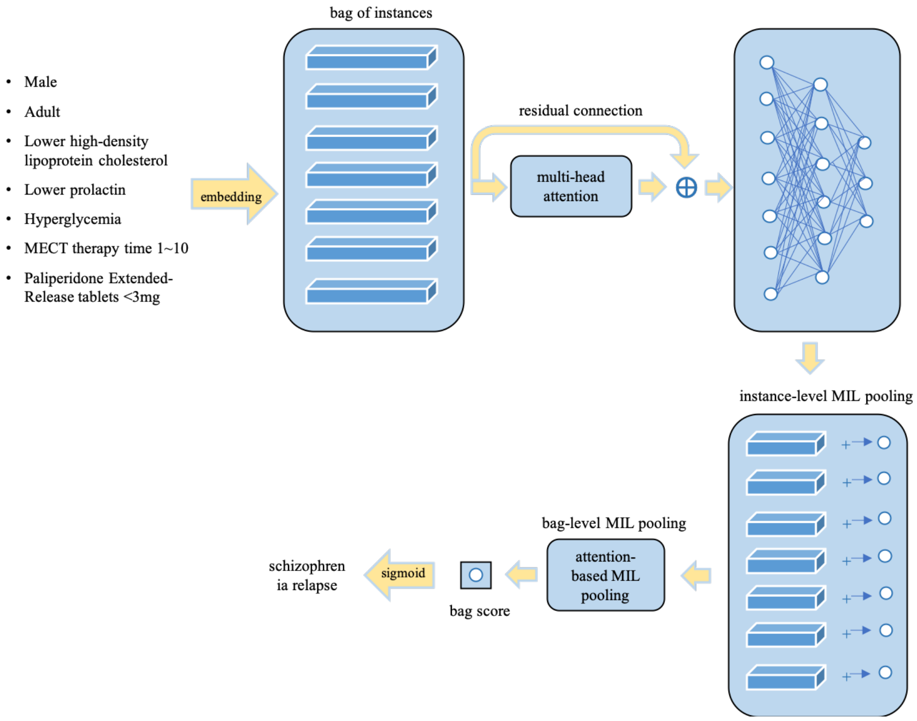


FIGURE 3.1: The overview of AMI-Net.

3.1.1 Multi-Instance Pooling (MIL Pooling)

Reflecting on the distinct attributes of both trainable and non-trainable MIL pooling methods, the proposed AMI-Net thoughtfully incorporates elements from both categories into its architecture. This strategic inclusion is guided by an understanding of the inherent strengths and limitations of each method and is designed to maximize the benefits accrued from their respective advantages.

The fusion of trainable and non-trainable MIL pooling methods in the model aims to optimize prediction capability and adaptability. This approach enables a more sophisticated analysis of instance data, effectively addressing scenarios where the optimal pooling strategy is either known or needs to be learned from the data. By striking a balance between the two, the model facilitates a comprehensive and accurate interpretation of the underlying medical data.

In the AMI-Net, inspiration has been taken from the document classification problem for the instance-level MIL pooling method. More specifically, the approach borrows from the way sentences are represented within the realm of text classification. Sum pooling, a non-trainable MIL pooling technique, is employed at the instance level.

The sum pooling technique essentially computes the sum of all instances within a bag to create a cumulative instance representation. This method is particularly effective as it allows for the aggregation of information from all instances, thereby capturing the holistic information present within each bag (i.e., patient). Thus, the sum pooling technique, formulated as below, in the instance-level MIL pooling stage, contributes to an overall richer representation of the patient's condition. This, in turn, can lead to a more accurate and robust prediction performance.

$$\forall_{m=1,2,\ldots,M} : v_m = \sum_{k=1}^{K} v_{m,k} \tag{3.1}$$

where M and K denotes the bag containing instances, and the embedding dimensions. Moreover, an attention-based MIL pooling approach is proposed for the bag-level to obtain the bag score, which is further mapped to the bag probability through a sigmoid function.

3.1.2 Attention-based MIL Pooling

The primary objective of attention-based MIL pooling mechanism is to assign a set of weights to instances within the bag. These weights, instead of being predetermined or fixed, are trained and optimized by the neural network itself during the learning process. In the proposed method, it is employed in the bag-level, which is formulated as follows:

$$v = \sum_{m=1}^{M} a_m v_m \tag{3.2}$$

where:

$$\begin{aligned} S &= W_1^T \left(\tanh\left(v_m W_2\right) \odot \operatorname{sigmoid}\left(v_m W_3\right)\right)\big) \\ a_m &= \operatorname{softmax}(S) \end{aligned} \tag{3.3}$$

where $W_1 \in \mathbb{R}^{d_{\text{model}} \times 1}$ and $W_2, W_3 \in \mathbb{R}^{d_{\text{model}} \times d_l}$ are parameters, and $\odot$ is the element-wise multiplication.

Considering the limitations of the $\tanh$ function, particularly its constrained capacity to capture complex relationships and express non-linearity, an additional operation is introduced in the model. Specifically, a $\operatorname{sigmoid}$ based function is applied in an element-wise multiplication subsequent to the $\tanh$ function. This operation forms part of what is commonly referred to as the gated mechanism, as discussed in the work by Dauphin et al. [57]. The gated mechanism enhances the model's capability to learn complex relations, allowing for a richer expression of non-linearities in the data.

Moreover, the attention mechanism effectively guides the network to focus predominantly on instances that are most likely to be labeled as positive, as outlined by Hu et al. [112]. This strategic focus not only enhances the model's performance but also its interpretability. It equips the model with the ability to sift through a vast amount of 'dirty' or low-quality data and zero in on the key information. Such an approach aligns closely with real-world medical diagnostic processes, wherein among a multitude of symptoms and data points, medical professionals must identify the most critical ones to arrive at an accurate diagnosis. In this sense, the attention mechanism allows the proposed model to mimic the discerning eye of a physician, thereby enhancing its efficacy in medical predictive learning.

3.1.3 Multi-Head Attention

In this method, the integration of the multi-head attention [255] on the AMI-Net is proposed. The principal aim of this integration is to capture the intricate intra-relationships among instances within different embedding subspaces.

This strategy is particularly suited to the medical domain, where symptoms often interrelate across various body parts or organs. In this context, each organ or body part can be considered as a distinct embedding subspace. Therefore, the multi-head attention mechanism can effectively uncover correlations and links among symptoms that standard linear methods might overlook.

Moreover, the multi-head attention mechanism can help bridge the gap between standard and non-standard expressions of symptoms, thereby enhancing the model's robustness when dealing with low-quality or inconsistent data.

The attention mechanism within the transformer model operates by taking a query (Q) and a set of key-value (K, V) pairs as input and generating a weighted sum of the values as output. The weights assigned to the values are computed based on the query and the corresponding key, utilizing a cosine similarity-based function.

In the proposed methodology, significant emphasis is placed on exploring the correlations among instances. Hence, in the context of the model, the query, key, and value are all derived from the instances themselves. In terms of its practical implementation, the multi-head attention mechanism consists of two main computational components: scaled dot-product attention and multi-head attention transformation.

The detailed architecture of the transformer, including its integration of the multi-head attention mechanism, is illustrated in Figure 3.2. This visual representation further clarifies the functioning of the transformer and its role in the overall structure of the model.

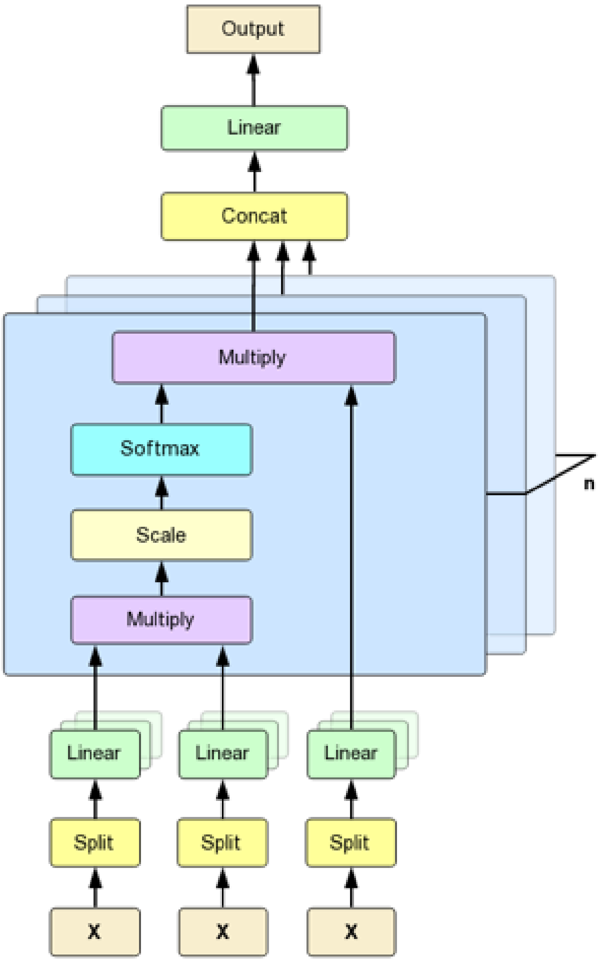


FIGURE 3.2: The architecture of multi-head attention.

Scaled dot-product attention Initially, cosine similarity is calculated within the subspace for the instances themselves. This is followed by the application of the softmax function to derive the final weights vector, which encapsulates the similarities and correlations among instances. Given that a large instance dimension, d_i could potentially result in extremely small gradients from the softmax function, a scaling factor of $\frac{1}{\sqrt{d_i}}$ is employed. The final output is computed as follows:

$$\text{Similarity}(b, c) = \frac{b \cdot c}{||b|| \, ||c||} = bc^T \tag{3.4}$$

$$\text{Att}(X, X, X) = \text{softmax}\left(\frac{\text{similarity}(X, X)}{\sqrt{d_i}}\right) X \tag{3.5}$$

where $\cdot$ is dot-product function and X denotes a bag of instances.

Multi-head transformation This process fractionates the instance dimensions into several subspaces, executing scaled dot-product attention independently on each subspace. This parallel operation allows for the capture of instance correlations across diverse subspaces. The resulting outputs from each subspace are subsequently concatenated to form the final output. Throughout this procedure, linear transformations are intermittently applied to facilitate the process. The entire procedure can be mathematically formulated as follows:

$$\text{MultiHead}\,(X, X, X) = \text{ Concat }\,(\text{ head }_1, \ldots, \text{ head }_n)\, W^m \tag{3.6}$$

$$\text{head}_i = \text{Att}\left(XW_i^1, XW_i^2, XW_i^3\right) \tag{3.7}$$

where $W_i^1, W_i^2, W_i^3 \in \mathbb{R}^{d_{\text{model}} \times d_k}$, $W^m \in \mathbb{R}^{hd_k \times d_{model}}$, h denotes the number of heads and $head_i$ denotes the i^{th} subspace.

3.2 Experiments

In the experimental evaluation, the proposed AMI-Net method was applied to two real-world medical datasets that were well-suited for the approach. The first dataset originated from the domain of Traditional Chinese Medicine (TCM), while the second was derived from the Western Medicine (WM) domain. Both datasets were used for diagnostic purposes. Representative examples from these two datasets are presented in Table 3.1.

TABLE 3.1: Examples of TCM and WM datasets

Dataset	Features	Diagnosis
TCM	Urine color yellow, Sweat, Pruritus, Coldness of extremities, Perspiration	Meridian obstruction
	Dark red tongue, Palpitation	Not meridian obstruction
WM	Personal income 3000 5000, Unmarried, LOS<10 days, MECT<=1, Onset age<17, Total course<1095 days, Lorazepam tablets=0.5mg	Schizophrenia relapse
	Personal income 1000 3000, Married, High levels of prolactin, hyperglycemia, High levels of corticotrophin, LOS 25 49 days, MECT 1 10, Onset age 1-10, Risperidone<=1mg, Total course 1095 5840 days, Haloperidol injection 5mg	Not relapse

3.2.1 Datasets

Traditional Chinese Medicine: Traditional Chinese Medicine (TCM): The TCM dataset was amassed from the medical records of diabetic patients at a Chinese Medical Hospital in Beijing. This dataset has been previously analyzed to discern crucial herb-herb interactions [194] and symptom-herb patterns [45]. The dataset comprises 1,617 outpatient records, each exhibiting one of 186 distinct symptoms. The number of symptoms per patient varies, ranging from 1 to 17. However, it's important to note that the symptom expressions are not standardized and lack consistency. For instance, the symptom 'sweat' is expressed in various ways across different records.

The binary classification task in this study focuses on determining whether a patient has meridian obstruction, a syndrome specific to Traditional Chinese Medicine (TCM). Out of all the patients in the dataset, 1,436 are labeled as negative (no meridian obstruction), while 181 are labeled as positive (presence of meridian obstruction). This results in a highly imbalanced dataset with a positive rate of 0.112. A significant challenge in this dataset is the presence of numerous missing values. This is primarily due to the difficulties clinicians face in completing patient examinations, which can be attributed to a lack of patient compliance and the non-standardization of TCM information collection.

Western Medicine (WM) Dataset: The Western Medicine (WM) dataset was provided by Medicinovo Inc. in real-world medical studies. This dataset comprises 3,927 inpatient records of schizophrenic patients who underwent modified electro-convulsive therapy (MECT) and showed improvement upon discharge. The objective of the model built on this dataset is to predict the likelihood of a schizophrenia relapse within three months. This prediction is based on 88 physical and clinical features, including marital status, employment status, high levels of prolactin, the number of MECT sessions (ranging from 1 to 10), and administration of 5mg haloperidol injections. For each patient, there are at least 5 and at most 21 features present, representing the individual patient's condition. Similar to the TCM dataset, the WM dataset is also highly imbalanced, with a positive label rate of only 0.057.

3.2.2 Experimental Setup

For both the TCM and WM datasets, preprocessing and parameter setting steps were implemented. Each input record was padded to the maximum size to ensure uniformity. The number of embedding dimensions was set at 128, which closely aligns with the number of human organs as per reference [50].

In the subsequent multi-head attention transformer, the number of heads was configured to be four. This setting was chosen to balance computational efficiency and the capacity to capture various feature interactions. For the instance-wise fully connected layers, hidden sizes of 64 and 32 were selected, respectively, to ensure a sufficient level of model complexity while avoiding overfitting.

For the final loss calculation, cross-entropy was employed as the loss function for binary classification tasks. The Adam optimizer [131] was used to minimize this loss over the training data. The hyperparameters for the Adam optimizer were set as follows: the learning rate was set at 0.01, the momentum parameters β_1 and β_2 were set at 0.9 and 0.98, respectively, and ε was set at $1e^{-8}$. These settings are generally accepted as effective starting points for many optimization tasks.

To evaluate and compare the model's performance, the binary threshold was set at 0.5, and several evaluation metrics were used, including AUC, Accuracy, Precision, Recall, and F1-score. These metrics provide a comprehensive assessment of the model's performance from different perspectives.

During the training process, the number of epochs was set at 1000. Early stopping based on the AUC score over the validation dataset in cross-validation was employed to prevent overfitting and select the best model. This strategy allows the training to be stopped as soon as the model's performance on the validation set starts to deteriorate.

To ensure a fair comparison of the model's performance, all experiments were conducted using 10-fold cross-validation with five repetitions. This rigorous evaluation methodology ensures that the results are robust and reliable, reducing the likelihood of overfitting and providing a more accurate estimate of the model's performance on unseen data.

3.2.3 Comparison with Baseline Models

The task at hand is not only viewed as a MIL problem, but also as a traditional binary classification problem. This perspective involves working with the dataset in a one-hot format, which is a common representation for categorical data. In this context, the aim is to learn a transformation function $g : X \to [0, 1]$, where $X = \{(\lambda_i, o_i)\}_{i=1}^{|X|}$ is a set of (feature, value) pairs as described in Grangier et al. [85]. The values in these pairs are binary, representing the presence or absence of a feature.

In cases where a value is missing, a common approach is to impute a 0, symbolizing the unknown condition. This method of handling missing data allows us to maintain the binary nature of the dataset while acknowledging the lack of information.

To evaluate the effectiveness of the proposed method, a comparison was made against several baseline models. The first four of these models were constructed using datasets transformed into the one-hot format. This transformation process involves converting categorical data into a format that can be more easily processed by machine learning algorithms. The purpose of comparing the proposed method against these baseline models was to demonstrate its superior performance in handling the binary classification task.

- **Logistic Regression (LR) [110, 274]:** LR is a classical linear model that has seen extensive use in various applications. These include binary classification tasks, the selection of risk factors, and the development of risk assessment scales. Its widespread adoption can be attributed to its simplicity, interpretability, and effectiveness in modeling the probability of a binary outcome.
- **SVM [257]:** The Support Vector Machine (SVM) method employs a nonlinear transformation to map the input space into a higher-dimensional space. Within this transformed space, SVM constructs a series of hyperplanes to carry out regression and classification tasks. This technique enables SVM to effectively manage intricate patterns and relationships within the data. As a result, SVM serves as a potent tool for both regression and classification problems, adept at handling complex data structures.
- **Random Forest [109] and XGBoost [46]:** Random Forest and XGBoost are quintessential decision tree-based algorithms that tackle classification and regression tasks using bagging and boosting methods, respectively. These algorithms have gained considerable recognition in the medical field owing to their interpretability, swift training speed, and outstanding performance. Random Forest, a bagging method, operates by creating an ensemble of decision trees and aggregating their predictions. This approach enhances the model's stability and reduces the likelihood of overfitting. On the other hand, XGBoost, a boosting method, builds decision trees sequentially, with each new tree aiming to correct the errors made by its predecessor. This strategy results in a robust model that can capture complex patterns in the data. Both Random Forest and XGBoost are celebrated for their capacity to handle diverse data types and complexities, making them versatile tools in predictive modeling.

- **mi-Net [259]:** A MIL neural network using the bag-level MIL pooling approach with the max operator.
- **MI-Net, MI-Net with DS and MI-Net with RC [259]:** They are all proposed by Wang et al. using the instance-level MIL pooling approaches, which have achieved state-of-art performance on several classic MIL datasets.
- **Att. Net and Gated Att. Net [118]:** Two recent state-of-art MIL neural networks, utilizing the attention based MIL pooling on instance level to capture the relations of instance attributes. This innovative approach allows for a more nuanced understanding of the data, enhancing the model's ability to make accurate predictions.

Additionally, the hyperparameters of all baseline models were optimized based on the AUC score with "Grid Search" method in cross-validation process. Specifically, LR, SVM and Random Forest were employed using scikit-learn package in Python and XGBoost was developed using xgboost package. The comprehensive parameter descriptions for SVM, Random Forest, and XGBoost, which have been fine-tuned using the Grid Search method, are presented below:

- **SVM:** In WM dataset, we the following hyperparameter set $\{kernel = poly, degree = 2, gamma = scale, coef = 0.0, decision\_function\_shape = over.\}$ In TCM dataset, the hyperpatameter set is $\{kernel = rbf, degree = 3, gamma = scale, coef = 0.0, decision\_function\_shape = over\}$.
- **Random Forest:** In WM dataset, we the following hyperparameter set $\{n\_estimators = 50, max\_depth = 3, min\_sample\_split = 2, max\_features = 1.0, boostrap = True\}$. In TCM dataset, the hyperpatameter set is $\{n\_estimators = 30, max\_depth = 2, min\_sample\_split = 3, max\_features = 1.0, boostrap = True\}$.

- **XGBoost:** In WM dataset, we the following hyperparameter set $\{n\_estimators = 120, max\_depth = 5, max\_leaves = 8, learning\_rate = 0.05, min\_child\_weight = 1, subsample = 0.7, colsample\_bytree = 0.7\}$. In TCM dataset, the hyperpatameter set is $\{n\_estimators = 120, max\_depth = 7, max\_leaves = 10, learning\_rate = 0.035, min\_child\_weight = 1, subsample = 0.6, colsample\_bytree = 1.0\}$

3.3 Results and Analysis

3.3.1 Comparison with Different Models

Table 2 and Table 3 present the performance results of AMI-Net and other models on the two medical datasets. When evaluating the WM dataset, AMI-Net demonstrated superior performance in terms of Precision and F1-score, as indicated in Table 2. Notably, the F1-score achieved by the model was significantly higher compared to other models, indicating a balanced performance between precision and recall. Although the AUC and Accuracy scores were slightly lower than those of the four classical machine learning algorithms, the overall performance of AMI-Net remained commendable.

In the context of the TCM dataset, which was highly imbalanced, AMI-Net outshone all other models in terms of Precision, Recall, and F1-score. This demonstrates the its robustness in handling imbalanced datasets, a common challenge in medical data analysis.

In spite of the incomplete and low-quality nature of the two datasets, which at times even lacked a sufficient number of positive samples, AMI-Net demonstrated greater resilience and dependability compared to other models in these demanding circumstances. This

TABLE 3.2: Model Performance on the WM Dataset

Models	AUC	Accuracy	Precision	Recall	F1
LR	0.732	0.954	0.250	0.019	0.035
SVM	0.657	**0.956**	0.232	0.270	0.249
Random Forest	**0.767**	**0.946**	0.132	0.171	0.148
XGBoost	0.706	0.945	0.100	0.007	0.013
mi-Net	0.556	0.621	0.089	**0.482**	0.150
MI-Net	0.554	0.782	0.151	0.253	0.189
MI-Net+DS	0.512	0.601	0.025	0.303	0.046
MI-Net+RC	0.586	0.837	0.323	0.228	0.267
Att. Net	0.608	0.849	0.342	0.143	0.202
Gated Att. Net	0.576	0.832	0.248	0.140	0.179
AMI-Net	0.702	0.907	**0.356**	0.283	**0.314**

TABLE 3.3: Model Performance on the TCM Dataset

Models	AUC	Accuracy	Precision	Recall	F1
LR	0.755	0.882	0.396	0.116	0.179
SVM	0.703	**0.889**	0.272	0.109	0.156
Random Forest	0.737	**0.889**	0.310	0.089	0.138
XGBoost	0.729	0.886	0.327	0.063	0.106
mi-Net	0.587	0.621	0.210	0.412	0.278
MI-Net	0.665	0.813	0.364	0.414	0.387
MI-Net+DS	0.586	0.731	0.358	0.290	0.320
MI-Net+RC	0.592	0.863	0.324	0.359	0.341
Att. Net	0.642	0.861	0.368	0.244	0.293
Gated Att. Net	0.607	0.755	0.319	0.354	0.336
AMI-Net	0.702	0.818	**0.399**	**0.468**	**0.431**

resilience sets AMI-Net apart from other models, showcasing its ability to handle challenging data conditions.

Furthermore, multi-instance neural networks exhibited efficient and effective performance in terms of Precision, Recall, and F1-score. These networks excelled in extracting crucial information from positive samples, underscoring the advantages of MIL in practical applications. This is particularly evident in the medical domain, where the ability to glean insights from sparse and imbalanced data is of paramount importance. The results of this study affirm the potential of MIL as a powerful tool for real-world medical data analysis.

3.3.2 Comparison of Different Number of Heads

To investigate the impact of varying the number of heads in the multi-head attention transformer on the performance of the proposed method, experiments were conducted with 0, 2, 4, 8, 16, and 32 heads on both the TCM and WM datasets. Here, 0 signifies a model without multi-head attention. The performance was evaluated using the F1-score.

As depicted in Figure 3.3, the optimal performance was achieved when the transformer was configured with four heads. This suggests that partitioning the data into four subspaces enabled the model to most effectively capture the intra-relations of symptoms. Interpreted through the lens of medical knowledge, this implies that a patient's condition is best understood when considered from four distinct aspects, within which symptoms exhibit high correlation.

Interestingly, this finding aligns with the data collection methodology of the TCM dataset. Symptoms in this dataset were gathered using four diagnostic methods: inspection, listening and smelling, inquiry, and pulse-taking. These methods represent four primary aspects of a patient's bodily condition, mirroring the four subspaces used in the model. This congruence between the experimental result and the TCM diagnostic approach underscores the relevance and applicability of the method in the context of TCM.

Furthermore, the results revealed that the model without multi-head attention performed the worst. This underscores the importance of identifying correlations among symptoms and bridging the gap between standardized and unstandardized symptom expressions. The multi-head attention mechanism plays a crucial role in achieving these objectives, further validating its inclusion in the method.

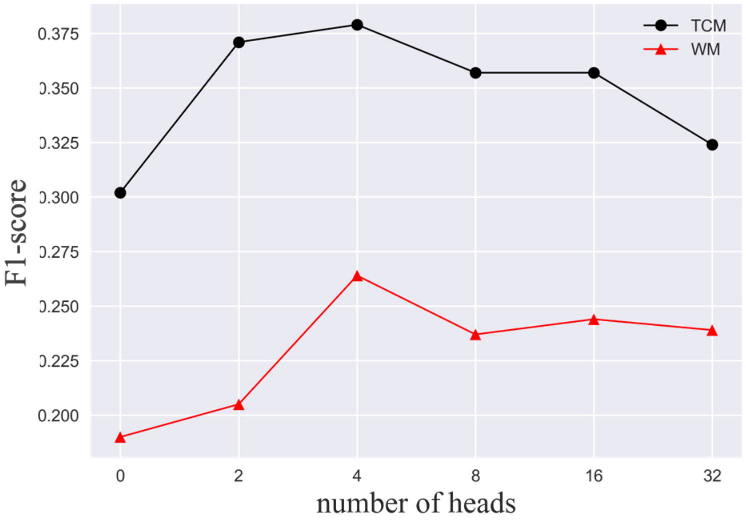


FIGURE 3.3: Comparison of different number of heads. 0 denotes the model without multi-head attention.

3.3.3 Comparison of Different MIL Pooling Methods

To evaluate the influence of various instance-level and bag-level Multi-Instance Learning (MIL) pooling methods, the F1-scores resulting from different combinations of these methods were compared. Prior studies have explored the use of max pooling [259] and attention-based pooling [118]. In this study, the scope was extended to include sum pooling, max pooling, and attention-based pooling. By comparing these different pooling strategies, the goal was to identify the most effective approach for the specific task.

The outcomes of these comparisons are presented in Figures 3.4 and 3.5. The performance of sum pooling and gated-attention based pooling at both the instance-level and bag-level outperformed other MIL pooling methods. This suggests that these methods are particularly effective at capturing the nuances of the data.

Interestingly, the model employing max pooling at the instance-level demonstrated the poorest performance. This result indicates that symptoms are interconnected across various embedding dimensions. Relying on information captured in a single dimension for diagnosis could lead to an incomplete understanding of the patient's condition. This underscores the importance of a multi-dimensional approach in capturing the complex relationships among symptoms.

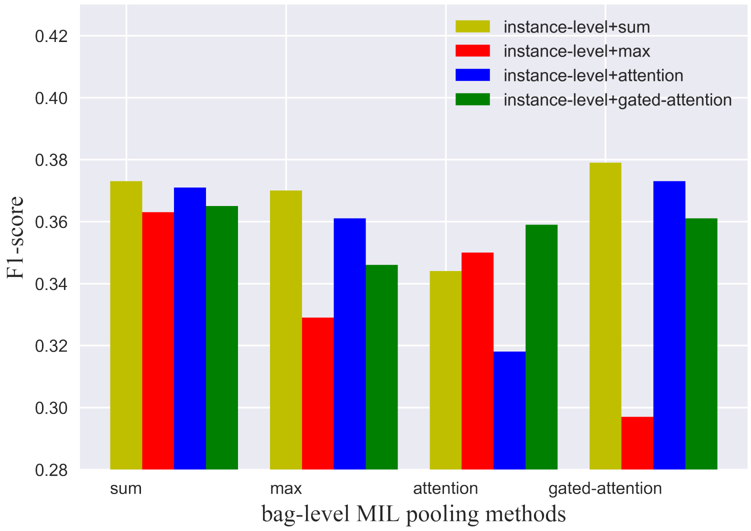


FIGURE 3.4: Comparison of different MIL pooling methods on the TCM dataset.

3.3.4 Influence of Data Noise

Given that data from real-world studies often exhibit varying degrees of inaccuracy and ambiguity, experiments were conducted to assess the robustness of the proposed model

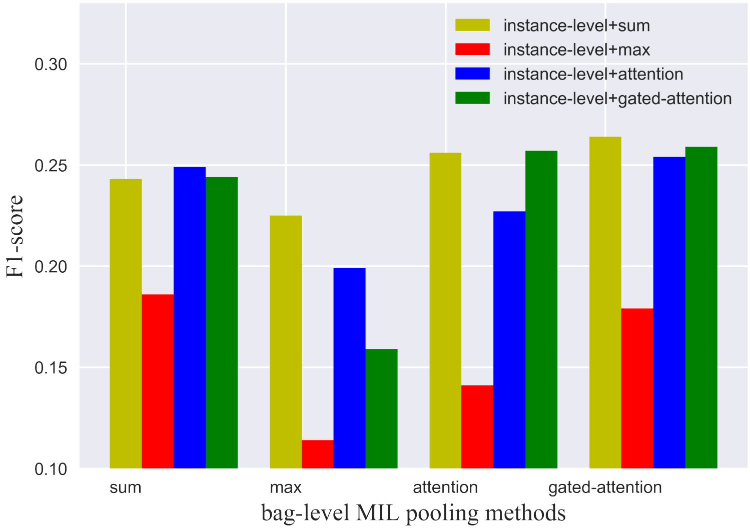


FIGURE 3.5: Comparison of different MIL pooling methods on the WM dataset.

against different data noise ratios, using the F1-score as the performance metric. Two dimensions of data noise were considered: feature noise and label noise.

To simulate feature noise, 1, 2, 3, 4, and 5 symptoms in each training sample were randomly altered. If a sample contained fewer symptoms than the required number to be changed, new symptoms were randomly added. For label noise, the ratio of labels in the training set was inverted from 0.1 to 1.0, incremented by 0.1 each time.

The impact of feature noise on different models is illustrated in Figure 3.6. Despite random changes to some symptoms, the proposed method maintained superior performance compared to all other models, with minimal fluctuations. The MIL methods demonstrated more stable performance than all other machine learning algorithms, attributable to their ability to capture and utilize effective information. This stability underscores their reliability in real-world applications.

Figure 3.6 also presents the influence of label noise. As the proportion of inverted labels increased, the F1-score converged around 0.2 and 0.1, respectively, when all samples in the validation set were labeled as positive. Despite this noise, the proposed method consistently outperformed others, suggesting its superior resistance to noise. This resilience further validates the robustness of the method in handling real-world data complexities.

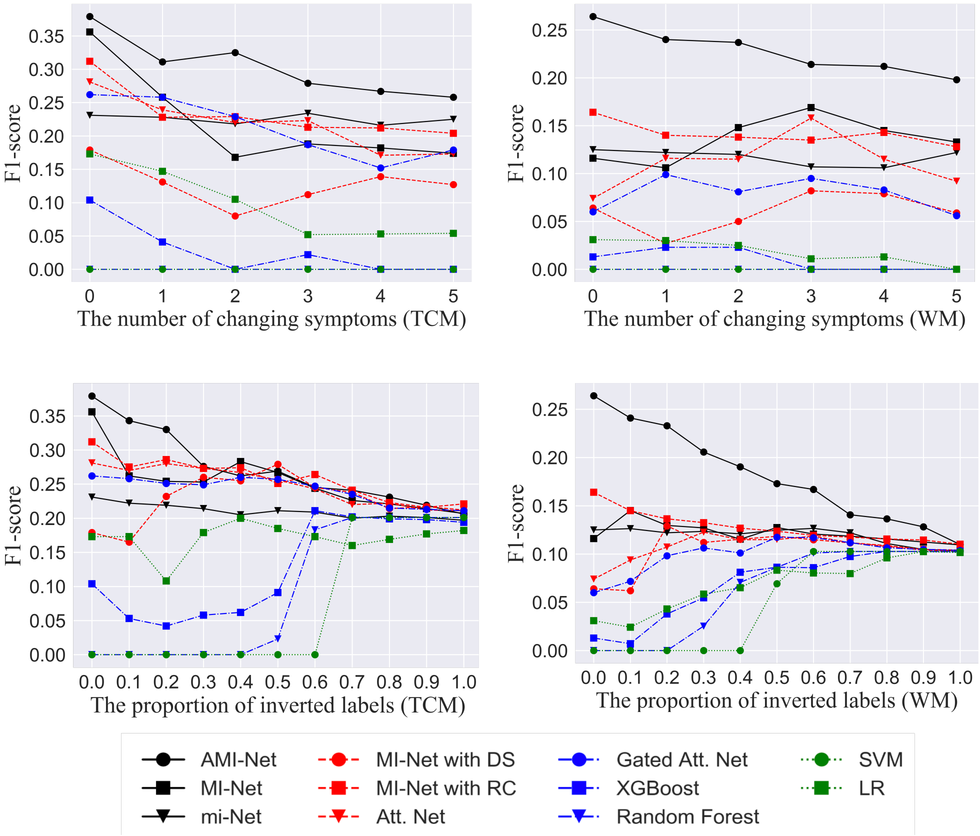


FIGURE 3.6: Test for the influence of feature and label noise

3.3.5 Influence of Incomplete Data

In this section, the performance of the proposed method, AMI-Net, was evaluated on incomplete datasets. Simulated incomplete data was created by randomly deleting 1, 2, 3, 4, and 5 symptoms from each training sample. If a sample had fewer symptoms than the number of deletions, all symptoms were removed. The F1-score was used as the performance metric.

The results, as depicted in Figure 3.7, indicated that AMI-Net exhibited greater robustness than all other models when confronted with incomplete data. This finding is particularly significant as real-world scenarios often involve missing information due to patients not undergoing all examinations or clinical measurements. Clinicians frequently need to infer missing information based on their experience and knowledge during the diagnostic process. The proposed method offers a promising strategy for handling such incomplete data, showcasing its resilience in the face of missing information.

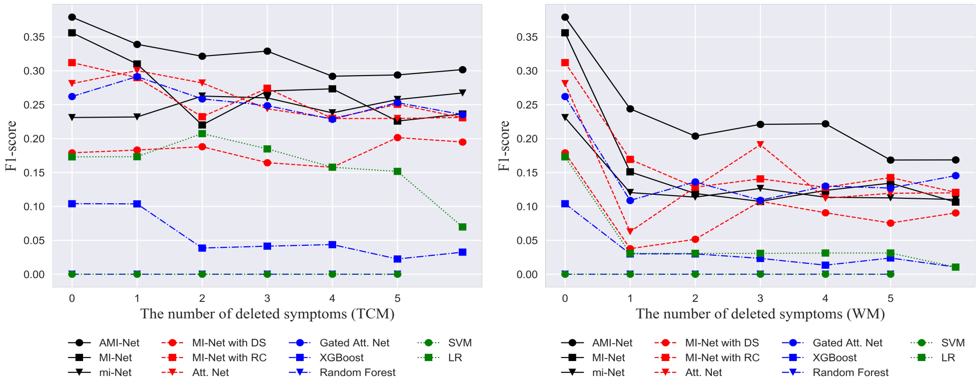


FIGURE 3.7: Test for the influence of incomplete data

3.3.6 Visualization of Attention

The gated attention-based MIL pooling layer in AMI-Net has the capability to select the most informative instances, in this case, symptoms. To enhance the interpretability of AMI-Net, the attention mechanism on two examples are visualized, as displayed in Figure 3.8. The importance of each symptom is represented by color intensity, with darker colors indicating higher importance.

In the WM dataset, factors such as "personal income 3000-5000", "unmarried", "length of stay<25 days" and "MECT<=1" were found to be dominant predictors of schizophrenia relapse.

In the TCM dataset, for the prediction of meridian obstruction, symptoms such as "decreased defecation", "urine color is yellow", "heavy legs" and "dropping" were assigned larger weights, signifying their importance.

These visualizations provide valuable insights into the decision-making process of AMI-Net, enhancing its interpretability and potential for practical application.

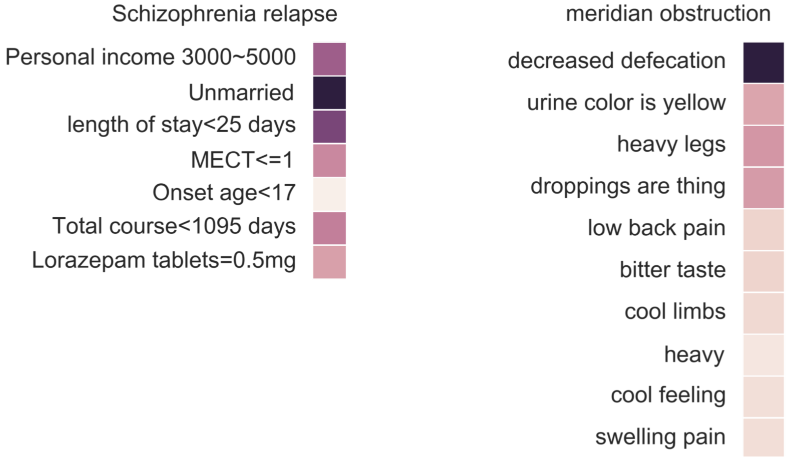


FIGURE 3.8: An example of informative instances selection

3.4 Summary

In this study, a novel attention-based multi-instance neural network, AMI-Net, was proposed to tackle the challenges of medical diagnosis using incomplete and low-quality data. The AMI-Net architecture encompasses two key components: capturing intra-relations among instances and selecting crucial instances for the final classification. This approach enables a more comprehensive understanding of the data, resulting in improved diagnostic accuracy.

The experimental results showcased the superiority of the proposed method compared to other models in real-world medical applications. AMI-Net outperformed alternative models in terms of Precision, Recall, and F1-score, indicating its efficacy and efficiency.

Additionally, the interpretability of AMI-Net was highlighted, offering valuable insights into the decision-making process of the model.

Notably, AMI-Net demonstrated robust performance even under challenging conditions characterized by noisy and incomplete data. This resilience makes it a valuable tool for real-world medical scenarios, where data quality and completeness can often be compromised.

In conclusion, this study establishes the effectiveness of AMI-Net in addressing real-life medical diagnosis challenges. It contributes to the advancement of real-world medical research by providing a promising solution for handling the intricacies and uncertainties inherent in medical data. The findings affirm the potential of Multi-Instance Learning methods in enhancing the accuracy and reliability of medical diagnoses, paving the way for further research in this field.

CHAPTER 4

AMI-Net+

The ability to effectively handle and learn from incomplete and low-quality data is pivotal in creating robust predictive models. The proposed AMI-Net demonstrates a superior capability to effectively handle and learn from incomplete and low-quality data in creating robust predictive models. However, despite its advantages, there are still substantial challenges that need to be addressed to enhance the efficacy of AMI-Net.

One of the key challenges in AMI-Net lies in its MIL pooling mechanism. The performance of predictive models depends heavily on the chosen MIL pooling method, and, unfortunately, no single method is universally suitable for all data types. The need to adjust the pooling method according to specific data cases is a significant constraint on the model's overall versatility, requiring additional customization and thereby limiting efficiency.

Furthermore, the issue of handling imbalanced data has been inadequately addressed in the current version of AMI-Net. As missing values in datasets increase, the sensitivity and accuracy of the model display a decline. Imbalanced datasets, particularly those with a high volume of missing values, significantly challenge the performance of the model, raising concern about its reliability and applicability in real-world settings, where such problems are commonplace.

In response to the identified limitations of AMI-Net, a cutting-edge enhancement known as AMI-Net+ is proposed. This novel algorithm uses AMI-Net as its foundational architecture, building on its proven strengths while introducing innovative improvements to address its inherent challenges.

The heart of AMI-Net+ is a newly developed, self-adaptive multi-instance pooling method. This methodology operates at the instance level, generating a robust representation of the data bag. In contrast to static pooling techniques, the self-adaptive method dynamically adjusts to the specifics of the data case, thereby providing a more tailored, data-sensitive approach to capturing information from diverse instances within the bag. This innovation enhances the accuracy of the bag representation and, consequently, the performance of the model.

In terms of addressing the problem of imbalanced data, the integration of the focal loss function into the neural network model is proposed. Traditionally, cross-entropy has been the common choice for dealing with classification tasks. However, the focal loss, originally proposed in the object detection community to address the issue of extreme foreground-background class imbalance, provides a more efficient solution for the purposes of this study. Focal loss has demonstrated superior performance in scenarios characterized by high class imbalance. It cleverly reduces the attention placed on well-classified instances while focusing more intently on the hard, less represented, and misclassified ones.

The incorporation of focal loss into the AMI-Net+ architecture is designed to bolster the model's resilience against data imbalance, thereby increasing its robustness and overall predictive accuracy. The integration of focal loss and the implementation of the self-adaptive multi-instance pooling method are pivotal to the advancement of

AMI-Net+, enabling it to navigate the complexities of real-world data more efficiently and effectively.

The detailed mechanics of these innovative elements, their integration into the model, and their impact on the overall performance of AMI-Net+ will be meticulously described in the ensuing chapter. It is believed that these innovations will make significant strides in addressing the limitations of AMI-Net, pushing the boundaries of predictive learning using incomplete and imbalanced datasets.

4.1 Model Architecture

The foundational architecture of AMI-Net+ is depicted in Figure 4.1, elucidating its methodical process of data computation. This model is designed to receive a "bag" of symptoms - referred to as instances - as its initial input, facilitating the model's understanding of the complexities inherent in each unique health scenario.

Each instance undergoes a transformative process beginning at the embedding layer, where it is mapped to a dense vector, thus generating instance embeddings. This process of embedding serves to convert the initial, often heterogeneous input data into a compact, unified format that facilitates the model's ability to draw meaningful inferences.

Subsequently, the model employs multi-head attention to analyze the instance embeddings. This mechanism is designed to scrutinize the information from multiple perspectives, thereby enhancing the breadth and depth of the model's understanding. This step is further enhanced by layer normalization [11] and residual connection [103], which work collectively to mine the correlations between the instances, thus yielding valuable contextual insights.

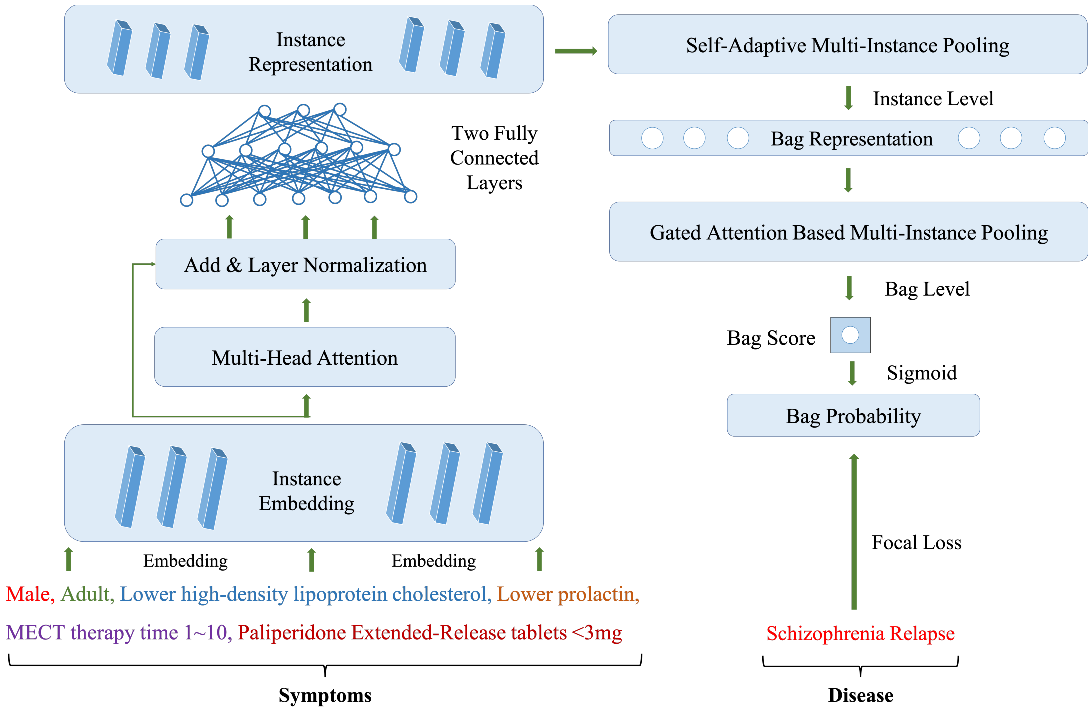

Instance Representation
Self-Adaptive Multi-Instance Pooling
Instance Level
Two Fully Connected Layers
Bag Representation
Gated Attention Based Multi-Instance Pooling
Add & Layer Normalization
Bag Level
Bag Score
Multi-Head Attention
Sigmoid
Bag Probability
Instance Embedding
Focal Loss
Embedding
Embedding
Male, Adult, Lower high-density lipoprotein cholesterol, Lower prolactin,
MECT therapy time 1~10, Paliperidone Extended-Release tablets <3mg
Schizophrenia Relapse
Symptoms
Disease

FIGURE 4.1: The overall architecture of AMI-Net+

Following these initial analytical steps, AMI-Net+ implements a series of fully connected layers, designed to estimate instance representations. This network of interconnected nodes facilitates a nuanced, comprehensive analysis of the instance embeddings, drawing on the collective intelligence of the layers to generate intricate, detailed representations.

The resulting instance representations are then processed using the innovative self-adaptive multi-instance pooling method at the instance level to construct the overall bag representation. This technique dynamically adjusts to each instance, resulting in a more precise and comprehensive representation of the bag of symptoms.

The computation of the bag score is handled by a unique, gated attention-based multi-instance pooling mechanism, specifically designed to function at the bag level. This advanced algorithm focuses on synthesizing the information in the bag representation to calculate a comprehensive bag score.

Ultimately, the model employs a sigmoid function and focal loss for supervision. The sigmoid function serves to convert the output into a probability, facilitating binary classification. In tandem with this, the focal loss is utilized to improve the model's handling of imbalanced data, as previously discussed.

4.1.1 Self-Adaptive Multi-Instance Pooling

In this study, a novel technique is proposed called self-adaptive multi-instance pooling, designed specifically for instance-level attributes processing. The overarching goal of this method is to effectively learn and construct a bag representation, which serves as a comprehensive summary of the input instances, and is utilized in subsequent classification tasks.

This process commences with the input of instance representations, each of which encapsulates a symptom or condition in the given medical scenario. These representations are then processed through a variety of untrainable pooling methods. Instead of being subject to change and refinement during the learning process, these pooling methods provide consistent, reliable frameworks for data manipulation, each extracting a unique bag representation from the instance inputs.

Conceptually, each bag representation can be seen as a 'view' or a unique descriptive perspective of the original bag of instances. The multiplicity of these views mirrors the diversity and complexity of the input data, providing a holistic and nuanced portrayal of the bag of instances.

Drawing inspiration from the principles of multi-view learning and ensemble learning, a strategy is designed to integrate these varied bag representations. In this innovative approach, all bag representations are concatenated, creating a unified stream of information. This concatenated data is then processed through a dense layer, which performs the essential function of calculating a weighted sum of the different views.

The resultant weighted sum integrates the strengths of each individual view, harmonizing them into a cohesive and balanced understanding of the bag of instances. In this way, the model can benefit from the insights each pooling method offers, thereby maximizing the informative value of the bag representation.

Through the deployment of self-adaptive multi-instance pooling, the aim is to enrich the capacity of the model to learn from and accurately classify real-world medical features.

Let X be a bag of N instances with K dimensions, and we formulate this proposed pooling method as:

$$\forall_{j=1,2,\ldots,N} : \text{ view }_{v} = \text{ Pooling }_{k=1,2,\ldots,K} \{x_{j,k}\} \tag{4.1}$$

$$\text{SelfAdaptive } = \text{ Concat }_{v=1,2,\ldots,V} \{ \text{ view }_{v} \} \, W^{v} \tag{4.2}$$

where $W^v \in \mathbb{R}^{V \times 1}, x \in X$ and V is the number of selected pooling methods. In this study, a diverse set of pooling techniques is performed, namely, max pooling, mean pooling, sum pooling and log-sum-exp pooling.

4.1.2 Focal Loss for Imbalanced Data

In the quest to address the pervasive problem of extreme data imbalance, an innovative modification to the loss function is proposed. Specifically, the conventional cross-entropy loss is suggested to be substituted with a more tailored alternative, the focal loss, as referenced in the literature[144]. The unique advantage of the focal loss function is its ability to steer the model's focus towards the difficult and misclassified samples.

Upon completion of each feed-forward pass in the network, a predicted bag probability, represented as y_{pred}, is obtained. The corresponding true bag label is given by y_{true}. This juxtaposition of predicted and true outcomes is a crucial component in determining the accuracy of the model's predictions and thereby its performance.

To optimize the AMI-Net+ model and further hone its predictive capability, the focal loss function is incorporated into the network's training process. This decision is driven by the focal loss function's unique ability to alleviate the problem of extreme data imbalance, a common and vexing issue in many medical applications. The focal loss function is specifically designed to down-weight the contribution of easy-to-classify examples and amplify the importance of those that are hard to classify or often misclassified.

The computation of the focal loss function in the AMI-Net+ model is as follows:

$$p_t = \begin{cases} y_{\text{pred}} & \text{if } y_{\text{true}} = 1 \\ 1 - y_{\text{pred}} & \text{if } y_{\text{true}} = 0 \end{cases} \tag{4.3}$$

$$\text{FocalLoss } = -\alpha \left(1 - p_t\right)^{\gamma} \log\left(p_t\right) \tag{4.4}$$

where $\gamma \geq 0$ reduces the loss contribution of easily classified samples, and α is a balance factor. In the experiments, it is found that $\gamma = 2$ and $\alpha = 0.25$ achieved the best performance. This revised loss function serves as a key ingredient in the AMI-Net+ model, enhancing its robustness and versatility in handling diverse and imbalanced datasets. By focusing on the challenging and often neglected samples, the model can yield more nuanced and accurate predictions, ultimately improving its overall performance in real-world medical applications.

4.2 Experiments

4.2.1 Data Description

To thoroughly evaluate the performance and resilience of the AMI-Net+ model, meticulous testing was carried out using the identical datasets deployed in the AMI-Net trials. These consist of two authentic medical datasets, each drawn from distinct medical paradigms — Traditional Chinese Medicine (TCM) and Western Medicine (WM). Illustrative examples of these datasets are provided in Table 4.1. It is notable that the TCM dataset exhibits a notable imbalance, with a total of 1436 control patients and a significantly smaller subset of 181 case patients. The WM dataset also displays a pronounced imbalance, possessing a mere 224 positive labels amongst a pool of 3927 patients, which translates to a scant positive rate of 0.057.

These diverse datasets, with their inherent complexity and imbalance, provide a challenging yet fitting environment to evaluate the performance and adaptability of the AMI-Net+ model. The results derived from these evaluations will offer valuable insights into the model's predictive performance and its ability to handle real-world medical data.

TABLE 4.1: Examples of TCM and WM datasets

Dataset	Features	Diagnosis
TCM	Urine color yellow, Sweat, Pruritus, Coldness of extremities, Perspiration	Meridian obstruction
	Dark red tongue, Palpitation	Not meridian obstruction
WM	Personal income 3000 5000, Unmarried, LOS<10 days, MECT<=1, Onset age<17, Total course<1095 days, Lorazepam tablets=0.5mg	Schizophrenia relapse
	Personal income 1000 3000, Married, High levels of prolactin, hyperglycemia, High levels of corticotrophin, LOS 25 49 days, MECT 1 10, Onset age 1-10, Risperidone<=1mg, Total course 1095 5840 days, Haloperidol injection 5mg	Not relapse

4.2.2 Experimental Setup

For the analysis, the first step involved padding each record to match the maximum length and then transforming each medical feature or symptom into a 512-dimensional dense vector using an embedding process. Multi-head attention mechanisms were employed in the model, utilizing 4 and 8 heads respectively for the Traditional Chinese Medicine (TCM) and Western Medicine (WM) datasets. Subsequently, two fully connected layers were implemented, each with hidden sizes of 256 and 128 respectively.

To manage the prevalent issue of extreme imbalance, focal loss was integrated, with the parameters α and γ set at 0.25 and 2, respectively. The Adam optimizer was applied to minimize the focal loss over the training data, with a learning rate of $1e^{-5}$, ϵ of $1e^{-8}$, and momentum parameters β_1 and β_2 designated as 0.9 and 0.98 respectively.

The evaluation metrics comprised the Area Under the Curve (AUC), Accuracy, Precision, and Recall. During the training process, the number of epochs was set at 500 and the batch size at 64. An early stopping strategy was also incorporated to select the optimal model based on the AUC score.

To facilitate a fair comparison of the model with other methodologies, experiments were conducted employing a 5-fold cross-validation approach. Several baseline models were used for comparison, including logistic regression (LR), support vector machine (SVM), random forest (RF), XGBoost (XGB), mi-Net, MI-Net, MI-Net with DS, MI-Net with RC, and attention and gated attention based multi-instance neural networks (Att. Net, Gated Att. Net). Among these, LR, SVM, RF, and XGB are traditional machine learning algorithms constructed on the dataset in a one-hot format with zero imputation. Moreover, the parameters of the baseline models were fine-tuned according to the AUC scores obtained from the validation dataset.

This diverse selection of comparison models and comprehensive training regimen enables a rigorous test of the robustness and efficacy of the proposed AMI-Net+ model. The results obtained will provide valuable insights into the model's performance, its relative strengths, and areas of potential improvement.

4.3 Results and Analysis

4.3.1 Comparison with Baseline Models

Table 4.2 presents a comparison of the proposed model's performance with that of several baseline models. It is noteworthy that the method outperforms the others in terms of both the Area Under the Curve (AUC) and recall scores. This demonstrates the model's exceptional ability to extract informative features, even from a very limited number of positive samples.

This capability is especially crucial in the realm of medical diagnosis. It is of paramount importance that diseases are not overlooked in any patient. However, the collection of sufficient positive samples presents a considerable challenge. Two widely used algorithms, Random Forest (RF) and XGBoost (XGB), were unable to identify any positive samples within the evaluation dataset, further underlining the significance of the model's performance.

Furthermore, a comparison of Precision, AUC, and Recall scores between Multi-Instance Neural Networks (MINNs), including mi-Net, MI-Net, and Attention Network (Att. Net), and classical machine learning methods such as Logistic Regression (LR), Support Vector Machine (SVM), RF, and XGB, indicates the superior performance of MINNs. This reinforces the assertion that Multi-Instance Learning (MIL) methods hold a greater potential for successful implementation in many real-world applications, especially within the medical domain.

The superior performance of the proposed AMI-Net+ model in terms of precision, AUC, and recall underscores its potential utility in medical applications. The results suggest that AMI-Net+ could potentially revolutionize the domain of medical diagnostics, significantly improving the detection and management of diseases, ultimately leading to improved patient outcomes.

4.3.2 Comparison of AMI-Net+ with Different Number of Heads

In this section, the objective is to evaluate the effect of varying the number of heads within the multi-head attention mechanism on model performance. This experiment was performed on both the Traditional Chinese Medicine (TCM) and Western Medicine (WM) datasets, employing configurations with 0, 4, 8, 16, and 32 heads. In this context,

TABLE 4.2: Performance comparison on TCM and WM datasets.

Models	TCM				WM			
	AUC	Accuracy	Precision	Recall	AUC	Accuracy	Precision	Recall
LR	0.760	0.944	0.200	0.017	0.755	0.882	0.396	0.116
SVM	0.657	**0.946**	0	0	0.703	**0.889**	0	0
RF	0.767	**0.946**	0	0	0.737	**0.889**	0	0
XGBoost	0.706	0.945	0.100	0.007	0.729	0.886	0.327	0.063
mi-Net	0.565	0.624	0.088	0.469	0.597	0.641	0.220	0.422
MI-Net	0.545	0.787	0.154	0.251	0.665	0.813	0.364	0.414
MI-Net with DS	0.510	0.621	0.045	0.383	0.586	0.731	0.358	0.290
MI-Net with RC	0.588	0.867	0.313	0.228	0596	0.861	0.353	0.358
Att. Net	0.608	0.849	0.342	0.143	0.642	0.861	0.368	0.244
Gated Att. Net	0.576	0.832	0.248	0.140	0.607	0.755	0.319	0.354
AMI-Net	0.702	0.907	**0.356**	0.283	0.702	0.818	**0.399**	0.468
AMI-Net+	**0.774**	0.779	0.301	**0.689**	**0.761**	0.802	0.165	**0.644**

’0’ signifies a model configuration that does not incorporate multi-head attention. The model’s performance was assessed using the AUC score as the metric.

As illustrated in Figure 4.2, the model lacking multi-head attention demonstrated a considerably weaker performance compared to the other configurations. This disparity highlights the critical role multi-head attention plays in identifying the correlations among clinical features prior to the classification task.

Moreover, when examining the performance on the WM dataset, the model equipped with 8 heads was found to be the optimal choice. This implies that clinical features in the WM dataset exhibit correlations predominantly along eight dimensions. In contrast, the model achieved peak performance on the TCM dataset when configured with 4 heads. This indicates an efficient exploration and identification of symptom correlations in the TCM context.

Hence, the findings support the application of multi-head attention in the model, while also emphasizing the importance of adjusting the number of attention heads to match the complexity and dimensional interdependencies of the dataset.

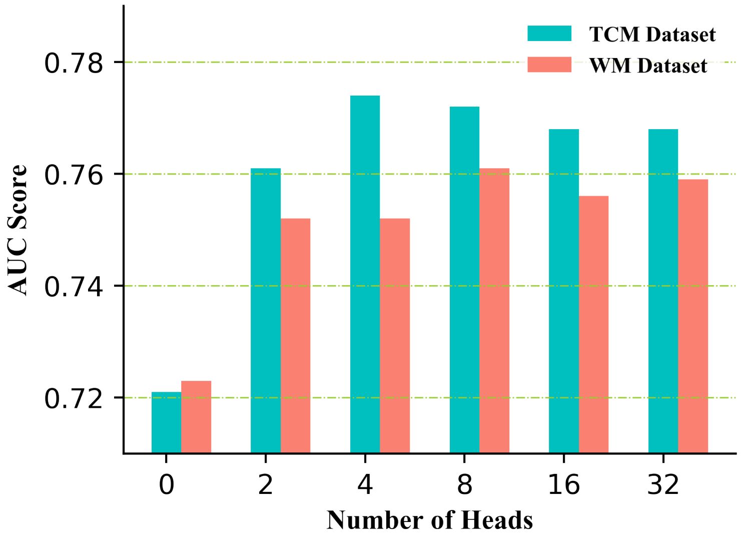


FIGURE 4.2: Comparison of AMI-Net+ with different number of heads in the multi-head attention.

4.3.3 Comparison of AMI-Net+ with Different Multi-Instance Pooling Methods

In this analysis, the impact of using various multi-instance pooling methods at the instance level on the model's performance, as measured by the AUC score, is investigated. Prior research [118, 263] has demonstrated the superiority of both max pooling and attention-based pooling, leading to the selection of these methods as baseline approaches for comparison.

The results, as illustrated in Figure 4.3, clearly demonstrate the effectiveness of the proposed pooling method, which exhibits superior performance compared to the others. This highlights the notion that a well-crafted pooling strategy can significantly enhance the efficiency and accuracy of multi-instance learning models in medical diagnosis tasks.

Interestingly, it is found that max pooling yields the least impressive performance. This underscores the limitation of relying solely on capturing information from a single embedding dimension when representing an instance. The simplistic approach of max pooling may disregard critical information present in other dimensions, resulting in suboptimal representations and poorer performance.

The proposed method effectively addresses this issue by employing a more complex and informative strategy that preserves the necessary multi-dimensionality of instances. Thus, these findings strongly advocate for the utilization of sophisticated multi-instance pooling methods to fully capture the nuanced information contained within medical instances and to maximize the potential of predictive models in healthcare applications.

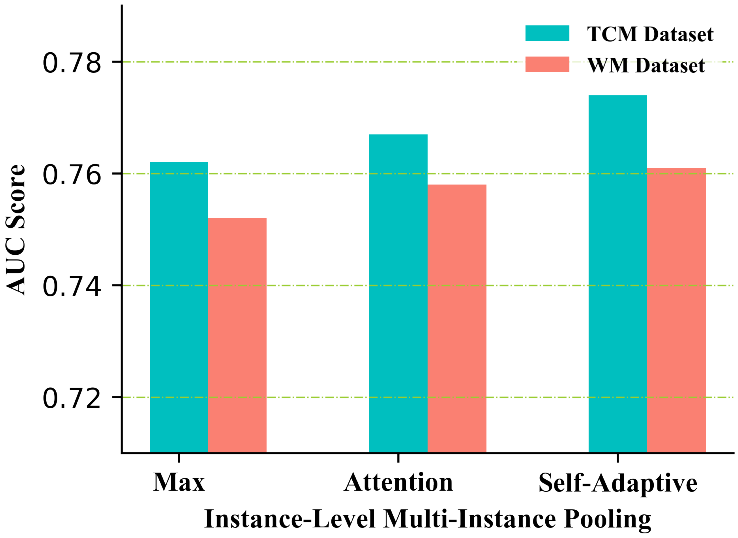


FIGURE 4.3: Comparison of AMI-Net+ with different multi-instance pooling methods on instance level.

4.3.4 Evaluation of Focal Loss

To elucidate the functionality and effectiveness of focal loss (FL), a comparative study is conducted on the model's performance with both focal loss and cross-entropy loss (CE) across the two datasets. Table 4.3 presents the comparative results, revealing the superior capabilities of the model utilizing FL in correctly identifying positive samples as compared to the one deploying CE.

Although the model with FL records lower Accuracy and Precision scores than the counterpart with CE, it's crucial to consider the context of the extreme data imbalance in the datasets. In such scenarios, the Recall score becomes a paramount indicator due to its focus on the capability of the model to correctly identify positive cases, which is a crucial requirement in medical diagnosis. If all predictions are naught, an Accuracy score of 0.946 can still be achieved, thereby illustrating its lack of reliability in such imbalanced contexts.

Generally, the incorporation of FL into the model manifests a simplified yet remarkably effective solution to address the challenge posed by extremely imbalanced data. By specifically mitigating the disproportionate focus on the majority class, FL assists the model in emphasizing harder, misclassified cases, thereby improving the performance in identifying rare, but often critical, positive instances. This strategy is particularly valuable in medical settings where missing a positive diagnosis could have significant consequences, thus asserting the superiority of FL in such applications.

TABLE 4.3: Performance comparison of focal loss and cross-entropy loss.

Loss	TCM				WM			
	AUC	Accuracy	Precision	Recall	AUC	Accuracy	Precision	Recall
FL	**0.774**	0.779	0.301	**0.689**	**0.761**	0.802	0.165	**0.644**
CE	0.746	**0.863**	**0.391**	0.394	0.707	**0.939**	**0.398**	0.204

4.4 Summary

This study addresses the dual challenge of incomplete and extremely imbalanced data through the development of AMI-Net+. The architecture of this network incorporates a multi-head attention mechanism coupled with a gated attention-based multi-instance pooling method. This integration facilitates efficient capture of symptom correlations and their informative elements. To enhance the model performance, a novel instance-level multi-instance pooling method is proposed to achieve improved bag representation. Additionally, the traditional cross-entropy loss is replaced with focal loss, a more effective method for handling severe class imbalance.

The experimental findings convincingly demonstrate the superior performance of the proposed method compared to a range of baseline models, as evidenced by higher AUC and Recall scores. This performance superiority highlights the importance of the AMI-Net+ model and its associated mechanisms in addressing the unique challenges of medical data.

Beyond theoretical constructs, this research validates the practical applicability and efficacy of AMI-Net+ in real-world medical applications. It proves to be a promising tool for handling the unique challenges of real-world medical data, characterized by imbalances and incompleteness. Furthermore, it sets a significant precedent for future research in this domain, providing a solid foundation and insightful direction for developing innovative solutions to address other complex data problems.

CHAPTER 5

AMI-Net3

Despite the notable performance of both AMI-Net and AMI-Net+, their practical application has been primarily confined to binary data, thus limiting their potential influence in real-world situations. In response to this constraint and with an aim to amplify their predictive prowess, a novel framework called AMI-Net3 is introduced in this chapter.

Within AMI-Net3, the problem is approached by viewing each patient as a bag populated with varying numbers of feature-value pairs, referred to as instances. Leveraging the proposed feature embedding technique, these instances are mapped to an embedding space as demonstrated in Figure 5.1. This direct learning methodology equips the predictive models with the capability to naturally attenuate the detrimental impact of missing data. Furthermore, an innovative architecture called the Multi-Instance Neural Network (MINN) is introduced, which excels at managing redundant and highly correlated features. MINN integrates an attention mechanism designed to discern informative features and their interconnections, providing empirical evidence within a clinical context.

To optimally tune AMI-Net3, all transformations are parameterized using neural networks under both primary and auxiliary supervision, harnessing the focal loss function as elucidated by Lin et al. [144]. Initially conceived for object detection tasks, focal loss adeptly addresses severe class imbalance by concentrating on difficult, infrequent, and misclassified instances.

In the experimental evaluations, the performance of AMI-Net3 is assessed across three diverse datasets extracted from real-world scenarios. Each dataset corresponds to distinct clinical risk prediction tasks, specifically: adverse drug reaction of risperidone, schizophrenia relapse, and invasive fungal infections. The results unambiguously demonstrate that the proposed AMI-Net3 framework substantially outperforms other competitive baseline models across all three medical datasets.

Beyond its performance, the AMI-Net3 framework presents an innovative paradigm capable of harnessing deep learning techniques drawn from various domains, such as computer vision (CV) and natural language processing (NLP), to address challenges posed by real-world clinical risk prediction applications. This novel approach promises to facilitate the development of advanced systems, including mortality prediction and adverse drug reaction warning systems. By addressing these pivotal challenges, AMI-Net3 contributes to the field of medical data analysis and paves the way for future research in real-world clinical risk prediction.

5.1 Proposed Method

To address the issue of low-quality medical data in real-world settings, a novel framework called AMI-Net3 is presented. The approach begins by introducing a unique methodology to transform and standardize the raw data. Consequently, each patient, denoted as (χ, γ), can be represented as a collection of n observable feature-value pairs, i.e., $\chi = (f_1, v_1), (f_2, v_2), \ldots, (f_n, v_n)$, where $v_j \in \mathbb{R}$ and f_j $(j = 1, 2, \ldots, n)$ represents a binary, nominal, ordinal, or continuous feature. The objective of the research is to train a classifier that can accurately predict $\gamma = 0, 1$ based on the input set χ.

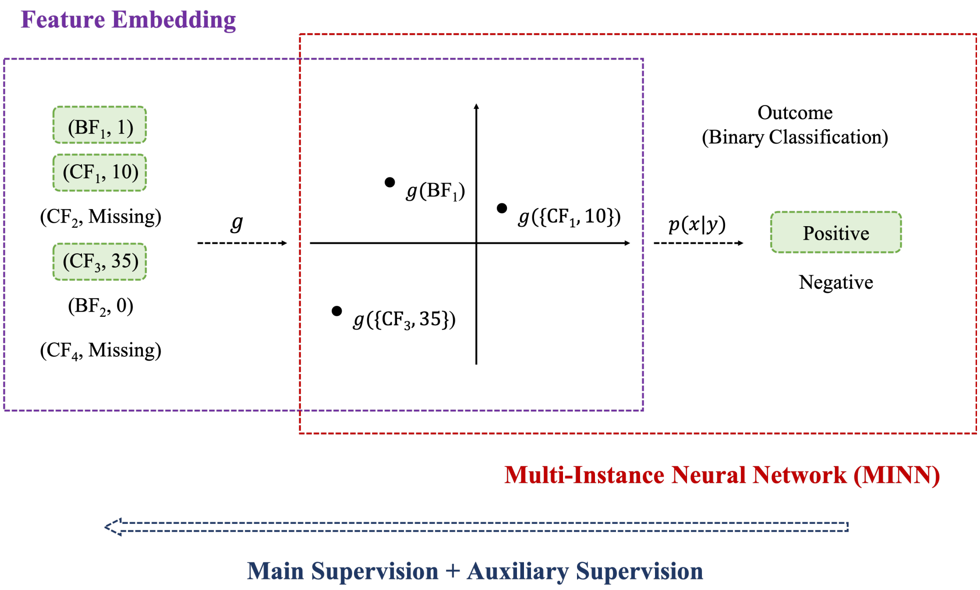


FIGURE 5.1: Continuous and ordinal features are referred to as CF, while binary features or those obtained through one-hot decoding from nominal features, such as headache, diagnosis, and medication history, are denoted as BF. The proposed feature embedding technique enables the method to effectively convey a comprehensive patient narrative in an informative embedding space using only observed data. Subsequently, the MINN algorithm analyzes the available information and generates the final output. The entire process is trained concurrently with both main and auxiliary supervision to ensure optimal performance.

The approach adopts the Multi-Instance Learning (MIL) paradigm, treating χ as a bag with the corresponding label γ, and the observed feature-value pairs as instances within the bag. To model this, a two-level strategy is employed. The first level involves generating embedding vectors, denoted as $G_f = G(f_j, v_j) \in \mathbb{R}^d$, of d dimensions for each instance using the proposed feature embedding method. The second level introduces a novel Multi-Instance Neural Network (MINN) that aggregates valuable information from the instances to compute the bag probability $p(\chi|\gamma)$. These two modeling components are parameterized and trained jointly in an end-to-end manner. The overall architecture is illustrated in Figure 5.2.

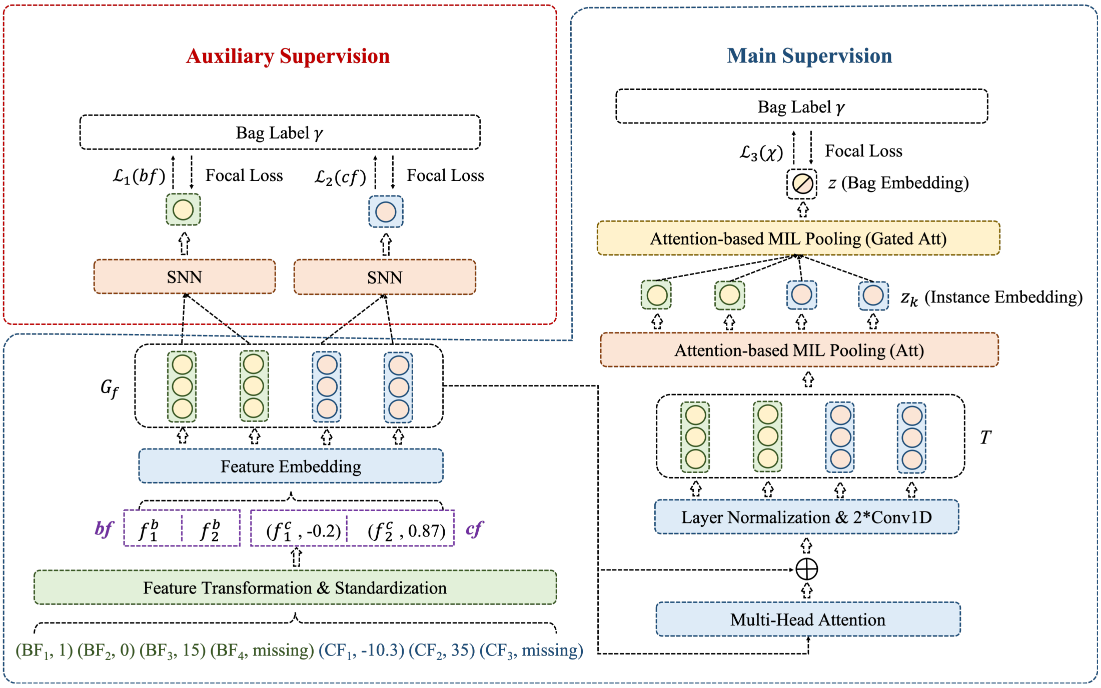


FIGURE 5.2: The comprehensive architecture of AMI-Net3 comprises two components for model training: auxiliary supervision and main supervision. Auxiliary supervision involves two shallow neural networks specifically designed for bf and cf respectively. On the other hand, the main supervision part consists of the primary computational modules, where 2*Conv1D represents the application of two-layer convolutions. Both components utilize the focal loss function to optimize the performance of AMI-Net3.

5.1.1 Feature Transformation and Standardization

AMI-Net3, the proposed model, primarily learns from static datasets, specifically tabular data, where the feature vectors are all 1-dimensional and can be either continuous or discrete. To ensure the suitability of these features for AMI-Net3, two strategies are used for feature transformation and standardization, taking into account their distinct types. Refer to Figure 5.3 for an illustration of these strategies.

- **Binary and Nominal Features (BF):** In order to incorporate nominal features into the analysis, a two-step process is followed. Initially, the nominal features are transformed into binary representations using one-hot encoding. These

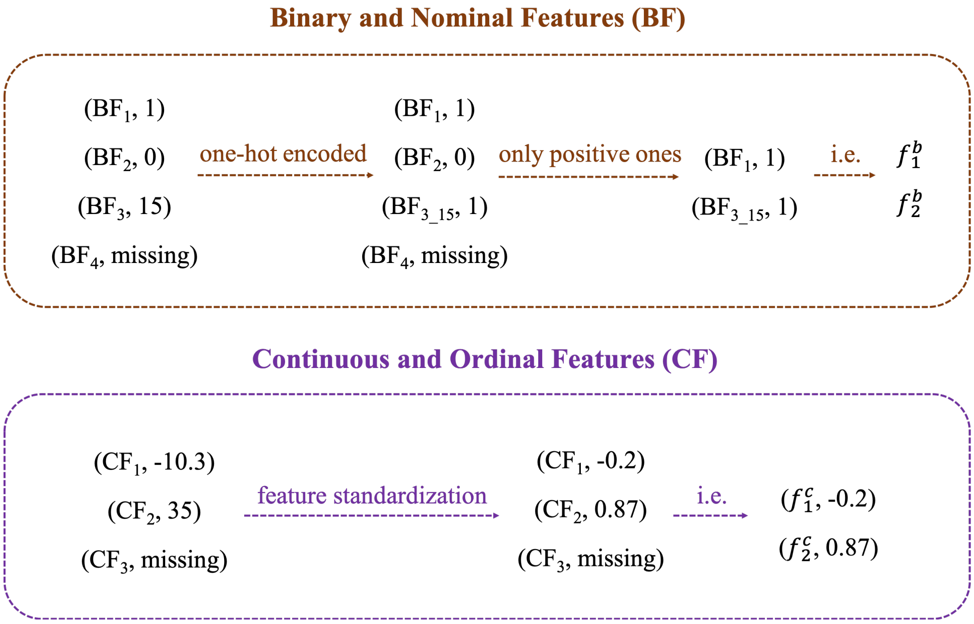
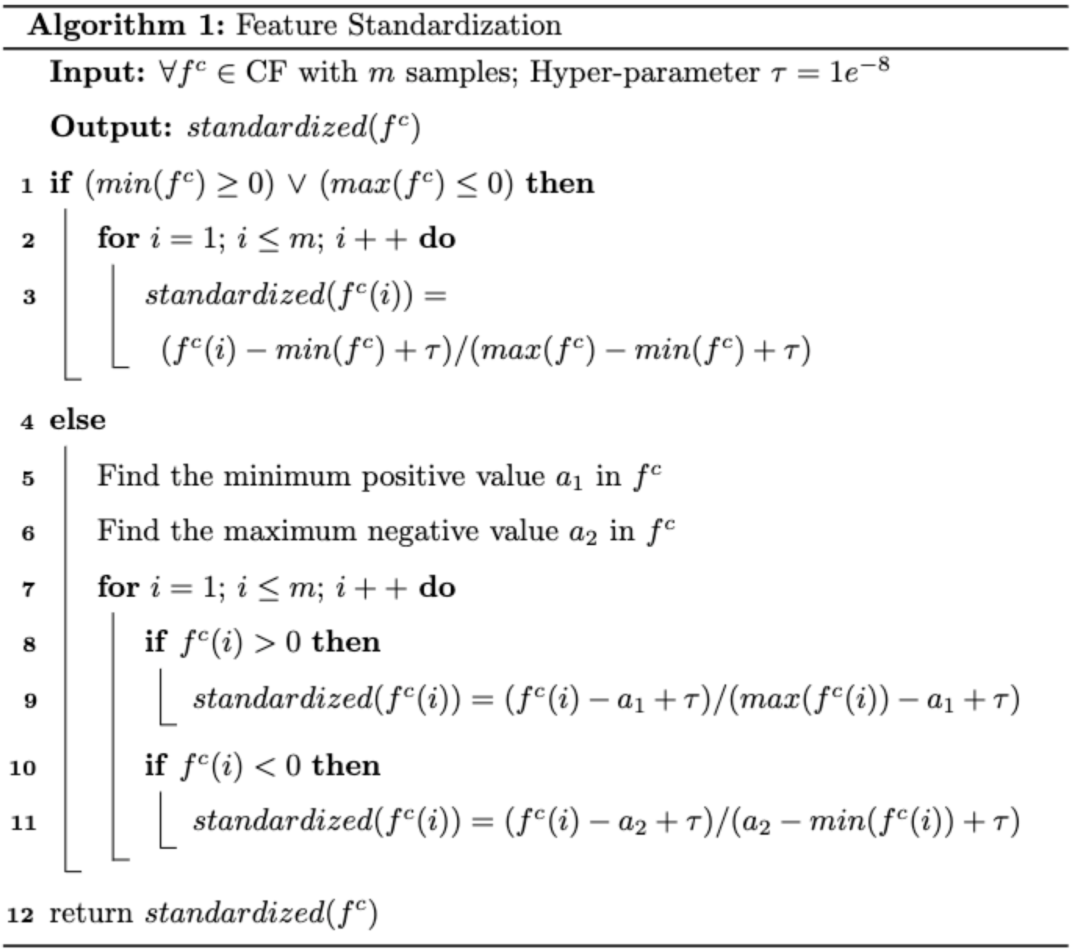


FIGURE 5.3: To illustrate the process of feature transformation and standardization for both binary features (BF) and continuous/ordinal features (CF), a concrete example is provided. It is worth noting that, in this analysis, only BF entries that have received positive responses are considered for further investigation.

newly created binary features are then combined with the existing binary features. The combined set is denoted as $f_1^b, v_1^b, f_2^b, v_2^b, \ldots, f_{n'}^b, v_{n'}^b$, where $v^b \in 0, 1$ and n' represents the number of features in the binary feature set (BF). For each patient, the value v^b determines the inclusion or omission of the corresponding binary feature f^b. Specifically, if $v^b = 0$, f^b is excluded from the feature set. Conversely, if $v^b = 1$, the pair f^b, v^b is replaced simply by f^b. This process ensures that only relevant binary features are retained for further analysis, taking into account their corresponding values.

- **Continuous and Ordinal Features (CF):** Regarding the continuous and ordinal instances, a designed feature standardization algorithm is employed to ensure uniform scaling across the data. This algorithm effectively standardizes the instances to a common scale ranging from -1 to 1. (Refer to Algorithm 1 for a detailed description of the standardization process.) It is important to note



```
Algorithm 1: Feature Standardization
  Input: $\forall f^c \in$ CF with $m$ samples; Hyper-parameter $\tau = 1e^{-8}$
  Output: $standardized(f^c)$
if $(min(f^c) \geq 0) \vee (max(f^c) \leq 0)$ then
    for $i = 1$; $i \leq m$; $i++$ do
        $standardized(f^c(i)) =$
        $(f^c(i) - min(f^c) + \tau)/(max(f^c) - min(f^c) + \tau)$
else
    Find the minimum positive value $a_1$ in $f^c$
    Find the maximum negative value $a_2$ in $f^c$
    for $i = 1$; $i \leq m$; $i++$ do
        if $f^c(i) > 0$ then
            $standardized(f^c(i)) = (f^c(i) - a_1 + \tau)/(max(f^c(i)) - a_1 + \tau)$
        if $f^c(i) < 0$ then
            $standardized(f^c(i)) = (f^c(i) - a_2 + \tau)/(a_2 - min(f^c(i)) + \tau)$
return $standardized(f^c)$
```

that missing values are not considered during the computation. Furthermore, to prevent any potential issues resulting from subtracting two elements that are both equal to 0 within each formula, a hyper-parameter denoted as $\tau = 1e^{-8}$ is introduced. This parameter serves the purpose of ensuring stability and mitigating any potential mathematical complications that could arise during the standardization process.

Following the process of feature transformation and standardization, each patient (χ, γ) can be represented as a collection of observed feature-value pairs, forming a bag of instances denoted as $\chi = f_1^b, f_2^b, \ldots, f_{n_1}^b, (f_1^c, v_1), (f_2^c, v_2), \ldots, (f_{n_2}^c, v_{n_2})$. The bag label, γ, is a binary value indicating the class label, taking values of either 0 or 1.

In this representation, $f_{n_1}^b$ belongs to the binary feature set (BF), f_{n2}^c corresponds to the continuous and ordinal feature set (CF), v_{n_2} represents the respective feature value which can be a real number, and n_1 and n_2 denote the number of instances from the BF and CF, respectively. This standardized format serves as the input structure for AMI-Net3 model, providing a consistent and unified representation of patient data.

5.1.2 Feature Embedding

The process of feature embedding, visually demonstrated in Figure 5.4, is crafted to compute a unique parameter vector, more specifically referred to as the 'embedding vector', for each instance nestled within the bag $\chi = f_1^b, f_2^b, \ldots, f_{n_1}^b, (f_1^c, v_1), (f_2^c, v_2), \ldots, (f_{n_2}^c, v_{n_2})$. Here, each embedded parameter is construed as a distinct attribute of the respective instance.

This methodology parallels feature transformation and standardization approaches in which $bf = f_j^b j = 1^{n1}$ and $cf = f_j^c, v_j j = 1^{n2}$ are separately addressed using two divergent strategies. For any $f^b \in bf$, a dense vector is systematically parameterized with d dimensions:

$$g(f^b) = L_{f^b} \tag{5.1}$$

where $L_{f^b} = \{w_1, w_2, \ldots, w_d\}$ and $\forall w_j \in L_{f^b}, w_j \in \mathbb{R}$. About $(f^c, v) \in cf$, it is mapped to an embedding space of d dimensions using the following transformation:

$$g(f^c, v) = W^c(vL_{f^c}/d) \tag{5.2}$$

where $W^c \in \mathbb{R}^{d \times d}$, $L_{f^c} = \{w_1, w_2, \ldots, w_d\}$ and $\forall w_j \in L_{f^c}, w_j \in \mathbb{R}$. W^c designates a weight matrix specifically constructed for the task of controlling distribution, which judiciously allocates disparate attention weights across instance attributes. This allocation, in turn, modulates their individual contributions towards a more effective depiction of the instance. Concurrently, L_{f^c} represents a parameter vector, configured to signify the presence of f^c, and then incorporates a multiplication by v to indicate its value. Within this context, the dimension d serves as a scaling factor, playing a crucial role in stabilizing the back propagation during the training process.

Once the embedding vectors from bf and cf have been obtained, they are combined and further processed to generate the output of the feature embedding module. To ensure appropriate normalization, layer normalization [11] is applied to the combined vectors. This step helps to stabilize and standardize the output, facilitating subsequent analysis and processing of the feature embeddings.

$$G_f = \text{LayerNorm}\left(\left[g(f^b), g(f^c, v)\right]\right) \tag{5.3}$$

The technique of feature embedding, as applied in AMI-Net3 methodology, exhibits superior flexibility and efficacy. It is uniquely equipped to handle different feature types, and adeptly encodes all discernible information into an embedding space.

Furthermore, it's important to note the parallels between AMI-Net3 and prevalent Natural Language Processing (NLP) methodologies. In many current NLP practices, input words are represented by embedding vectors through techniques such as Word2Vec [170] or various pre-trained models [1]. These representations serve as the basis for

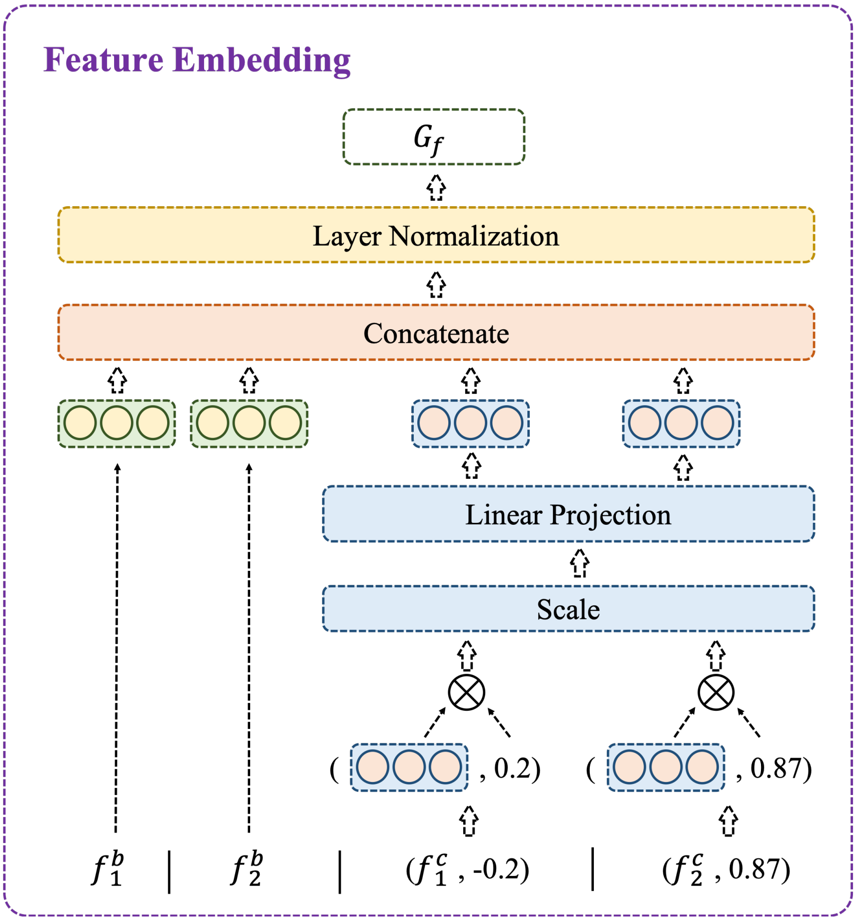


FIGURE 5.4: The objective of the feature embedding module is to assign a unique parameter vector to each input instance. These parameter vectors represent the instance attributes and are subsequently subjected to multi-head attention for further processing. The aim is to capture the specific characteristics of each instance and enable effective attention-based operations to extract meaningful information from the input data.

subsequent processing stages, mirroring the method employed. Thus, the feature embedding technique demonstrates a promising capacity to synergize with NLP techniques. A prime example is multi-head attention, a tool seamlessly integrated into the methodology to reveal concealed correlations between instances.

5.1.3 Gated Attention-based MIL Pooling

As previously discussed, trainable Multi-Instance Learning (MIL) pooling methods provide an effective means of consolidating instance-level or bag-level information within the neural network. As such, in the proposed Multi-Instance Neural Network (MINN), attention-based MIL pooling is incorporated. This method, heralded as a cutting-edge trainable pooling approach [118], offers notable benefits.

The primary objective of gated attention-based MIL pooling is to assign weights and compute the weighted sum across the attributes of an instance (that is, the instance embedding) or the instances contained within a bag (i.e., bag embedding). Let's consider an example bag T, consisting of K instances resulting from previous operations, such that $T = t_1, t_2, \ldots, t_K$. Each instance t_k is defined as a set of attributes $t_k = t_{k,1}, t_{k,2}, \ldots, t_{k,S}$ where k spans from 1 to K and S represents the total number of instance attributes. Utilizing the attention mechanism (Att), the instance embedding z_k is computed as follows:

$$z_k = \sum_{j=1}^{S} a_{k,s} t_{k,s} \tag{5.4}$$

$$a_{k,s} = \frac{\exp\left\{\tanh\left(w_1 t_{k,s}\right) w_2\right\}}{\sum_{s'=1}^{S} \exp\left\{\tanh\left(w_1 t_{s'}\right) w_2\right\}} \tag{5.5}$$

where w_1, w_2 are parameters. Furthermore, the bag embedding z can be obtained by introducing an additional gate mechanism (Gated Att) [57]:

$$z = \sum_{k=1}^{K} a_k z_k \tag{5.6}$$

$$a_k = \frac{\exp\left\{\left(\tanh\left(w_3 z_k\right) \odot \operatorname{sigmoid}\left(w_4 z_k\right)\right) w_5\right\}}{\sum_{k'=1}^{K} \exp\left\{\left(\tanh\left(w_3 z_{k'}\right) \odot \operatorname{sigmoid}\left(w_4 z_{k'}\right)\right) w_5\right\}} \tag{5.7}$$

where w_3, w_4, w_5 are also parameters, optimized by neural network. The element-wise multiplication $\odot$ and $sigmoid$ function form the gate mechanism that improves the non-linearity learning ability via information flow controlling and data adjusting. Also, it eliminates the troubling linearity that the $tanh$ function brings [57].

During the computation of bag embeddings, attention-based MIL pooling is utilized to assign higher weights to instances that are more likely to be positive. This mechanism not only enables the selection of important feature-value pairs but also enhances the interpretability of the results generated by MINN. This interpretability is crucial for clinical risk prediction tasks where understanding the contributing factors is necessary.

Finally, the bag embedding z is fed into a sigmoid function to predict the positive probability $p(\chi|\gamma)$. This step ensures that the output probability is within the range of 0 to 1, providing a meaningful and interpretable prediction of the likelihood of positive instances in the bag.

5.1.4 Model Training

The objective of model learning in AMI-Net3 is to optimize the selection of parameter matrices, which are initially randomly initialized, in both the feature embedding and MINN modules. To enhance the learning process, AMI-Net3 is trained using two complementary strategies: the main supervision and a novel approach named as auxiliary supervision. Furthermore, to address the challenge of imbalanced data, AMI-Net3 employs the focal loss as the chosen loss function. This loss function effectively handles the imbalanced nature of the data and aids in achieving improved performance for learning task.

5.1.4.1 Auxiliary Supervision

The concept of auxiliary supervision is borrowed from the multiple teacher network methodology [279], where both auxiliary and main supervision act as dual educators, offering a comprehensive guidance system for model training. Unlike main supervision, which is implemented across the entire framework, auxiliary supervision is targeted specifically at the feature embedding portion of the model. This focus helps to expedite feedback turnaround and aids in the learning of more suitable embedding weights.

Furthermore, in the realm of auxiliary supervision, separate treatment is maintained for bf and cf. Distinct supervisions for each are provided by utilizing two separate shallow neural networks (SNNs). This approach allows for a more customized and effective supervisory system:

$$\mathrm{SSN}(x) = (\text{Flatten}\left(xW_1 + b_1\right)W_2 + b_2)W_3 + b_3 \tag{5.8}$$

where $W_1, W_2 \in \mathbb{R}^{d\times(d/4)}$ and $W_3 \in \mathbb{R}^{(d/4)\times 1}$.

At last, to optimize AMI-Net3, the auxiliary and main supervisions lead to the following training loss function:

$$\mathcal{L}\left(\chi\right) = \delta\mathcal{L}_1\left(bf\right) + \eta\mathcal{L}_2\left(cf\right) + \mu\mathcal{L}_3\left(\chi\right) \tag{5.9}$$

where $\chi = [bf, cf]$, $bf = \{f_j^b\}_{j=1}^{n_1}$ and $cf = \left\{f_j^c, v_j\right\}_{j=1}^{n_2}$. To fine-tune the model's attention on different sub-tasks, three balancing factors, namely δ, μ, and η, are introduced. These factors play a crucial role in adjusting the model's focus during training. Moreover, to optimize the model's performance, three loss functions, denoted as $\mathcal{L}1$, $\mathcal{L}2$, and $\mathcal{L}_3$, are computed using the focal loss. This choice of loss functions allows for effective handling of imbalanced data and facilitates the achievement of desired results across multiple sub-tasks.

5.1.4.2 Focal Loss

Focal loss, as proposed by Lin et al. [144], was originally designed to address the issue of severe class imbalance commonly encountered in object detection tasks. It accomplishes this by reformulating the standard cross-entropy loss function, consequently reducing the influence of well-classified instances while assigning greater importance to those instances that are difficult to classify. Guided by this approach, focal loss is employed in the optimization of AMI-Net3, aiming to enhance the model's proficiency in the detection of positive bags.

Following each forward propagation, the bag probability, represented as $p\left(\chi|\gamma\right)$, is derived either from the main supervision or the auxiliary supervision, utilizing the ground truth label γ. For the scope of this work, focal loss is implemented for binary classification. The execution of this process is as follows:

$$p_t = \begin{cases} p\left(\chi|\gamma\right) & \text{if } \gamma = 1 \\ 1 - p\left(\chi|\gamma\right) & \text{if } \gamma = 0 \end{cases} \tag{5.10}$$

$$\text{FocalLoss} = -\alpha\left(1 - p_t\right)^{\gamma} \log\left(p_t\right) \tag{5.11}$$

where α is a weighting factor, balancing the importance of positive and negative labels. Additionally, the term $(1 - p_t)^{\gamma}$ serves as a modulating element, where $\gamma \geq 0$ is a tunable parameter. This element effectively reduces the loss contribution of examples that are relatively easier to classify. By incorporating this mechanism, the impact of easily classified examples on the overall loss is attenuated, allowing the model to focus more on challenging instances and improving its overall performance.

5.2 Experiments

5.2.1 Data Description

Distinct from AMI-Net and AMI-Net+, which are confined to modeling binary features, AMI-Net3 demonstrates versatility in handling multiple feature types. Thus, to ensure an equitable evaluation of the proposed methodology, three static inpatient datasets, collected by Medicinovo Inc. from hospitals in Beijing and Shanghai, China, are employed. These datasets incorporate non-sequential attributes such as diagnosis codes, demographics, and physical test results at the time of admission. Each dataset aligns with a separate clinical risk prediction task, namely, adverse drug reaction of risperidone (ADR), schizophrenia relapse (SR), and invasive fungi infection (IFI). Unlike its predecessors that exclusively process binary features, the AMI-Net3 model possesses the capability to process a variety of feature types concurrently. Consequently, the datasets utilized in AMI-Net3 experiments are newly assembled, distinguishing them from those utilized in the AMI-Net and AMI-Net+ experiments. Despite these differences, all datasets retain a common trait: they epitomize low-quality, real-world data beset by comparable data quality challenges.

The datasets originate from real-world settings and exhibit several low-quality issues, including:

- **Extreme Class Imbalance:** The positive class rates for the three datasets are only 0.10, 0.23, and 0.03, respectively.
- **High-Dimensional Feature Space:** Each of the three datasets contains a large number of features that exceed the true number of features relevant for clinical risk prediction tasks. Specifically, the datasets consist of 82, 614, and 67 features, respectively.

- **Incomplete Data:** The average rates of missing values in the feature vectors are 14.5%, 86.3%, and 35.4% for the three datasets, respectively. Furthermore, within each patient, the maximum observed features account for 57, 72, and 34 out of the total 82, 614, and 67 features, respectively.

The detailed statistics of three data sets are shown in Table 5.1.

TABLE 5.1: Statistics of Data sets

Data set	**ADR**	**SR**	**IFI**
Total Patients	5644	1032	4899
Positive Labels	548	240	136
Negative Labels	5096	792	4763
Total Features	82	614	67
Number of BF	33	525	50
Number of CF	49	89	17
Max. Observable Features	57	72	34
Avg. Missing Rate	14.5%	86.3%	35.4%

5.2.2 Experimental Setup

Upon transformation from raw data, instances within each bag are symbolized by embedding vectors comprising 512 dimensions, with their interrelations apprehended through a multi-head attention configuration consisting of eight heads. Under the purview of main supervision, the two-layer convolutional operation comprises of 256 and 128 dimensions respectively. With regards to the auxiliary supervision via the Shallow Neural Network (SNN), both layers maintain hidden sizes of 128 dimensions.

Focal loss parameters are denoted as α and γ, with values set at 0.25 and 2 respectively. The training loss function incorporates weight balancing factors δ, μ, and η, assigned the values 0.2, 0.3, and 0.5 respectively. Ultimately, the proposed method is optimized using the Adam optimization algorithm [131], setting ϵ to $1e^{-8}$, with the momentum parameters β_1 and β_2 defined at 0.9 and 0.98.

In terms of the application of the proposed method across the three datasets mentioned earlier, parameters remain consistent barring variations in the number of epochs and the learning rate. For the Adverse Drug Reaction (ADR), Schizophrenia Relapse (SR), and Invasive Fungal Infection (IFI) datasets, the respective epoch, learning rate pairs are as follows: $600, 1e^{-6}$, $300, 1e^{-5}$, and $300, 1e^{-6}$, with a batch size of 64 being maintained. To ensure an equitable comparison, a 5-fold cross-validation is utilized, in tandem with the implementation of "early stopping" during training as dictated by the AUC score.

5.2.3 Baseline Methods

To validate the effectiveness of AMI-Net3, it is compared with a series of baseline methodologies, namely:

- **Without Data Imputation:** Missing values can be considered as a condition in developing decision rules, and under this assumption, we implement three advanced tree based methods for comparison, LightGBM, XGBoost, and CatBoost. To optimise their performance, these methods are paired with an *automated machine learning (AutoML)*[1] strategy [242], ensuring automatic hyper-parameter selection.
- **With Data Imputation:** In this scenario, missing values are addressed using a variety of imputation techniques including zero, median, mice [36],

[1] https://github.com/ClimbsRocks/auto_ml

random forest (RF) [186], and KNN [17]. Subsequently, *AutoML*[2] is independently applied to the three datasets to identify optimal base models and their corresponding hyper-parameter sets.

- **MINNs with Feature Embedding:** Beyond the classification methods mentioned, AMI-Net3 is further contrasted against leading-edge Multi-Instance Neural Networks (MINNs), including mi-Net, MI-Net [259], Att-Net and Gated Att-Net [118]. In order to facilitate their operation, these networks are integrated with the proposed feature embedding approach. All leverage the focal loss function, boosting their resilience to imbalanced data, thus elevating their competitiveness.

5.3 Results and Analysis

5.3.1 Performance on Clinical Risk Prediction

The comparative results of AMI-Net3 with baseline methods across three clinical risk prediction tasks are presented in Table 5.2. Considering the substantial imbalance in the three training datasets, AUC and F1-score are employed as the evaluation metrics for all models to gauge their overall performance and capacity to identify positive samples, respectively.

As depicted, AMI-Net3 consistently outperforms all baseline methodologies. Navigating imbalanced datasets to achieve a balanced focus on positive and negative samples is challenging. However, the method exhibits superior capability in this aspect, as evidenced by its F1-score that significantly surpasses that of the other methods. Interestingly, models operating without data imputation fare better than those implementing

[2] https://github.com/EpistasisLab/tpot

imputation, suggesting that imputation may introduce significant bias when dealing with highly incomplete data and is, therefore, less desirable. In the context of MINNs, their performance suffers due to the underutilization of information stemming from the sole employment of either instance embedding or bag embedding in MIL pooling. However, the "Auto-XGBoost," "Auto-LightGBM," and "Auto-CatBoost" methodologies yield AUC scores comparable to AMI-Net3. Yet, their F1-scores fall short of matching ours, indicating that while advanced tree-based methods possess adequate internal mechanisms to confront the challenges in low-quality medical data, they could serve as alternative solutions under certain conditions.

In conclusion, it can be confidently asserted that the proposed AMI-Net3 exhibits superior efficacy and robustness in learning from low-quality medical data, thereby outclassing the other baseline approaches.

TABLE 5.2: Comparison with Baseline Methods on Three Low Quality Medical Data sets (95% CI)

Strategy	**Model**	**SR**		**IFI**		**ADR**	
		AUC	**F1-score**	**AUC**	**F1-score**	**AUC**	**F1-score**
With Data Imputation	Zero-AutoML	0.558 ± 0.082	0.197 ± 0.061	0.610 ± 0.093	0.294 ± 0.063	0.572 ± 0.096	0.246 ± 0.055
	Median-AutoML	0.557 ± 0.083	0.208 ± 0.057	0.613 ± 0.097	0.289 ± 0.043	0.587 ± 0.106	0.285 ± 0.061
	KNN-AutoML	0.540 ± 0.074	0.155 ± 0.042	0.651 ± 0.089	0.401 ± 0.060	0.595 ± 0.093	0.302 ± 0.049
	Mice-AutoML	0.554 ± 0.112	0.187 ± 0.064	0.614 ± 0.108	0.294 ± 0.067	0.586 ± 0.092	0.285 ± 0.051
	RF-AutoML	0.568 ± 0.082	0.261 ± 0.043	0.625 ± 0.089	0.318 ± 0.046	0.569 ± 0.093	0.238 ± 0.058
Without Data Imputation	Auto-XGBoost	0.669 ± 0.073	0.360 ± 0.036	0.924 ± 0.065	0.346 ± 0.029	0.844 ± 0.071	0.282 ± 0.035
	Auto-LightGBM	0.704 ± 0.065	0.350 ± 0.023	0.930 ± 0.061	0.317 ± 0.026	$\mathbf{0.845 \pm 0.064}$	0.323 ± 0.019
	Auto-CatBoost	0.688 ± 0.088	0.404 ± 0.053	0.920 ± 0.072	0.373 ± 0.037	0.802 ± 0.091	0.136 ± 0.040
MINNs with Feature Embedding	mi-Net	0.648 ± 0.052	0.385 ± 0.021	0.876 ± 0.034	0.293 ± 0.052	0.652 ± 0.046	0.265 ± 0.025
	MI-Net	0.612 ± 0.031	0.394 ± 0.017	0.843 ± 0.023	0.244 ± 0.026	0.790 ± 0.044	0.360 ± 0.031
	Att-Net	0.610 ± 0.023	0.375 ± 0.019	0.863 ± 0.020	0.236 ± 0.017	0.795 ± 0.022	0.393 ± 0.018
	Gated Att-Net	0.629 ± 0.027	0.400 ± 0.032	0.853 ± 0.038	0.227 ± 0.018	0.796 ± 0.026	0.388 ± 0.029
Our Method	AMI-Net3	$\mathbf{0.716 \pm 0.014}$	$\mathbf{0.468 \pm 0.024}$	$\mathbf{0.937 \pm 0.026}$	$\mathbf{0.433 \pm 0.012}$	0.837 ± 0.028	$\mathbf{0.448 \pm 0.017}$

5.3.2 Evaluation of Integrated Modules in MINN

One of the main contributions in this work is the approach taken to address the challenges of class imbalance and highly correlated features. In the designed MINN, focal loss and multi-head attention are employed to address these challenges and improve the overall model performance. Experiments are conducted on AMI-Net3 with and without these components to evaluate their impact and performance. It is important to note that the necessity of focal loss is tested by replacing it with the commonly used cross-entropy.

As shown in Figure 5.5, when multi-head attention is removed from AMI-Net3, feature correlations cannot be captured anymore, resulting in a significant drop in AUC and F1-score, particularly on the SR and IFI datasets. Moreover, focal loss demonstrates its efficacy in overcoming the challenge of severe class imbalance when compared to cross-entropy. Replacing focal loss with cross-entropy leads to a drop of over 50% in the average F1-scores across the three datasets, highlighting the benefits of focal loss. Additionally, auxiliary supervision helps to improve the AUC and F1-score, although the improvement is not significant.

Overall, the comparison results indicate that the computational modules used, including multi-head attention, auxiliary supervision, and focal loss, each play a crucial role in enhancing the predictive performance.

A key innovation in this research lies in effectively addressing the challenges of class imbalance and high feature correlation. This is achieved by incorporating focal loss and multi-head attention in the uniquely designed MINN, while also utilizing auxiliary supervision to strengthen the overall model performance. In order to evaluate the impact of these components and their operational efficiency, experiments were conducted on

AMI-Net3 with and without these elements. It is important to note that the necessity of focal loss was assessed by substituting it with the conventional cross-entropy.

Figure 5.5 illustrates that when multi-head attention is excluded from AMI-Net3, the model fails to capture feature correlations, resulting in a significant decrease in AUC and F1-score, particularly in the SR and IFI datasets. Additionally, focal loss proves to be highly effective in addressing the challenge of severe class imbalance compared to cross-entropy. Replacing focal loss with cross-entropy leads to an average reduction of more than 50% in F1-scores across the three datasets, highlighting the advantages of focal loss. Furthermore, auxiliary supervision contributes to improvements in the AUC and F1-score, although the improvements are not substantial.

In summary, the comparative results indicate that the computational modules utilized in this research, including multi-head attention, auxiliary supervision, and focal loss, each play a significant role in enhancing the predictive performance.

5.3.3 Scenario Analysis: Employing BF Exclusively

Prior research [260, 263] has established the effectiveness of feature embedding methods for binary feature (BF) representations. To assess the impact of feature embedding on continuous/ordinal feature (CF) representations and its effect on model performance, we conduct experiments by excluding all CF during the training of AMI-Net3, and observe any potential changes in AUC and F1-score.

Figure 5.6 presents a comparative analysis, indicating that the use of feature embedding for both BF and CF significantly improves the ability of AMI-Net3 to capture the comprehensive patient narrative, compared to using only BF as the input. This is particularly evident in the ADR dataset, where the proportion of CF is considerably

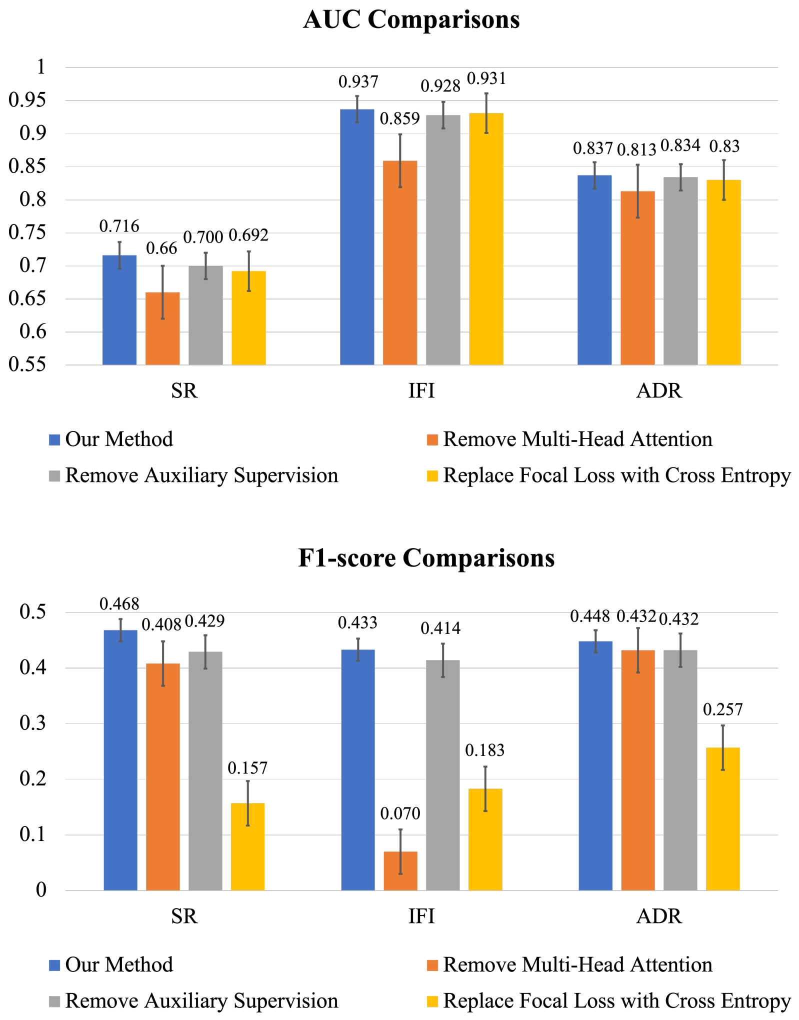
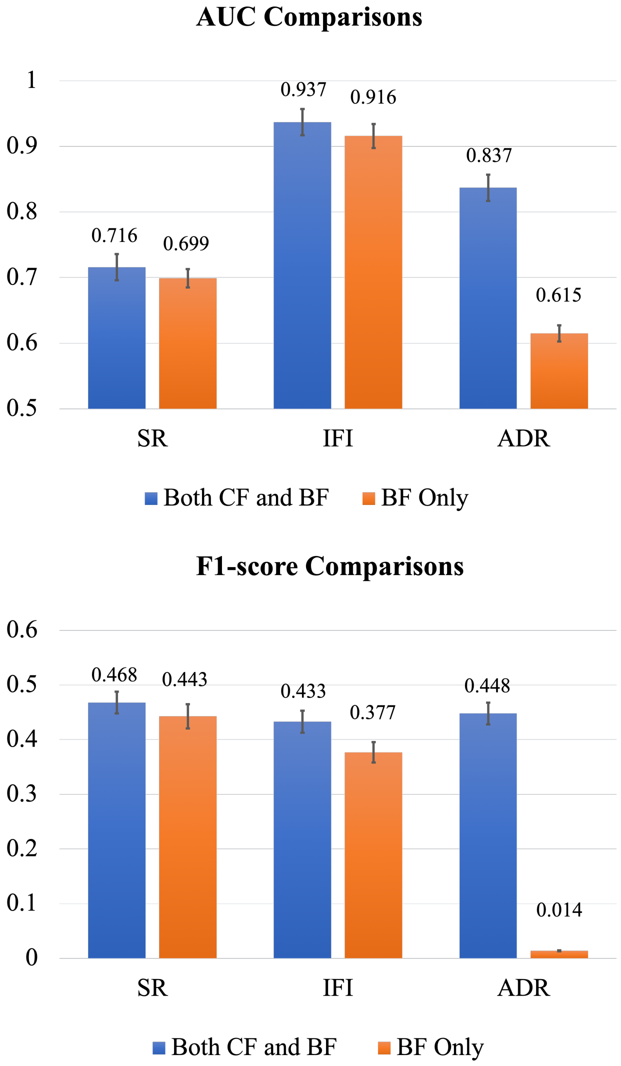


FIGURE 5.5: We run experiments to evaluate the value of the computational modules we integrate in MINN. (95% CI)

higher than the other two datasets. When only BF is utilized, the model performance suffers a notable decline. These findings support the effectiveness of a flexible approach and demonstrate that the proposed feature embedding effectively represents observed features in an information-rich embedding space, thus making progress in the right direction. In the IFI and SR datasets, where the CF ratios are only 25.4% and 14.5% respectively, and BF provides sufficient feature information for predictions, selecting BF alone as the input results in only a marginal decline in model performance.



AUC Comparisons

F1-score Comparisons

FIGURE 5.6: The experiment is to evaluate the effectiveness of the proposed feature embedding. (95% CI)

5.3.4 Assessing the Influence of MIL Pooling

In the context of extracting instance or bag embeddings, a fully connected layer (FC) can be used as an alternative to MIL pooling when it is equipped with a single output neuron. In order to evaluate the effectiveness of our customized MIL pooling approach, we compare attention-based MIL pooling methods (Att and Gated Att) with FC to obtain instance embeddings (Ins Emb), bag embeddings (Bag Emb), or both. These comparisons are performed on three datasets, and the evaluation metrics used are AUC and F1-score.

The results, as presented in Table 5.3, demonstrate the superior performance of AMI-Net3. This suggests that FC alone does not adequately learn instance and bag representations and lacks the non-linear expressive capacity required for effective modeling. This limitation becomes particularly evident when FC is used to replace both MIL pooling layers, resulting in the least optimal model performance, as indicated by both AUC and F1-score. On average, there is a 4.5% decrease in AUC and an 11.1% reduction in F1-score. Additionally, models that utilize FC for bag embeddings perform worse than those using FC for instance embeddings, indicating the inherent complexity and importance of bag representations. Therefore, the importance of employing a computational module with strong non-linear expressive capacity, such as attention-based MIL pooling, is underscored.

TABLE 5.3: Comparison of MIL Pooling and FC (95% CI)

Ins Emb	**Bag Emb**	**SR**		**IFI**		**ADR**	
		AUC	**F1-score**	**AUC**	**F1-score**	**AUC**	**F1-score**
FC	FC	0.667 ± 0.042	0.412 ± 0.034	0.896 ± 0.040	0.389 ± 0.029	0.816 ± 0.039	0.398 ± 0.033
	Gated Att	0.709 ± 0.035	0.448 ± 0.023	0.928 ± 0.031	0.416 ± 0.026	0.835 ± 0.037	0.444 ± 0.024
Att	FC	0.659 ± 0.038	0.398 ± 0.026	0.915 ± 0.030	0.405 ± 0.035	0.818 ± 0.040	0.431 ± 0.022
	Gated Att (Our Method)	**0.716 ± 0.014**	**0.468 ± 0.024**	**0.937 ± 0.026**	**0.433 ± 0.012**	**0.837 ± 0.028**	**0.448 ± 0.017**

5.3.5 Comparison of Different MIL Pooling Techniques

To corroborate the effectiveness of our chosen MIL pooling techniques, we conduct a comparison of max-pooling (Max), Att, and Gated Att on our MINN. These methods have previously been established as superior in identifying informative instances [118, 259].

Table 5.4 illustrates that the MIL pooling technique utilized in AMI-Net3 outperforms the others. The gated attention mechanism is hampered by the limited data volume in the SR dataset, leading to a performance inferior to that of the simple attention mechanism. However, with an increased data volume, the gating mechanism would significantly boost the capacity to learn complex relationships, as demonstrated in the other two datasets. It is worth noting that max-pooling generally underperforms in comparison to the others, exposing the limitations of non-trainable pooling methods. Although they might be viable options for instance-level MIL methods, they could potentially hinder the computation of bag and instance representations for superior results. As for trainable MIL pooling methods, they excel at learning complex relations and enhance model performance through task- and data-specific adaptations.

TABLE 5.4: Comparison of Different MIL Pooling Methods (95% CI)

Ins Emb	**Bag Emb**	**SR**		**IFI**		**ADR**	
		AUC	**F1-score**	**AUC**	**F1-score**	**AUC**	**F1-score**
Max	Max	0.675 ± 0.069	0.399 ± 0.047	0.929 ± 0.072	0.284 ± 0.049	0.758 ± 0.069	0.416 ± 0.062
	Att	0.712 ± 0.047	0.457 ± 0.038	0.921 ± 0.042	0.344 ± 0.037	0.799 ± 0.051	0.409 ± 0.048
	Gated Att	0.709 ± 0.028	0.455 ± 0.033	0.926 ± 0.040	$\mathbf{0.433 \pm 0.030}$	0.802 ± 0.031	0.373 ± 0.028
Gated Att	Max	0.685 ± 0.042	0.421 ± 0.036	0.933 ± 0.045	0.264 ± 0.029	0.802 ± 0.038	0.446 ± 0.041
	Att	0.680 ± 0.031	0.436 ± 0.026	0.929 ± 0.032	0.368 ± 0.022	0.825 ± 0.037	0.359 ± 0.024
	Gated Att	0.679 ± 0.019	0.451 ± 0.023	0.935 ± 0.030	0.424 ± 0.026	0.832 ± 0.029	0.447 ± 0.023
Att	Max	0.684 ± 0.038	0.423 ± 0.032	0.930 ± 0.034	0.398 ± 0.025	0.815 ± 0.032	0.438 ± 0.026
	Att	0.678 ± 0.021	0.436 ± 0.027	0.924 ± 0.031	0.311 ± 0.024	0.829 ± 0.029	0.444 ± 0.022
	Gated Att (Our Method)	$\mathbf{0.716 \pm 0.014}$	$\mathbf{0.468 \pm 0.024}$	$\mathbf{0.937 \pm 0.026}$	$\mathbf{0.433 \pm 0.012}$	$\mathbf{0.837 \pm 0.028}$	$\mathbf{0.448 \pm 0.017}$

5.4 Limitations

The study encounters certain limitations, particularly regarding the criteria for data collection. In the SR dataset, there is a significant imbalance between the number of binary features (BFs) and continuous features (CFs). This imbalance may hinder effective learning from CFs, thus reducing the sensitivity of predictive models. Future studies could address this issue by curating data based on prior knowledge or specific inclusion and exclusion criteria. Another limitation is the focus on static data, while the omission of temporal data, which is more prevalent in hospital settings and contains valuable information such as frequent lab test results and physical examination records, limits the study's scope. A future direction is to explore a suitable framework for temporal data, enabling continuous and predictive monitoring of clinical outcomes and treatment guidance.

5.5 Summary

AMI-Net3 is introduced as a novel framework for clinical risk prediction using low-quality data. It incorporates a feature embedding module and an innovative MINN, enabling direct learning from observed data without the need for data imputation techniques. The framework effectively addresses challenges related to redundant, correlated features and extreme class imbalance. Comparative experiments conducted on three low-quality medical datasets demonstrate the effectiveness, scalability, and superiority of the proposed method when compared to state-of-the-art MINNs and AutoML methods employing different data imputation strategies.

The main advantage of the AMI-Net3 is its capability to represent patient features in the embedding space through feature embedding. This enables predictive models to effectively utilize incomplete data and allows for flexibility in integrating techniques from various deep learning domains, including CV, NLP, and MINN, for handling real-world medical data. By learning from meta-features defined by their attributes, the proposed framework empowers predictive models to capture essential characteristics. Moreover, the use of feature embedding opens up possibilities for feature selection and interpretation, as MINN and attention mechanism can identify significant features in clinical risk prediction for individual patients. Future work will delve deeper into exploring these aspects.

CHAPTER 6

AMI-Net for Regression

In light of the exceptional learning capacities and predictive results exhibited by the AMI-Net series algorithms through multi-instance learning in the context of low-quality real-world medical data, these algorithms have predominantly been utilized for classification tasks. However, the medical field encompasses numerous regression tasks of considerable importance, such as drug dosage prediction or hospitalization duration estimation. These tasks necessitate predictive models capable of generating continuous output rather than discrete labels. The development of effective models for these tasks is crucial due to their potential impact on patient care.

Recognizing this, the aim is to expand the applicability of the approach by adapting and refining the methodology for regression tasks. The objective extends beyond simply applying the same model architecture to a new task type; it involves exploring how the unique characteristics of regression problems may require adjustments to the model's design or training process. Specifically, considerations are given to the appropriate loss function for regression tasks, the impact of imbalanced datasets on regression performance, and the optimal interpretation of regression model outputs within the medical domain.

To concretize this exploration, the task of warfarin dosage prediction is selected as a practical clinical application. Warfarin, an anticoagulant drug, demands careful dosage management to balance therapeutic effects with the potential for serious side effects.

Accurately predicting the appropriate dosage for individual patients poses a critical challenge, aligning well with the regression-focused approach. Detailed description of this task is provided in the following chapter.

6.1 Background of Warfarin Dose Prediction

Warfarin, an anticoagulant of international repute, plays a pivotal role in managing non-valvular atrial fibrillation and venous thromboembolism [52]. Its global acceptance and recurrent usage in various clinical settings are undeniable, yet the challenges in its administration cannot be understated. One of the most distinctive features of warfarin's pharmacodynamic profile is its narrow therapeutic index, a characteristic that necessitates precise dosing to maintain an effective yet safe level of the drug in the body.

Adding complexity to this delicate balancing act is the broad spectrum of responses displayed by individual patients. The dynamic interplay of genetic factors, lifestyle habits, and concomitant medications often leads to considerable variability in drug metabolism and responsiveness, with a subset of patients exhibiting heightened sensitivity to warfarin [300]. In these warfarin-sensitive patients, standard doses may lead to an increased risk of bleeding, emphasizing the necessity for careful dosage adjustments and close monitoring [108].

Despite the frequent use of International Normalized Ratio (INR) monitoring as a strategy to maintain therapeutic anticoagulation, it is a sobering reality that less than 60% of patients manage to maintain warfarin levels within the therapeutic window [222]. This statistic highlights the complexity of warfarin dose optimization and underscores the urgent need for refined dosing strategies. With this in mind, the task of achieving optimal warfarin dosing lies at the forefront of therapeutic goals, ultimately influencing the balance between efficacy and safety in a clinical scenario [52].

This challenge calls for the development of novel, data-driven approaches capable of deciphering the complexity of warfarin pharmacokinetics and pharmacodynamics. By facilitating individualized dosage adjustments, these strategies aim to bring the majority of patients within the desired therapeutic window, thereby improving the overall safety and efficacy profile of warfarin.

In recent decades, a concentrated effort has been deployed towards formulating comprehensive and effective warfarin dose prediction models. These sophisticated models amalgamate various patient-specific factors to create individualized therapeutic strategies, considering each patient's unique clinical, demographic, and genetic profile [52, 89, 140, 151]. They embody a significant stride towards personalized warfarin dosing, with the aim of improving patient safety and efficacy. Among the multitude of predictive algorithms, multivariate linear regression (MLR) has gained favor due to its ease of implementation and interpretability [52, 72, 73]. However, MLR's inability to effectively model non-linear relationships between predictors and outcomes limits its utility, particularly within the complex non-linear interactions often occurring between demographic, clinical, and genetic factors influencing warfarin response [159]. These constraints could potentially compromise the performance of MLR models in diverse patient populations, potentially leading to inaccurate dose predictions for certain subsets of patients. To circumvent these limitations, researchers have ventured into the more advanced domains of machine learning and deep learning. These evolving computational technologies offer a range of tools such as support vector machines (SVM), decision tree-based algorithms, and neural networks. They provide significant promise in personalized warfarin dosing by excelling at identifying and capturing intricate interrelationships between variables [53, 89, 150]. For example, SVM identifies the hyperplane in an N-dimensional space that classifies data points distinctly, decision tree-based algorithms offer a robust framework for classification and regression tasks, while neural networks emulate the human brain's architecture to learn complex patterns through interconnected nodes.

These methodologies represent a paradigm shift in the modelling of warfarin dosing. Instead of making assumptions about relationships between variables, they learn these relationships directly from the data, allowing them to adapt to the non-linear and interactive nature of the factors affecting warfarin dose [159]. This characteristic, in particular, is expected to increase the accuracy of predictions, especially in patients for whom traditional linear models may fall short.

Through their adept handling of high-dimensional and diverse data, machine learning and deep learning algorithms could pave the way for more precise warfarin dosing models. The implications of such advancements could be far-reaching, enhancing the predictability of anticoagulation control, reducing the risk of adverse events, and ultimately leading to improved patient outcomes.

However, both linear and machine learning models are built upon 1-dimensional vectors in a given feature space, as illustrated in Figure 6.1a. These models routinely face two significant challenges: high dimensionality and missing values [143]. High dimensionality stems from the multitude of variables that characterize each patient's medical history, including their clinical features, medications, and indications. Missing values, on the other hand, arise from the inevitable reality that patients will not undergo all potential examinations during their hospital stay. Addressing these two primary concerns – high dimensionality and missing values – forms the cornerstone of the method proposed in our study.

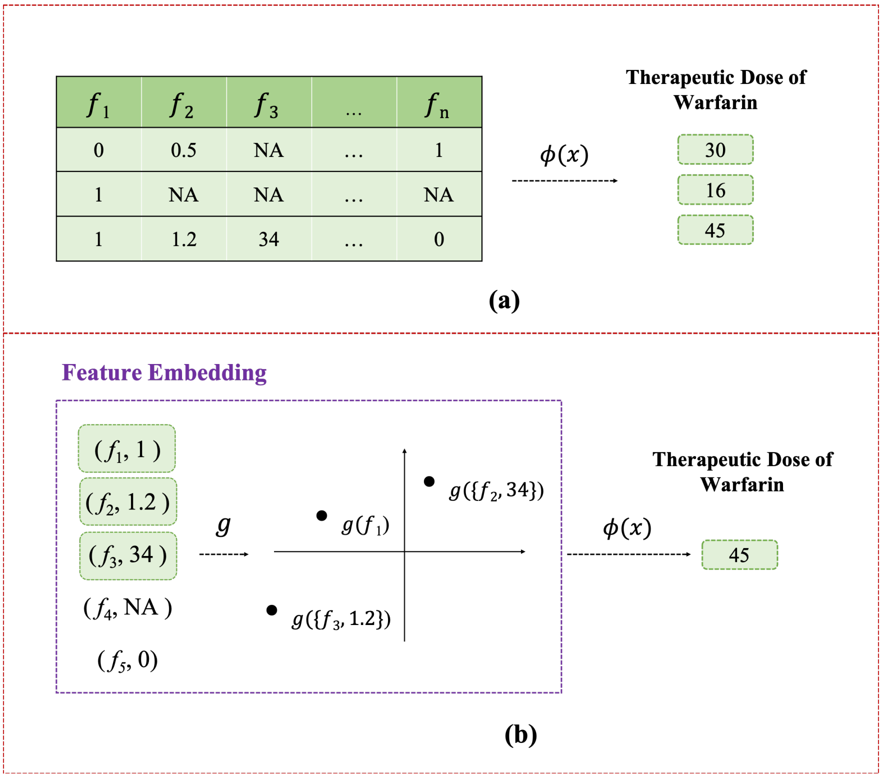


FIGURE 6.1: **(a)** An example of common data processing way. **(b)** An example of how transforming and processing the original data from each individual patient.

This work presents an innovative solution to address the challenges inherent in high-dimensional feature space and missing values. Rather than relying on conventional methodologies, the proposed approach involves treating each patient's record as a distinct combination of observed features and converting them into an embedding space through feature embedding (as illustrated in Figure 6.1b). This strategy enables the direct and seamless handling of incomplete and high-dimensional data without the need for additional preprocessing steps. Furthermore, the capability of AMI-Net approach is extended to encompass this particular regression task.

6.1.1 Methodology

To accurately estimate the appropriate dosage of warfarin, even when dealing with incomplete and high-dimensional data, an innovative framework is proposed. Within this framework, each patient (X, y) is initially converted into a series of observed feature-value pairs, denoted as $X = (f_1, v_1), (f_2, v_2), \ldots, (f_n, v_n)$. These pairs correspond to the optimal warfarin dose y, and each feature f_j $(j = 1, 2, \ldots, n)$ is either binary f_j^b, ordinal, or continuous f_j^c. Crucially, during the transformation process, all nominal features are one-hot encoded into binary ones.

The goal is to train a regressor to predict y based on the set of features X. The strategy for achieving this is twofold. The first level of the proposed approach involves representing each observed feature-pair (f_j, v_j) as a d-dimensional embedded vector $g_f = g(f_j, v_j) \in \mathbb{R}^d$.

The second level employs a novel neural network to discern and capture intricate correlations present within the group G. It also aims to identify valuable information within G for estimating the optimal warfarin dose $y \in \mathbb{R}$. The group G includes g_{rep}, an embedded vector of a representative feature which embodies all observed information - essentially the overall body condition. The rest of G comprises $g(f_1, v_1), g(f_2, v_2), \ldots, g(f_n, v_n)$.

Both these components of the proposed model are parameterized and trained together in a unified, end-to-end fashion. The complete architecture is depicted in Figure 6.2a.

6.1.2 Multi-Head Attention

The multi-head attention module employed closely follows the original definition by Vaswani et al. [255]. However, the position encoding component is excluded in this

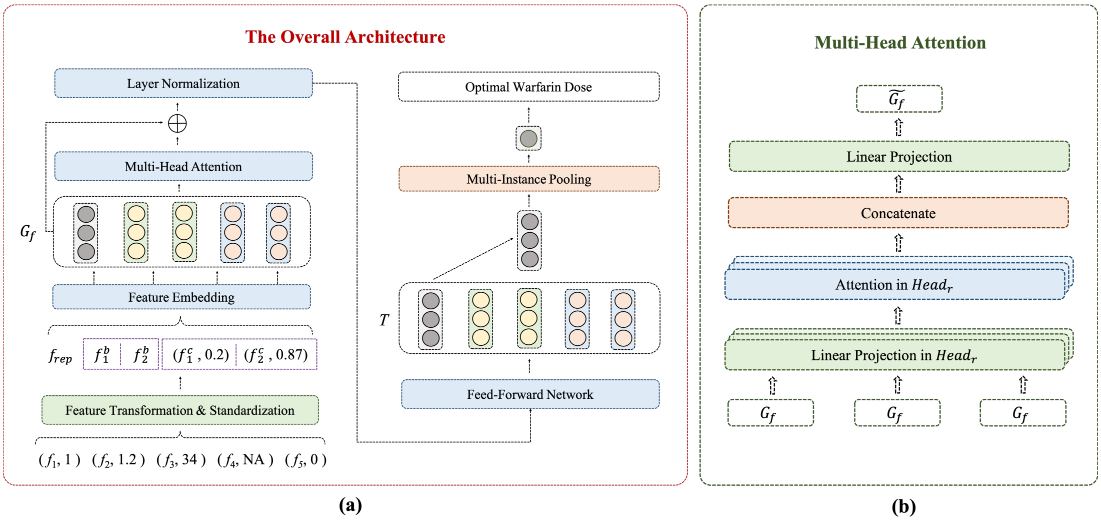


FIGURE 6.2: **(a)** The figure outlines the overall structure of the proposed methodology, which consists of two primary levels. The first level incorporates a feature embedding module designed to convert the observed information into an embedded space. The second level comprises a sophisticated neural network armed with multi-head attention, a feed-forward network, and multi-instance pooling, all intended to decipher and unravel hidden patterns among features for an accurate final prediction. Notably, these two levels are trained in a unified, end-to-end manner. **(b)** The methodology employs multi-head attention to reveal correlations among embedded vectors, i.e., observed features. Importantly, multi-head attention enables the detection of relationships between G_{rep} and other embedded vectors. G_{rep} acts as a representative of overall patient information, facilitating further processing and analysis.

work, as the data being dealt with does not possess sequential information. The multi-head attention module consists of two computational components: scaled dot-product attention and multi-head transformation.

In scaled dot-product attention, three input vectors with dimensions d_k are utilized: a query vector Q, a key vector K, and the corresponding value vector V. The output is obtained through the following computations:

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V \tag{6.1}$$

This research primarily concentrates on uncovering potential associations among observed features. As such, for the computations in question, all the vectors Q, K, and V correspond to the embedded vector G_f obtained from the preceding feature embedding step:

$$\text{Attention}(G_f, G_f, G_f) = \text{softmax}\left(\frac{G_f G_f^T}{\sqrt{d_k}}\right) G_f \tag{6.2}$$

Furthermore, to thoroughly explore the underlying relationships within the embedded vector G_f, multi-head attention is used to access several sub-embedding spaces via multi-head transformation:

$$H_r = \text{ Attention } \left(W_r^1 G_f, W_r^2 G_f, W_3^1 G_f\right) \tag{6.3}$$

where W^r is the output of a single attention head and W_r^1, W_r^2, W_r^3 are three linear projections for G_f. Then they are concatenate as the final output of multi-head attention:

$$\text{MultiHead}(G_f, G_f, G_f) = [H_1; \ldots; H_R]\, W^4 \tag{6.4}$$

where W^4 is the output projection. By adopting this approach, model can effectively explore the associations between the self-defined representation vector g_{rep} and other observed features. This enables g_{rep} to serve as a comprehensive representation of the patient's current medical condition. Furthermore, this representation can be utilized for the estimation of warfarin dosage. The exploration of associations and the utilization of g_{rep} enhance the understanding of the patient's medical state and provide valuable insights for subsequent medical decision-making processes.

6.1.3 Feed-Forward Network

After multi-head attention, the feed-forward network is adopted, consisting of two 1-dimensional convolution layers with kernel size equals to 1 and ReLu activation in between, to further enhance the representation capability of G_f:

$$\mathrm{FFN}(x) = \mathrm{Conv1D}(\max(0, \mathrm{Conv1D}(x))) \tag{6.5}$$

By utilizing a feed-forward network, the final vector set $T = t_{rep}, t_1, t_2, \ldots, t_n$ can be obtained. For the subsequent operation of multi-instance pooling, only the self-defined representation vector t_{rep} with d dimensions is included.

6.1.4 Multi-Instance Pooling

Multi-instance pooling plays a crucial role in the MINN framework by effectively filtering out noise and irrelevant information [263]. In light of this, multi-instance pooling is chosen over the commonly used fully connected layer as the final output layer for predicting the warfarin dose. Furthermore, since the objective is to perform regression rather than classification, it is worth noting that trainable multi-instance pooling methods typically generate weights within the range of 0 and 1, which may be better suited for classification tasks. To address this, a non-trainable approach is used, specifically max pooling, which allows us to aggregate the most influential instances within the bag and obtain a robust prediction for the warfarin dose.

$$\text{Output} = \max(t_k = \{w_1, w_2, \ldots, w_d\}) \tag{6.6}$$

$w_i \in \mathbb{R}$ is a parameter in the representation vector t_k.

Once the final output is obtained, the loss is calculated by comparing it with the true value y. The model is then trained using backpropagation in an end-to-end manner.

6.2 Experiments

6.2.1 Data Description

The modeling data utilized in this study is derived from the publicly available IWPC cohort, previously described by [52]. The IWPC data set can be downloaded from the PharmGKB website (http://www.pharmgkb.org/downloads/). This comprehensive data set comprises information from 6256 warfarin users spanning four continents. It encompasses a range of demographic factors, including age, weight, height, as well as clinical features such as indications, united medication, and genotypes of CYP2C9 and VKORC1. For data integrity and consistency, subjects who did not reach stable warfarin doses and instances with missing therapeutic dose information were excluded. As a result, a total of 5410 subjects were included in the study. The dataset comprises eight indications, 1458 comorbidities, and 1917 medications, leading to a high-dimensional dataset. It is worth noting that the dataset contains a considerable number of missing values, with an average missing rate of 41.8%. Table 6.1 provides a detailed breakdown of the statistics for the included data.

6.2.2 Experimental Settings

Feature-pairs in this approach are mapped to an embedding space of 512 dimensions and processed using multi-head attention with 8 heads. This allows for the exploration of underlying relationships across 8 distinct embedding sub-spaces. In the feed-forward network, two convolution layers with dimensions of 1024 and 512 are incorporated.

TABLE 6.1: Statistics of the Data Set

Included Patients	5410
Binary or Nominal Features	3395
Continuous or Ordinal Features	4
Max. Observed Features	58
Min. Observed Features	5
Average Missing Rate	41.8%
Max. Missing Rate	83.8%

To prevent overfitting, a dropout layer [234] with a dropout rate of 0.3 is included in each computational module. During training, the Adam optimizer [131] is used with a learning rate of 5×10^{-6}, an epsilon value (ϵ) of 1×10^{-8}, and momentum parameters β_1 and β_2 set to 0.9 and 0.98, respectively. Additionally, to ensure a fair comparison, an "early stopping" mechanism is implemented based on five-fold cross-validation, using R^2, MAE (Mean Absolute Error), and MSE (Mean Squared Error) as evaluation metrics. The employed loss function is the Log-Cosh loss, defined as follows:

$$loss(y, f(x)) = \sum_{i=1}^{n} \log\cosh(y_{true} - y_{pred}) \tag{6.7}$$

where y_{true} is the true value and y_{pred} denotes the predicted value on the i^{th} sample.

6.2.3 Baseline Models

To comprehensively evaluate the effectiveness of the proposed method, comparisons are conducted with three advanced machine learning algorithms: XGBoost [46], LightGBM [127], CatBoost [62] and FT-transformer [83]. To optimize the performance of these machine learning methods, an AutoML approach [242] is employed to automatically select the best parameter set. It should be noted that these methods are all decision tree-based algorithms, which allow for direct learning from incomplete data and exhibit robustness to sparse data, addressing two key challenges in the research.

Furthermore, the effectiveness of the chosen multi-instance pooling method is demonstrated by comparing its performance with several alternative pooling methods in the proposed framework. These include fully connected layer, max pooling, mean pooling [263], attention-based pooling, and gated attention-based pooling methods [118]. By conducting these performance comparisons, the aim is to showcase the superiority of the selected multi-instance pooling method in capturing relevant information from the instances.

6.3 Results and Analysis

6.3.1 Performance Comparisons

The proposed method is compared with three advanced machine learning techniques, and the performance of different pooling methods is evaluated to generate the final output. The detailed comparison results can be found in Table 6.2, using R^2, MAE, and MSE as evaluation metrics. A higher R^2 indicates better performance, while MAE and MSE are considered in the opposite direction.

The proposed method consistently outperforms all baseline methods, as evidenced by the R^2, MAE, and MSE values of 0.437, 8.471, and 160.016, respectively. Notably, FT Transformer demonstrates superior performance compared to XGBoost, LightGBM and CatBoost, highlighting its effectiveness in handling categorical and sparse matrix data, which are prominent characteristics of the dataset.

TABLE 6.2: Performance Comparison with Baseline Methods

Strategy	**Models**	**R^2**	**MAE**	**MSE**
Machine Learning Techniques	XGBoost-AutoML	0.420	8.979	167.972
	LightGBM-AutoML	0.327	9.495	190.347
	CatBoost-AutoML	0.427	8.773	163.884
	FT-Transformer	0.431	8.592	162.714
With Different Pooling Methods	Fully Connected Layer	0.405	8.742	168.549
	Mean Pooling	0.418	8.626	165.069
	Att. Pooling	0.426	8.639	163.122
	Gated Att. Pooling	0.424	8.580	163.558
This Work	Max Pooling	**0.437**	**8.471**	**160.016**

When comparing the fully connected layer with other multi-instance pooling methods, limitations of the fully connected layer in effectively locating relevant information are observed, as its R^2 value is only 0.405. On the other hand, attention-based pooling (Att. Pooling) and gated attention-based pooling (Gated Att. Pooling) assign different weights to instances to adjust their contributions. However, as the weights range from 0 to 1, they may restrict the model's performance on regression tasks. Specifically, the R^2 values for Att. Pooling and Gated Att. Pooling are 0.426 and 0.424, respectively, both lower than the performance achieved by the chosen max pooling method.

6.3.2 Impact of Multi-Head Attention

To assess the impact of employing multi-head attention, experiments were conducted with different numbers of heads in the method: 0, 2, 4, 6, 8, 10, and 12. When 0 heads were used, multi-head attention was not integrated into the proposed neural network. The evaluation results can be seen in Figure 3.

Among the different configurations, it was found that utilizing 8 heads yielded the best model performance. This suggests that the relations captured in these 8 subspaces fully

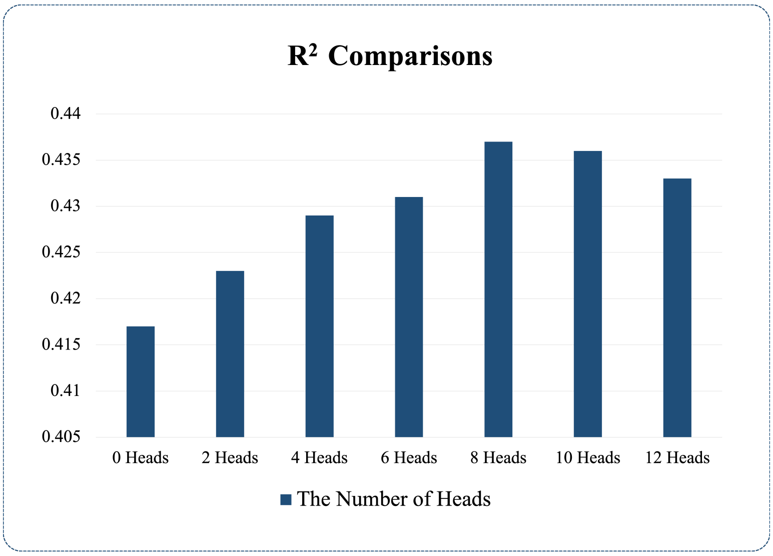


FIGURE 6.3: The Performance Comparisons of Different Number of Heads

uncover the underlying associations present in the clinical features. This finding has practical implications, as it implies that the employed features can be examined from 8 distinct perspectives, each with its own set of hidden connections.

Furthermore, the experiments confirmed the effectiveness of multi-head attention in the method. Removing it led to a significant decrease in R^2 to 0.417, indicating the crucial role of multi-head attention in capturing the complex relationships among the clinical features of warfarin users. These results underscore the importance of exploring correlations among the features and highlight the necessity of incorporating multi-head attention for improved performance.

6.3.3 Dose Subgroup Analysis

In this subsection, the primary focus is to evaluate the clinical applicability of the proposed method across different dose subgroups. Following the criteria described in [52], the warfarin doses are categorized into three groups: low dose group ($\leq 21\mathrm{mg/wk}$), medium dose group (> 21 to $< 49\mathrm{mg/wk}$), and high dose group ($\geq$

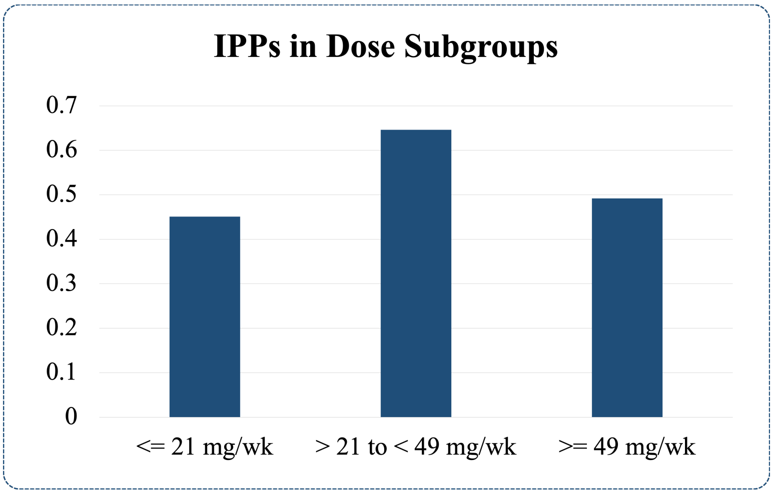


FIGURE 6.4: The IPPs Comparison in Different Dose Subgroups

$49\text{mg}/\text{wk}$). To assess the model's clinical applicability within these subgroups, the ideal predictive percentage (IPP) [52] is employed, which indicates the percentage of predicted doses falling within a 20% interval of the actual dose.

Figure 6.4 illustrates the results, indicating that the proposed method exhibits high clinical applicability, particularly in the medium and high dose groups, with IPPs of 0.646 and 0.492, respectively. In comparison, the low dose group has an IPP of 0.451. Notably, the medium dose group outperforms both the low and high dose groups in terms of IPP. This suggests that patients in the medium dose group exhibit less clinical variability and possess a more stable disease condition, making it easier to obtain the optimal warfarin dose from the predictive models. On the other hand, the lower IPP values in the low and high dose groups imply greater challenges in accurately predicting the optimal warfarin dose for patients in these subgroups, likely due to their higher clinical variability and disease complexity.

TABLE 6.3: Comprehensive Evaluation of the Proposed Method Using Different Feature Sets

Feature Set	R^2	**MAE**	**MSE**
CF	0.412	8.612	163.559
Cl	0.383	8.972	174.648
GF	0.395	8.910	169.513
This Work (Mixed)	**0.437**	**8.471**	**160.016**

6.3.4 Different Feature Sets

Finally, an evaluation was conducted on the method using four different feature combinations: (1) only continuous or ordinal features (CF), (2) clinical factors excluding continuous and ordinal features and genetic variables (Cl), (3) only genetic features (GF), and (4) all features combined (Mixed). The detailed comparison results can be found in Table 6.3.

By examining the results in Table 6.3, valuable insights can be gained regarding the performance of the method across these feature combinations.

The results demonstrate that among the various clinical features, the continuous or ordinal features, such as age, weight, height, target INR, and genetic variables, play a crucial role in guiding the optimal warfarin dose prediction. These features provide valuable insights and serve as primary indicators for accurate dose estimation. However, it is important to note that clinical features, including other factors such as demographic and medical history, also contribute significant information to the dose prediction task. Therefore, combining both continuous/ordinal features and clinical features proves to be the optimal approach, as it allows for the integration of comprehensive clinical guidance and leads to improved modeling performance. By incorporating a diverse range of features, the proposed method effectively leverages the synergistic power of these combined inputs, resulting in enhanced accuracy and robustness in warfarin dose estimation.

6.4 Summary

This work presents a groundbreaking and viable approach for warfarin dose estimation on incomplete and high-dimensional data, eliminating the need for data imputation and feature selection prior to analysis. Additionally, the AMI-Net (Adaptive Multi-level Integration Network) approach is extended to regression tasks, which hold significant importance in clinical applications. The methodology comprises two levels, with the first level utilizing a feature embedding module to seamlessly map all available information to an embedding space, effectively addressing missing values and redundant features. Leveraging the embedded vectors, the second modeling level employs a novel neural network architecture capable of capturing intricate and underlying relationships among features. This network not only facilitates the discovery of complex relationships but also isolates invalid information and noise, leading to enhanced warfarin dose determination.

Furthermore, future work aims to expand upon the existing methodology by addressing the challenges associated with temporal data. This involves incorporating frequent physical test results and lab test results to uncover hidden patterns and gain deeper insights into the dynamics of warfarin dose requirements. Exploring temporal data promises to provide a comprehensive understanding of patient profiles and enable more accurate predictions and personalized dosage recommendations. This extension is anticipated to advance the field of warfarin dose estimation and improve patient care outcomes.

CHAPTER 7

Conclusions

The nature of real-world data (RWD) in TCM and WM is complex, with issues related to data quality, inconsistency, and representation significantly affecting the usability and reliability of predictive learning models. For example, TCM relies heavily on holistic and subjective assessments, with diagnoses often based on qualitative observations like tongue appearance and pulse. The data from TCM might be more heterogeneous and less standardized compared to Western datasets. AMI-Net's ability to handle incomplete datasets and varying feature sets would be critical here. However, converting qualitative TCM diagnostics into quantifiable data for the model could be a significant challenge. WM datasets are typically more standardized and quantitative, with an emphasis on lab results, imaging data, and electronic health records. The challenge here would be in dealing with the volume of data, its complexity, and ensuring the model can understand and utilize the diverse range of clinical measurements and observations.

Despite these challenges, the utility of RWD in clinical decision-making cannot be overstated, as it provides an invaluable supplement to the evidence garnered from randomized controlled trials (RCTs), especially in special populations where RCTs are challenging to implement.

In this thesis, an exploration and proposal of the AMI-Net series – AMI-Net, AMI-Net+, and AMI-Net3, are presented as innovative approaches to overcoming RWD's limitations, thus facilitating its effective use in predictive modeling. By mapping data with varied features to a unified embedding space and emphasizing informative instances, the AMI-Net series offers a robust mechanism for handling incomplete data, noise, and extreme class imbalances, which are inherent issues when working with RWD. Additionally, the development of AMI-Net3 extends the series' applicability to a broader range of features, while the exploration of regression further enhances the versatility of these models.

Moreover, to effectively implement the AMI-Net series in clinical research and patient care, a structured approach is recommended. Firstly, identify the specific clinical objective, be it disease risk prediction, differential diagnosis, or prognostic analysis. Then, gather relevant real-world data (RWD), such as electronic medical records and input from portable devices, ensuring a broad and representative sample. The next step is to preprocess this data using AMI-Net's efficient data preprocessing tools, addressing common data quality issues like incomplete datasets and noise. This ensures the extraction of reliable insights even from imperfect data. Incorporate the preprocessed data into the AMI-Net algorithm, utilizing its robustness and versatility with various data types (binary, nominal, ordinal, and continuous). This facilitates its applicability across different medical domains. The automatic identification of informative data points by AMI-Net streamlines the decision-making process, enhancing patient care effectiveness. For multi-task learning objectives, such as advancing disease management and treatment methodologies, leverage AMI-Net's multi-task learning capabilities to concurrently address multiple research questions or clinical problems. This not only enhances the efficiency and comprehensiveness of clinical studies but also reduces research time and costs. Regularly evaluate the outcomes and insights generated by AMI-Net, adjusting the approach as needed based on real-world application feedback. This iterative process ensures continual improvement and adaptation of the AMI-Net series to specific clinical needs, maximizing its potential in harnessing RWD for more informed, efficient, and personalized healthcare solutions.

However, it is crucial to acknowledge that while the proposed methods display promise, they are not without limitations. In particular, the issues of missing values, data inconsistencies, and labeling errors, coupled with the binary-input-feature constraint of the AMI-Net and AMI-Net+, must be kept in mind. Furthermore, like all models dealing with real-world data, there exists a constant risk of overfitting and bias, warranting further research to refine understanding of these challenges and enhance the models.

7.1 Future Work

Efforts to develop AMI-Net models have primarily focused on single-task scenarios, including classification and regression. However, it is important to acknowledge that the realm of medical applications is extensive and often requires a multi-task or multi-label approach. There are numerous real-world medical scenarios where multiple outcomes or labels are required, such as in the generation of prescriptions or medication therapy management. In these complex scenarios, the prediction of multiple, often interrelated outcomes is paramount. Current AMI-Net models, while performing admirably in single-task contexts, do not fully cater to these multi-task or multi-label demands, thus marking a limitation in this work. However, the adaptability of AMI-Net to handle various types of data (binary, nominal, ordinal, continuous) indicates a potential for further development in handling complex tasks. The next steps could involve enhancing the model's ability to process and learn from multi-dimensional data, integrating temporal dynamics for time-series analysis (useful in patient monitoring), or even combining various data types (like imaging and textual data) for comprehensive analyses, such as multi-label classification or even multi-task prediction.

In addition, another challenge ubiquitous in the healthcare field is recognized: the extreme multi-label problem. This problem arises when a large number of potential labels exist, and only a few are relevant for a given instance. The extreme multi-label problem is frequently encountered in medical contexts where a patient can have several distinct medical conditions or symptoms concurrently, each requiring its label. Addressing this problem requires models capable of discerning the intricate interplay of multiple labels within a highly-dimensional label space. Here, too, it is acknowledged that current AMI-Net models fall short.

In light of these identified limitations, an exploratory study utilizing traditional Chinese Medicine data, as detailed by Wang et al. (2019) [261], has been undertaken. The primary task in this context is to predict prescriptions based on a multitude of patient symptoms, which involves dealing with a label set approaching one thousand in number. This scenario provides a rich ground for tackling the extreme multi-label problem inherent in such situations.

In this exploration, an innovative approach is introduced wherein the labels of a patient are treated as a sequence, similar to the strategy adopted in sequence-based language processing tasks. This unique perspective allows both the input (patient symptoms) and the output (prescriptions) to be viewed as two distinct language sequences. This innovative approach involves using a translator model to transform one sequence into the other. Through this translation mechanism, the issue of label sparsity that often complicates multi-label prediction tasks can be effectively sidestepped. In fact, preliminary results indicate a marked improvement, reflecting the potential efficacy of this method.

These promising findings provide an inspiring foundation for future research. They underscore the potential of novel and creative approaches like sequence translation in addressing complex prediction tasks within the medical domain. This will certainly guide and inform subsequent research endeavors aimed at improving the capabilities of the models in multi-task and extreme multi-label environments.

The discussion thus far has primarily focused on limitations at the label level, namely the challenges surrounding multi-task applications and the extreme multi-label problem. However, it's equally important to consider constraints arising from the nature of the data the models are designed to handle.

In the current state, work is predominantly based on static data, that is, data points that are fixed or unchanging over time. The static nature of data inputs provides a snapshot of conditions or symptoms at a specific point in time, which has proven useful in many applications. However, this approach may not fully capture the intricacies and dynamism inherent in many real-world medical scenarios. Medical conditions often present as temporal phenomena, changing and evolving over time, sometimes in predictable patterns and other times in ways that are more erratic and less easily anticipated. Consequently, the ability to work with temporal data, including time-series data, is critical in providing comprehensive and accurate insights. Unfortunately, existing AMI-Net models do not incorporate temporal data, which is a significant limitation. The models' architecture and functionality are not equipped to handle the dynamic nature of time-series data, which could offer valuable insights into how conditions evolve and respond to interventions over time. This lack of capacity to incorporate and learn from temporal patterns can limit the utility of these models in predicting future states or identifying trends.

This identified limitation forms a crucial area for future research. There is an aspiration to extend the models' capabilities to handle and learn from temporal data in the future, thus enhancing their ability to capture and reflect the dynamic nature of medical conditions. Incorporating this temporal perspective will potentially strengthen the predictive performance and generalizability of the models, making them more adaptable and useful in a broader range of medical applications.

The identified gaps in the models' capabilities are not meant to diminish their accomplishments, but rather to highlight the expansive potential for further research and development. They underline the necessity for continued evolution and adaptation of the models to meet the multifaceted and complex demands of real-world medical applications. In subsequent research, the intention is to explore these areas of multi-task learning and extreme multi-label problems, enhancing the models to address these challenges more effectively, thus moving closer to a comprehensive solution for predictive learning in medicine.

Furthermore, while AMI-Net has demonstrated initial success in the medical field, its underlying algorithms and methodology hold significant potential for application across various sectors, including finance, environmental science, and retail. Each of these fields presents unique challenges concerning data quality and noise, yet the core principles and functionalities of AMI-Net are well-suited to address these issues. In the finance sector, despite the complexities introduced by erratic market data and fraudulent transactions, AMI-Net's robustness to noise and capability to identify informative instances can be leveraged to enhance risk assessments and fraud detection. The method's ability to manage binary, nominal, ordinal, and continuous features makes it particularly adaptable to the diverse data types encountered in finance. Environmental science, faced with the challenge of heterogeneous and sometimes incomplete climate data, can benefit from AMI-Net's ability to learn from incomplete datasets. This feature is crucial in improving the accuracy and reliability of climate models, even when faced with data inconsistencies. In the retail sector, where analyzing customer behavior is often hindered by unstructured and voluminous data, AMI-Net's capacity to automatically identify important instances from a large dataset can play a pivotal role in refining customer behavior analysis and predictions.

Moreover, the integration of AMI-Net with IoT and wearable health technologies has the potential to revolutionize tele-health and remote patient monitoring. This integration would require the handling and interpretation of vast amounts of real-time health data, but AMI-Net's foundational algorithms are well-equipped to manage and process low-quality and noisy data effectively. In summary, the AMI-Net suite, while initially developed for medical applications, has underlying algorithms that are highly relevant and applicable to a broad range of sectors. This adaptability is key to addressing the specific challenges related to data quality and noise in each of these domains.

In conclusion, it can be argued that the AMI-Net series represents a significant step forward in the utilization of RWD for predictive learning. The demonstrated capacity of these models to navigate the inherent limitations of RWD underscores their potential value in medical studies and other applications. Nonetheless, it is crucial that future research continues to scrutinize and improve upon these algorithms, ensuring that the generated insights maintain the highest level of accuracy and reliability. The judicious application of these models, in tandem with rigorous quality control of data, could propel the field towards more efficient, effective, and personalized medical care.

Bibliography

[1] Alan Akbik et al. ‘FLAIR: An easy-to-use framework for state-of-the-art NLP’. In: *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics (Demonstrations)*. 2019, pp. 54–59.

[2] Kathy S Albain et al. ‘Prognostic and predictive value of the 21-gene recurrence score assay in postmenopausal women with node-positive, oestrogen-receptor-positive breast cancer on chemotherapy: a retrospective analysis of a randomised trial’. In: *The lancet oncology* 11.1 (2010), pp. 55–65.

[3] Paul D Allison et al. *Missing data*. Vol. 200210. 9781412985079.31. Sage Thousand Oaks, CA, 2010.

[4] Paul D Allison. ‘Missing data techniques for structural equation modeling.’ In: *Journal of abnormal psychology* 112.4 (2003), p. 545.

[5] Naomi S Altman. ‘An introduction to kernel and nearest-neighbor nonparametric regression’. In: *The American Statistician* 46.3 (1992), pp. 175–185.

[6] Robert A Amar et al. ‘Multiple-instance learning of real-valued data’. In: *ICML*. Citeseer. 2001, pp. 3–10.

[7] Jaume Amores. ‘Multiple instance classification: Review, taxonomy and comparative study’. In: *Artificial intelligence* 201 (2013), pp. 81–105.

[8] Stuart Andrews, Ioannis Tsochantaridis and Thomas Hofmann. ‘Support vector machines for multiple-instance learning’. In: *Advances in neural information processing systems*. 2003, pp. 577–584.

[9] Diego Ardila et al. ‘End-to-end lung cancer screening with three-dimensional deep learning on low-dose chest computed tomography’. In: *Nature medicine* 25.6 (2019), pp. 954–961.

[10] Sercan Ö Arik and Tomas Pfister. ‘Tabnet: Attentive interpretable tabular learning’. In: *Proceedings of the AAAI Conference on Artificial Intelligence*. Vol. 35. 8. 2021, pp. 6679–6687.

[11] Jimmy Lei Ba, Jamie Ryan Kiros and Geoffrey E Hinton. ‘Layer normalization’. In: *arXiv preprint arXiv:1607.06450* (2016).

[12] Stephen Bacchi et al. ‘Machine learning in the prediction of medical inpatient length of stay’. In: *Internal medicine journal* 52.2 (2022), pp. 176–185.

[13] Dzmitry Bahdanau, Kyunghyun Cho and Yoshua Bengio. ‘Neural machine translation by jointly learning to align and translate’. In: *arXiv preprint arXiv:1409.0473* (2014).

[14] Alastair Baker. *Crossing the quality chasm: a new health system for the 21st century*. Vol. 323. 7322. British Medical Journal Publishing Group, 2001.

[15] Miriam Barnum. ‘Dealing with missing and incomplete data’. In: *Handbook of Research Methods in International Relations*. Edward Elgar Publishing, 2022, pp. 425–445.

[16] David W Bates et al. ‘Big data in health care: using analytics to identify and manage high-risk and high-cost patients’. In: *Health affairs* 33.7 (2014), pp. 1123–1131.

[17] Gustavo EAPA Batista, Maria Carolina Monard et al. ‘A Study of K-Nearest Neighbour as an Imputation Method.’ In: *HIS* 87.251-260 (2002), p. 48.

[18] Kornelia Batko and Andrzej Ślęzak. ‘The use of Big Data Analytics in healthcare’. In: *Journal of big Data* 9.1 (2022), p. 3.

[19] Brett K Beaulieu-Jones, Jason H Moore and POOLED RESOURCE OPEN-ACCESS ALS CLINICAL TRIALS CONSORTIUM. ‘Missing data imputation

in the electronic health record using deeply learned autoencoders'. In: *Pacific symposium on biocomputing 2017*. World Scientific. 2017, pp. 207–218.

[20] Vanesa Bellou et al. 'Prognostic models for outcome prediction in patients with chronic obstructive pulmonary disease: systematic review and critical appraisal'. In: *Bmj* 367 (2019).

[21] Tirimula Rao Benala and Karunya Tantati. 'Efficiency of oversampling methods for enhancing software defect prediction by using imbalanced data'. In: *Innovations in Systems and Software Engineering* (2022), pp. 1–17.

[22] Eric I Benchimol et al. 'The REporting of studies Conducted using Observational Routinely-collected health Data (RECORD) statement'. In: *PLoS medicine* 12.10 (2015), e1001885.

[23] Yoshua Bengio, Aaron Courville and Pascal Vincent. 'Representation learning: A review and new perspectives'. In: *IEEE transactions on pattern analysis and machine intelligence* 35.8 (2013), pp. 1798–1828.

[24] Lorenzo Beretta and Alessandro Santaniello. 'Nearest neighbor imputation algorithms: a critical evaluation'. In: *BMC medical informatics and decision making* 16.3 (2016), pp. 197–208.

[25] Marc L Berger et al. 'Good practices for real-world data studies of treatment and/or comparative effectiveness: recommendations from the joint ISPOR-ISPE Special Task Force on real-world evidence in health care decision making'. In: *Value in Health* 20.8 (2017), pp. 1003–1008.

[26] Charles Bergeron et al. 'Fast bundle algorithm for multiple-instance learning'. In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 34.6 (2011), pp. 1068–1079.

[27] Dimitris Bertsimas, Colin Pawlowski and Ying Daisy Zhuo. 'From predictive methods to missing data imputation: an optimization approach.' In: *J. Mach. Learn. Res.* 18.1 (2017), pp. 7133–7171.

[28] Rok Blagus and Lara Lusa. ‘SMOTE for high-dimensional class-imbalanced data’. In: *BMC bioinformatics* 14 (2013), pp. 1–16.

[29] Adam L Booth, Elizabeth Abels and Peter McCaffrey. ‘Development of a prognostic model for mortality in COVID-19 infection using machine learning’. In: *Modern Pathology* 34.3 (2021), pp. 522–531.

[30] Giorgos Borboudakis and Ioannis Tsamardinos. ‘Forward-backward selection with early dropping’. In: *The Journal of Machine Learning Research* 20.1 (2019), pp. 276–314.

[31] Shyam Boriah, Varun Chandola and Vipin Kumar. ‘Similarity measures for categorical data: A comparative evaluation’. In: *Proceedings of the 2008 SIAM international conference on data mining*. SIAM. 2008, pp. 243–254.

[32] Y-Lan Boureau, Jean Ponce and Yann LeCun. ‘A theoretical analysis of feature pooling in visual recognition’. In: *Proceedings of the 27th international conference on machine learning (ICML-10)*. 2010, pp. 111–118.

[33] Stavroula Bourou et al. ‘A review of tabular data synthesis using GANs on an IDS dataset’. In: *Information* 12.09 (2021), p. 375.

[34] Omar Boursalie, Reza Samavi and Thomas E Doyle. ‘Evaluation metrics for deep learning imputation models’. In: *AI for Disease Surveillance and Pandemic Intelligence: Intelligent Disease Detection in Action*. Springer, 2022, pp. 309–322.

[35] Stephen Brooks. ‘Markov chain Monte Carlo method and its application’. In: *Journal of the royal statistical society: series D (the Statistician)* 47.1 (1998), pp. 69–100.

[36] S van Buuren and Karin Groothuis-Oudshoorn. ‘mice: Multivariate imputation by chained equations in R’. In: *Journal of statistical software* (2010), pp. 1–68.

[37] Enrico Capobianco. ‘High-dimensional role of AI and machine learning in cancer research’. In: *British journal of cancer* 126.4 (2022), pp. 523–532.

[38] Marc-André Carbonneau et al. ‘Multiple instance learning: A survey of problem characteristics and applications’. In: *Pattern Recognition* 77 (2018), pp. 329–353.

[39] Girish Chandrashekar and Ferat Sahin. ‘A survey on feature selection methods’. In: *Computers & Electrical Engineering* 40.1 (2014), pp. 16–28.

[40] Olivier Chapelle, Bernhard Scholkopf and Alexander Zien. ‘Semi-supervised learning (chapelle, o. et al., eds.; 2006)[book reviews]’. In: *IEEE Transactions on Neural Networks* 20.3 (2009), pp. 542–542.

[41] Jeremy Charlier et al. ‘SynGAN: Towards generating synthetic network attacks using GANs’. In: *arXiv preprint arXiv:1908.09899* (2019).

[42] Nitesh V Chawla et al. ‘SMOTE: synthetic minority over-sampling technique’. In: *Journal of artificial intelligence research* 16 (2002), pp. 321–357.

[43] Daohong Chen. ‘Real-world studies: Bridging the gap between trial-assessed efficacy and routine care’. In: *Journal of Biomedical Research* 36.3 (2022), p. 147.

[44] Jiahua Chen and Jun Shao. ‘Nearest neighbor imputation for survey data’. In: *Journal of official statistics* 16.2 (2000), p. 113.

[45] Jinpeng Chen et al. ‘Mining symptom-herb patterns from patient records using tripartite graph’. In: *Evidence-Based Complementary and Alternative Medicine* 2015 (2015).

[46] Tianqi Chen and Carlos Guestrin. ‘Xgboost: A scalable tree boosting system’. In: *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining*. 2016, pp. 785–794.

[47] Tianqi Chen et al. ‘Xgboost: extreme gradient boosting’. In: *R package version 0.4-2* 1.4 (2015), pp. 1–4.

[48] Veronika Cheplygina, David MJ Tax and Marco Loog. ‘Multiple instance learning with bag dissimilarities’. In: *Pattern recognition* 48.1 (2015), pp. 264–275.

[49] Edward Choi et al. ‘Generating multi-label discrete patient records using generative adversarial networks’. In: *Machine learning for healthcare conference*. PMLR. 2017, pp. 286–305.

[50] J Calvin Coffey and D Peter O’Leary. ‘The mesentery: structure, function, and role in disease’. In: *The lancet Gastroenterology & hepatology* 1.3 (2016), pp. 238–247.

[51] John Concato, Nirav Shah and Ralph I Horwitz. ‘Randomized, controlled trials, observational studies, and the hierarchy of research designs’. In: *New England journal of medicine* 342.25 (2000), pp. 1887–1892.

[52] International Warfarin Pharmacogenetics Consortium. ‘Estimation of the warfarin dose with clinical and pharmacogenetic data’. In: *New England Journal of Medicine* 360.8 (2009), pp. 753–764.

[53] Erdal Cosgun, Nita A Limdi and Christine W Duarte. ‘High-dimensional pharmacogenetic prediction of a continuous trait using machine learning techniques with application to warfarin dose prediction in African Americans’. In: *Bioinformatics* 27.10 (2011), pp. 1384–1389.

[54] Thomas M Cover. *Elements of information theory*. John Wiley & Sons, 1999.

[55] Ralph B D’Agostino Sr et al. ‘General cardiovascular risk profile for use in primary care: the Framingham Heart Study’. In: *Circulation* 117.6 (2008), pp. 743–753.

[56] Sabyasachi Dash et al. ‘Big data in healthcare: management, analysis and future prospects’. In: *Journal of Big Data* 6.1 (2019), pp. 1–25.

[57] Yann N Dauphin et al. ‘Language modeling with gated convolutional networks’. In: *Proceedings of the 34th International Conference on Machine Learning-Volume 70*. JMLR. org. 2017, pp. 933–941.

[58] David L DeMets and Susan S Ellenberg. 'Data monitoring committees—expect the unexpected'. In: *New England Journal of Medicine* 375.14 (2016), pp. 1365–1371.

[59] Thomas G Dietterich, Richard H Lathrop and Tomás Lozano-Pérez. 'Solving the multiple instance problem with axis-parallel rectangles'. In: *Artificial intelligence* 89.1-2 (1997), pp. 31–71.

[60] Xiaojian Ding, Fan Yang and Fuming Ma. 'An efficient model selection for linear discriminant function-based recursive feature elimination'. In: *Journal of Biomedical Informatics* 129 (2022), p. 104070.

[61] Carsten F Dormann et al. 'Collinearity: a review of methods to deal with it and a simulation study evaluating their performance'. In: *Ecography* 36.1 (2013), pp. 27–46.

[62] Anna Veronika Dorogush, Vasily Ershov and Andrey Gulin. 'CatBoost: gradient boosting with categorical features support'. In: *arXiv preprint arXiv:1810.11363* (2018).

[63] Sahibsingh A Dudani. 'The distance-weighted k-nearest-neighbor rule'. In: *IEEE Transactions on Systems, Man, and Cybernetics* 4 (1976), pp. 325–327.

[64] Charles Elkan and Keith Noto. 'Learning classifiers from only positive and unlabeled data'. In: *Proceedings of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining*. 2008, pp. 213–220.

[65] Joshua Elliott et al. 'Predictive accuracy of a polygenic risk score–enhanced prediction model vs a clinical risk score for coronary artery disease'. In: *Jama* 323.7 (2020), pp. 636–645.

[66] Craig K Enders. *Applied missing data analysis*. Guilford Publications, 2022.

[67] Bart S Ferket et al. 'Systematic review of guidelines on cardiovascular risk assessment: which recommendations should clinicians follow for a cardiovascular health check?' In: *Archives of internal medicine* 170.1 (2010), pp. 27–40.

[68] James Foulds and Eibe Frank. ‘A review of multi-instance learning assumptions’. In: *The Knowledge Engineering Review* 25.1 (2010), pp. 1–25.

[69] Jessica M Franklin and Sebastian Schneeweiss. ‘When and how can real world data analyses substitute for randomized controlled trials?’ In: *Clinical Pharmacology & Therapeutics* 102.6 (2017), pp. 924–933.

[70] Maayan Frid-Adar et al. ‘GAN-based synthetic medical image augmentation for increased CNN performance in liver lesion classification’. In: *Neurocomputing* 321 (2018), pp. 321–331.

[71] Thomas R Frieden. ‘Evidence for health decision making—beyond randomized, controlled trials’. In: *New England Journal of Medicine* 377.5 (2017), pp. 465–475.

[72] Brian F Gage et al. ‘Use of pharmacogenetics and clinical factors to predict the maintenance dose of warfarin’. In: *Thrombosis and haemostasis* 91.01 (2004), pp. 87–94.

[73] Tejasvita Gaikwad et al. ‘Warfarin dose model for the prediction of stable maintenance dose in indian patients’. In: *Clinical and Applied Thrombosis/Hemostasis* 24.2 (2018), pp. 353–359.

[74] Salvador Garcı et al. ‘Evolutionary-based selection of generalized instances for imbalanced classification’. In: *Knowledge-Based Systems* 25.1 (2012), pp. 3–12.

[75] Thomas Gärtner, Peter Flach and Stefan Wrobel. ‘On graph kernels: Hardness results and efficient alternatives’. In: *Learning Theory and Kernel Machines: 16th Annual Conference on Learning Theory and 7th Kernel Workshop, COLT/Kernel 2003, Washington, DC, USA, August 24-27, 2003. Proceedings*. Springer. 2003, pp. 129–143.

[76] Thomas Gärtner et al. ‘Multi-instance kernels’. In: *ICML*. Vol. 2. 3. 2002, p. 7.

[77] Andrew Gelman et al. *Bayesian data analysis*. CRC press, 2013.

[78] Zoubin Ghahramani and Michael I Jordan. ‘Supervised learning from incomplete data via an EM approach’. In: *Advances in neural information processing systems*. 1994, pp. 120–127.

[79] Ross Girshick. ‘Fast r-cnn’. In: *Proceedings of the IEEE international conference on computer vision*. 2015, pp. 1440–1448.

[80] Benjamin A Goldstein and Michael J Pencina. ‘Developing Implementable Risk Prediction Models with Electronic Health Records Data’. In: *Wiley StatsRef: Statistics Reference Online* (2014), pp. 1–8.

[81] Ian Goodfellow, Yoshua Bengio and Aaron Courville. *Deep learning*. MIT press, 2016.

[82] Ian Goodfellow et al. ‘Generative adversarial networks’. In: *Communications of the ACM* 63.11 (2020), pp. 139–144.

[83] Yury Gorishniy et al. ‘Revisiting deep learning models for tabular data’. In: *Advances in Neural Information Processing Systems* 34 (2021), pp. 18932–18943.

[84] Somya Goyal. ‘Handling class-imbalance with KNN (neighbourhood) undersampling for software defect prediction’. In: *Artificial Intelligence Review* 55.3 (2022), pp. 2023–2064.

[85] David Grangier and Iain Melvin. ‘Feature set embedding for incomplete data’. In: *Advances in Neural Information Processing Systems* 23 (2010).

[86] Alex Graves and Alex Graves. ‘Long short-term memory’. In: *Supervised sequence labelling with recurrent neural networks* (2012), pp. 37–45.

[87] Alex Graves, Abdel-rahman Mohamed and Geoffrey Hinton. ‘Speech recognition with deep recurrent neural networks’. In: *2013 IEEE international conference on acoustics, speech and signal processing*. Ieee. 2013, pp. 6645–6649.

[88] Lawrence W Green. ‘Public health asks of systems science: to advance our evidence-based practice, can you help us get more practice-based evidence?’ In: *American journal of public health* 96.3 (2006), pp. 406–409.

[89] Enzo Grossi et al. ‘Prediction of optimal warfarin maintenance dose using advanced artificial neural networks’. In: *Pharmacogenomics* 15.1 (2014), pp. 29–37.

[90] Steve R Gunn et al. ‘Support vector machines for classification and regression’. In: *ISIS technical report* 14.1 (1998), pp. 5–16.

[91] Wei Guo et al. ‘A machine learning model to predict risperidone active moiety concentration based on initial therapeutic drug monitoring’. In: *Frontiers in Psychiatry* 12 (2021), p. 711868.

[92] Yu Guo et al. ‘Combating imbalance in network traffic classification using GAN based oversampling’. In: *2021 IFIP Networking Conference (IFIP Networking)*. IEEE. 2021, pp. 1–9.

[93] Gordon H Guyatt et al. ‘GRADE guidelines 6. Rating the quality of evidence—imprecision’. In: *Journal of clinical epidemiology* 64.12 (2011), pp. 1283–1293.

[94] Isabelle Guyon and André Elisseeff. ‘An introduction to variable and feature selection’. In: *Journal of machine learning research* 3.Mar (2003), pp. 1157–1182.

[95] Hui Han, Wen-Yuan Wang and Bing-Huan Mao. ‘Borderline-SMOTE: a new over-sampling method in imbalanced data sets learning’. In: *Advances in Intelligent Computing: International Conference on Intelligent Computing, ICIC 2005, Hefei, China, August 23-26, 2005, Proceedings, Part I 1*. Springer. 2005, pp. 878–887.

[96] Yufei Han et al. ‘Multi-label learning with highly incomplete data via collaborative embedding’. In: *Proceedings of the 24th ACM SIGKDD international conference on knowledge discovery & data mining*. 2018, pp. 1494–1503.

[97] Peter Hart. ‘The condensed nearest neighbor rule (corresp.)’ In: *IEEE transactions on information theory* 14.3 (1968), pp. 515–516.

[98] Trevor Hastie et al. *The elements of statistical learning: data mining, inference, and prediction*. Vol. 2. Springer, 2009.

[99] Mardhiya Hayaty, Siti Muthmainah and Syed Muhammad Ghufran. ‘Random and synthetic over-sampling approach to resolve data imbalance in classification’. In: *International Journal of Artificial Intelligence Research* 4.2 (2020), pp. 86–94.

[100] R Brian Haynes. ‘Of studies, syntheses, synopses, summaries, and systems: the “5S” evolution of information services for evidence-based healthcare decisions’. In: *BMJ Evidence-Based Medicine* 11.6 (2006), pp. 162–164.

[101] Haibo He and Edwardo A Garcia. ‘Learning from imbalanced data’. In: *IEEE Transactions on knowledge and data engineering* 21.9 (2009), pp. 1263–1284.

[102] Haibo He et al. ‘ADASYN: Adaptive synthetic sampling approach for imbalanced learning’. In: *2008 IEEE international joint conference on neural networks (IEEE world congress on computational intelligence)*. IEEE. 2008, pp. 1322–1328.

[103] Kaiming He et al. ‘Deep residual learning for image recognition’. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2016, pp. 770–778.

[104] Kaiming He et al. ‘Mask r-cnn’. In: *Proceedings of the IEEE international conference on computer vision*. 2017, pp. 2961–2969.

[105] Xia He et al. ‘A Risk Scoring Model for High-Dose Methotrexate-Induced Liver Injury in Children With Acute Lymphoblastic Leukemia Based on Gene Polymorphism Study’. In: *Frontiers in Pharmacology* 12 (2021), p. 726229.

[106] Francisco Herrera et al. ‘Multiple Instance Multiple Label Learning’. In: *Multiple Instance Learning: Foundations and Algorithms* (2016), pp. 209–230.

[107] Geoffrey E Hinton. ‘Deep belief networks’. In: *Scholarpedia* 4.5 (2009), p. 5947.

[108] Jack Hirsh et al. ‘American Heart Association/American College of Cardiology foundation guide to warfarin therapy’. In: *Circulation* 107.12 (2003), pp. 1692–1711.

[109] Tin Kam Ho. ‘Random decision forests’. In: *Proceedings of 3rd international conference on document analysis and recognition*. Vol. 1. IEEE. 1995, pp. 278–282.

[110] David W Hosmer Jr, Stanley Lemeshow and Rodney X Sturdivant. *Applied logistic regression*. Vol. 398. John Wiley & Sons, 2013.

[111] Chang-Hua Hu et al. ‘A prognostic model based on DBN and diffusion process for degrading bearing’. In: *IEEE Transactions on Industrial Electronics* 67.10 (2019), pp. 8767–8777.

[112] Dichao Hu. ‘An introductory survey on attention mechanisms in NLP problems’. In: *Intelligent Systems and Applications: Proceedings of the 2019 Intelligent Systems Conference (IntelliSys) Volume 2*. Springer. 2020, pp. 432–448.

[113] Weiwei Hu and Ying Tan. ‘Generating adversarial malware examples for black-box attacks based on GAN’. In: *Data Mining and Big Data: 7th International Conference, DMBD 2022, Beijing, China, November 21–24, 2022, Proceedings, Part II*. Springer. 2023, pp. 409–423.

[114] Yang Hu, Mingjing Li and Nenghai Yu. ‘Multiple-instance ranking: Learning to rank images for image retrieval’. In: *2008 IEEE Conference on Computer Vision and Pattern Recognition*. IEEE. 2008, pp. 1–8.

[115] David Hua et al. ‘Using AI-Driven Triaging to Optimise Clinical Workflows in Non-Emergency Outpatient Settings: A Real-World Case Study Concerning the Screening of Tuberculosis’. In: *Proceedings of the 2023 Australasian Computer Science Week*. 2023, pp. 240–243.

[116] Guilin Huang. ‘Missing data filling method based on linear interpolation and lightgbm’. In: *Journal of Physics: Conference Series*. Vol. 1754. 1. IOP Publishing. 2021, p. 012187.

[117] Xiaohui Huang et al. ‘Prediction of vancomycin dose on high-dimensional data using machine learning techniques’. In: *Expert Review of Clinical Pharmacology* 14.6 (2021), pp. 761–771.

[118] Maximilian Ilse, Jakub Tomczak and Max Welling. ‘Attention-based deep multiple instance learning’. In: *International conference on machine learning*. PMLR. 2018, pp. 2127–2136.

[119] Md Aminul Islam and Nusrat Jahan. ‘Prediction of onset diabetes using machine learning techniques’. In: *International Journal of Computer Applications* 180.5 (2017), pp. 7–11.

[120] José M Jerez et al. ‘Missing data imputation using statistical and machine learning methods in a real breast cancer problem’. In: *Artificial intelligence in medicine* 50.2 (2010), pp. 105–115.

[121] Tammy Jiang et al. ‘Addressing measurement error in random forests using quantitative bias analysis’. In: *American Journal of Epidemiology* 190.9 (2021), pp. 1830–1840.

[122] Justin M Johnson and Taghi M Khoshgoftaar. ‘Survey on deep learning with class imbalance’. In: *Journal of Big Data* 6.1 (2019), pp. 1–54.

[123] James M Joyce. ‘Kullback-leibler divergence’. In: *International encyclopedia of statistical science*. Springer, 2011, pp. 720–722.

[124] Hyun Kang. ‘The prevention and handling of the missing data’. In: *Korean journal of anesthesiology* 64.5 (2013), pp. 402–406.

[125] Devan Kansagara et al. ‘Risk prediction models for hospital readmission: a systematic review’. In: *Jama* 306.15 (2011), pp. 1688–1698.

[126] Nazmul Karim et al. ‘Unicon: Combating label noise through uniform selection and contrastive learning’. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2022, pp. 9676–9686.

[127] Guolin Ke et al. ‘Lightgbm: A highly efficient gradient boosting decision tree’. In: *Advances in neural information processing systems* 30 (2017).

[128] Salman Khan et al. ‘Transformers in vision: A survey’. In: *ACM computing surveys (CSUR)* 54.10s (2022), pp. 1–41.

[129] Han-Gyu Kim et al. ‘Recurrent neural networks with missing information imputation for medical examination data prediction’. In: *2017 IEEE International Conference on Big Data and Smart Computing (BigComp)*. IEEE. 2017, pp. 317–323.

[130] Jaeyoon Kim, Donghyun Tae and Junhee Seok. ‘A survey of missing data imputation using generative adversarial networks’. In: *2020 International conference on artificial intelligence in information and communication (ICAIIC)*. IEEE. 2020, pp. 454–456.

[131] Diederik P Kingma and Jimmy Ba. ‘Adam: A method for stochastic optimization’. In: *arXiv preprint arXiv:1412.6980* (2014).

[132] William A Knaus et al. ‘APACHE II: a severity of disease classification system.’ In: *Critical care medicine* 13.10 (1985), pp. 818–829.

[133] R Kohavi and GH John. *Wrappers for feature subset selection, Artificial Intelligence, vol. 97, no. 1-2*. 1997.

[134] Konstantina Kourou et al. ‘Machine learning applications in cancer prognosis and prediction’. In: *Computational and structural biotechnology journal* 13 (2015), pp. 8–17.

[135] Miroslav Kubat, Stan Matwin et al. ‘Addressing the curse of imbalanced training sets: one-sided selection’. In: *Icml*. Vol. 97. 1. Citeseer. 1997, p. 179.

[136] Christine Laine et al. ‘Reproducible research: moving toward research the public can really trust’. In: *Annals of Internal Medicine* 146.6 (2007), pp. 450–453.

[137] Edmund C Lau et al. ‘Use of electronic medical records (EMR) for oncology outcomes research: assessing the comparability of EMR information to patient registry and health claims data’. In: *Clinical epidemiology* 3 (2011), p. 259.

[138] Jorma Laurikkala. ‘Improving identification of difficult small classes by balancing class distribution’. In: *Artificial Intelligence in Medicine: 8th Conference on Artificial Intelligence in Medicine in Europe, AIME 2001 Cascais, Portugal, July 1–4, 2001, Proceedings 8*. Springer. 2001, pp. 63–66.

[139] Yann LeCun, Yoshua Bengio and Geoffrey Hinton. ‘Deep learning’. In: *nature* 521.7553 (2015), pp. 436–444.

[140] P Lenzini et al. ‘Integration of genetic, clinical, and INR data to refine warfarin dosing’. In: *Clinical Pharmacology & Therapeutics* 87.5 (2010), pp. 572–578.

[141] Wenhua Liang et al. ‘Development and validation of a clinical risk score to predict the occurrence of critical illness in hospitalized patients with COVID-19’. In: *JAMA internal medicine* 180.8 (2020), pp. 1081–1089.

[142] Xuejun Liao, Hui Li and Lawrence Carin. ‘Quadratically gated mixture of experts for incomplete data classification’. In: *Proceedings of the 24th International Conference on Machine learning*. 2007, pp. 553–560.

[143] Junji Lin et al. ‘Application of electronic medical record data for health outcomes research: a review of recent literature’. In: *Expert review of pharmacoeconomics & outcomes research* 13.2 (2013), pp. 191–200.

[144] Tsung-Yi Lin et al. ‘Focal loss for dense object detection’. In: *Proceedings of the IEEE international conference on computer vision*. 2017, pp. 2980–2988.

[145] Wei-Chao Lin et al. ‘Clustering-based undersampling in class-imbalanced data’. In: *Information Sciences* 409 (2017), pp. 17–26.

[146] Zilong Lin, Yong Shi and Zhi Xue. ‘Idsgan: Generative adversarial networks for attack generation against intrusion detection’. In: *Advances in Knowledge Discovery and Data Mining: 26th Pacific-Asia Conference, PAKDD 2022, Chengdu, China, May 16–19, 2022, Proceedings, Part III*. Springer. 2022, pp. 79–91.

[147] Geert Litjens et al. ‘A survey on deep learning in medical image analysis’. In: *Medical image analysis* 42 (2017), pp. 60–88.

[148] Roderick JA Little and Donald B Rubin. *Statistical analysis with missing data*. Vol. 793. John Wiley & Sons, 2019.

[149] Guoqing Liu, Jianxin Wu and Zhi-Hua Zhou. ‘Key instance detection in multi-instance learning’. In: *Asian Conference on Machine Learning*. PMLR. 2012, pp. 253–268.

[150] KE Liu, C-L Lo and Y-H Hu. ‘Improvement of adequate use of warfarin for the elderly using decision tree-based approaches’. In: *Methods of Information in Medicine* 53.01 (2014), pp. 47–53.

[151] Rong Liu et al. ‘Comparison of nine statistical model based warfarin pharmacogenetic dosing algorithms using the racially diverse international warfarin pharmacogenetic consortium cohort database’. In: *PloS one* 10.8 (2015), e0135784.

[152] Steven Liu et al. ‘Diverse image generation via self-conditioned gans’. In: *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 2020, pp. 14286–14295.

[153] Tao Liu et al. ‘Non-instinct detection of cellphone usage from lane-keeping performance based on eXtreme gradient boosting and optimal sliding windows’. In: *IET Intelligent Transport Systems* 16.11 (2022), pp. 1600–1610.

[154] Xiaodong Liu et al. ‘A GAN and feature selection-based oversampling technique for intrusion detection’. In: *Security and communication networks* 2021 (2021), pp. 1–15.

[155] Xu-Ying Liu, Jianxin Wu and Zhi-Hua Zhou. ‘Exploratory undersampling for class-imbalance learning’. In: *IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics)* 39.2 (2008), pp. 539–550.

[156] Yushan Liu and Steven D Brown. ‘Comparison of five iterative imputation methods for multivariate classification’. In: *Chemometrics and Intelligent Laboratory Systems* 120 (2013), pp. 106–115.

[157] Zhun-ga Liu et al. ‘Adaptive imputation of missing values for incomplete pattern classification’. In: *Pattern Recognition* 52 (2016), pp. 85–95.

[158] Lin Lu et al. ‘Wearable health devices in health care: narrative systematic review’. In: *JMIR mHealth and uHealth* 8.11 (2020), e18907.

[159] Zhiyuan Ma et al. ‘Ensemble of machine learning algorithms using the stacked generalization approach to estimate the warfarin dose’. In: *PloS one* 13.10 (2018), e0205872.

[160] Tomasz Maciejewski and Jerzy Stefanowski. ‘Local neighbourhood extension of SMOTE for mining imbalanced data’. In: *2011 IEEE symposium on computational intelligence and data mining (CIDM)*. IEEE. 2011, pp. 104–111.

[161] David MacKay. ‘Information theory, pattern recognition and neural networks’. In: *Proceedings of the 1st International Conference on Evolutionary Computation*. Cambridge University Press Cambridge, UK. 2003.

[162] Ruth Macklin. ‘Enrolling pregnant women in biomedical research’. In: *The Lancet* 375.9715 (2010), pp. 632–633.

[163] Amr Makady et al. ‘What is real-world data? A review of definitions based on literature and stakeholder interviews’. In: *Value in health* 20.7 (2017), pp. 858–865.

[164] Inderjeet Mani and I Zhang. ‘kNN approach to unbalanced data distributions: a case study involving information extraction’. In: *Proceedings of workshop on learning from imbalanced datasets*. Vol. 126. ICML. 2003, pp. 1–7.

[165] Elaine R Mardis. ‘Next-generation sequencing platforms’. In: *Annual review of analytical chemistry* 6 (2013), pp. 287–303.

[166] Oded Maron and Tomás Lozano-Pérez. ‘A framework for multiple-instance learning’. In: *Advances in neural information processing systems* 10 (1997).

[167] Fawad Masood et al. ‘Novel approach to evaluate classification algorithms and feature selection filter algorithms using medical data’. In: *Journal of Computational and Cognitive Engineering* 2.1 (2023), pp. 57–67.

[168] Gary C McDonald. ‘Ridge regression’. In: *Wiley Interdisciplinary Reviews: Computational Statistics* 1.1 (2009), pp. 93–100.

[169] Roxana Mehran et al. ‘A risk score to predict bleeding in patients with acute coronary syndromes’. In: *Journal of the American College of Cardiology* 55.23 (2010), pp. 2556–2566.

[170] Tomas Mikolov et al. ‘Efficient estimation of word representations in vector space’. In: *arXiv preprint arXiv:1301.3781* (2013).

[171] Roweida Mohammed, Jumanah Rawashdeh and Malak Abdullah. ‘Machine learning with oversampling and undersampling techniques: overview study and experimental results’. In: *2020 11th international conference on information and communication systems (ICICS)*. IEEE. 2020, pp. 243–248.

[172] Fantine Mordelet and J-P Vert. ‘A bagging SVM to learn from positive and unlabeled examples’. In: *Pattern Recognition Letters* 37 (2014), pp. 201–209.

[173] Bobak J Mortazavi et al. ‘Analysis of machine learning techniques for heart failure readmissions’. In: *Circulation: Cardiovascular Quality and Outcomes* 9.6 (2016), pp. 629–640.

[174] David B Morton and PH Griffiths. ‘Guidelines on the recognition of pain, distress and discomfort in experimental animals and an hypothesis for assessment’. In: *Vet Rec* 116.16 (1985), pp. 431–6.

[175] Alejandro Mottini, Alix Lheritier and Rodrigo Acuna-Agost. ‘Airline passenger name record generation using generative adversarial networks’. In: *arXiv preprint arXiv:1807.06657* (2018).

[176] Jared S Murray. ‘Multiple imputation: a review of practical and theoretical findings’. In: (2018).

[177] Jared S Murray and Jerome P Reiter. ‘Multiple imputation of missing categorical and continuous values via Bayesian mixture models with local dependence’. In: *Journal of the American Statistical Association* 111.516 (2016), pp. 1466–1479.

[178] Carol M Musil et al. ‘A comparison of imputation techniques for handling missing data’. In: *Western journal of nursing research* 24.7 (2002), pp. 815–829.

[179] Nagarajan Natarajan et al. ‘Learning with noisy labels’. In: *Advances in neural information processing systems* 26 (2013).

[180] Alexey Natekin and Alois Knoll. ‘Gradient boosting machines, a tutorial’. In: *Frontiers in neurorobotics* 7 (2013), p. 21.

[181] Nonso Nnamoko, Abir Hussain and David England. ‘Predicting diabetes onset: an ensemble supervised learning approach’. In: *2018 IEEE Congress on evolutionary computation (CEC)*. IEEE. 2018, pp. 1–7.

[182] Arie Nouwen et al. ‘Type 2 diabetes mellitus as a risk factor for the onset of depression: a systematic review and meta-analysis’. In: *Diabetologia* 53 (2010), pp. 2480–2486.

[183] Ziad Obermeyer and Ezekiel J Emanuel. ‘Predicting the future—big data, machine learning, and clinical medicine’. In: *The New England journal of medicine* 375.13 (2016), p. 1216.

[184] Joo-Hyuk Oh, Jae Yeol Hong and Jun-Geol Baek. ‘Oversampling method using outlier detectable generative adversarial network’. In: *Expert Systems with Applications* 133 (2019), pp. 1–8.

[185] Soonmyung Paik et al. ‘A multigene assay to predict recurrence of tamoxifen-treated, node-negative breast cancer’. In: *New England Journal of Medicine* 351.27 (2004), pp. 2817–2826.

[186] Adam Pantanowitz and Tshilidzi Marwala. ‘Missing data imputation through the use of the Random Forest Algorithm’. In: *Advances in Computational Intelligence*. Springer, 2009, pp. 53–62.

[187] Noseong Park et al. ‘Data synthesis based on generative adversarial networks’. In: *arXiv preprint arXiv:1806.03384* (2018).

[188] Arshi Parvaiz et al. ‘Vision transformers in medical computer vision—A contemplative retrospection’. In: *Engineering Applications of Artificial Intelligence* 122 (2023), p. 106126.

[189] Karl Pearson. ‘VII. Note on regression and inheritance in the case of two parents’. In: *proceedings of the royal society of London* 58.347-352 (1895), pp. 240–242.

[190] Ekachai Phaisangittisagul. ‘An analysis of the regularization between L2 and dropout in single hidden layer neural network’. In: *2016 7th International Conference on Intelligent Systems, Modelling and Simulation (ISMS)*. IEEE. 2016, pp. 174–179.

[191] Edward F Philbin and Thomas G DiSalvo. ‘Prediction of hospital readmission for heart failure: development of a simple risk score based on administrative

data'. In: *Journal of the American College of Cardiology* 33.6 (1999), pp. 1560–1566.

[192] Robert S Phillips, Bas Vaarwerk and Jessica E Morgan. 'Using Evidence-Based Medicine to Support Clinical Decision-Making in RMS'. In: *Cancers* 15.1 (2022), p. 66.

[193] Walter Hugo Lopez Pinaya et al. 'Autoencoders'. In: *Machine learning*. Elsevier, 2020, pp. 193–208.

[194] Simon K Poon et al. 'A novel approach in discovering significant interactions from TCM patient prescription data'. In: *International journal of data mining and bioinformatics* 5.4 (2011), pp. 353–368.

[195] Mohammad H Poursaeidi and O Erhun Kundakcioglu. 'Robust support vector machines for multiple instance learning'. In: *Annals of Operations Research* 216.1 (2014), pp. 205–227.

[196] Yanjun Qi. 'Random forest for bioinformatics'. In: *Ensemble machine learning: Methods and applications*. Springer, 2012, pp. 307–323.

[197] Jiaohua Qin et al. 'A biological image classification method based on improved CNN'. In: *Ecological Informatics* 58 (2020), p. 101093.

[198] J. Ross Quinlan. 'Induction of decision trees'. In: *Machine learning* 1 (1986), pp. 81–106.

[199] Alec Radford, Luke Metz and Soumith Chintala. 'Unsupervised representation learning with deep convolutional generative adversarial networks'. In: *arXiv preprint arXiv:1511.06434* (2015).

[200] Alvin Rajkomar, Jeffrey Dean and Isaac Kohane. 'Machine learning in medicine'. In: *New England Journal of Medicine* 380.14 (2019), pp. 1347–1358.

[201] Jan Ramon and Luc De Raedt. 'Multi instance neural networks'. In: *Proceedings of the ICML-2000 workshop on attribute-value and relational learning*. 2000, pp. 53–60.

[202] René Ranftl, Alexey Bochkovskiy and Vladlen Koltun. 'Vision transformers for dense prediction'. In: *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 2021, pp. 12179–12188.

[203] Jonas Ranstam and JA Cook. 'LASSO regression'. In: *Journal of British Surgery* 105.10 (2018), pp. 1348–1348.

[204] Noémie Resseguier, Roch Giorgi and Xavier Paoletti. 'Sensitivity analysis when data are missing not-at-random'. In: *Epidemiology* 22.2 (2011), p. 282.

[205] Sherri Rose. 'Mortality risk score prediction in an elderly population using machine learning'. In: *American journal of epidemiology* 177.5 (2013), pp. 443–452.

[206] Mehrdad Rostami and Mourad Oussalah. 'A novel explainable COVID-19 diagnosis method by integration of feature selection with random forest'. In: *Informatics in Medicine Unlocked* 30 (2022), p. 100941.

[207] Peter M Rothwell. 'External validity of randomised controlled trials:"to whom do the results of this trial apply?"' In: *The Lancet* 365.9453 (2005), pp. 82–93.

[208] Patrick Royston. 'Multiple imputation of missing values'. In: *The Stata Journal* 4.3 (2004), pp. 227–241.

[209] Donald B Rubin. *Multiple imputation for nonresponse in surveys*. Vol. 81. John Wiley & Sons, 2004.

[210] Vivek A Rudrapatna, Atul J Butte et al. 'Opportunities and challenges in using real-world data for health care'. In: *The Journal of Clinical Investigation* 130.2 (2020), pp. 565–574.

[211] David L Sackett et al. *Evidence based medicine: what it is and what it isn't*. 1996.

[212] Yvan Saeys, Inaki Inza and Pedro Larranaga. 'A review of feature selection techniques in bioinformatics'. In: *bioinformatics* 23.19 (2007), pp. 2507–2517.

[213] Jorge Sánchez et al. ‘Image classification with the fisher vector: Theory and practice’. In: *International journal of computer vision* 105 (2013), pp. 222–245.

[214] Joseph L Schafer. *Analysis of incomplete multivariate data*. CRC press, 1997.

[215] Joseph L Schafer and John W Graham. ‘Missing data: our view of the state of the art.’ In: *Psychological methods* 7.2 (2002), p. 147.

[216] Kenneth F Schulz and David A Grimes. ‘Allocation concealment in randomised trials: defending against deciphering’. In: *The Lancet* 359.9306 (2002), pp. 614–618.

[217] Karen A Schwarz and Cheryl S Elman. ‘Identification of factors predictive of hospital readmissions for patients with heart failure’. In: *Heart & Lung* 32.2 (2003), pp. 88–99.

[218] Chiranjibi Shah, Qian Du and Yan Xu. ‘Enhanced TabNet: Attentive interpretable tabular learning for hyperspectral image classification’. In: *Remote Sensing* 14.3 (2022), p. 716.

[219] Rachel E Sherman et al. ‘Real-world evidence—what is it and what can it tell us’. In: *N Engl J Med* 375.23 (2016), pp. 2293–2297.

[220] Pannagadatta K Shivaswamy, Chiranjib Bhattacharyya and Alexander J Smola. ‘Second order cone programming approaches for handling missing and uncertain data’. In: *Journal of Machine Learning Research* (2006), pp. 1283–1314.

[221] Edward H Shortliffe and Martin J Sepúlveda. ‘Clinical decision support in the era of artificial intelligence’. In: *Jama* 320.21 (2018), pp. 2199–2200.

[222] Wen-ying Shu et al. ‘Pharmacogenomics and personalized medicine: a review focused on their application in the Chinese population’. In: *Acta pharmacologica Sinica* 36.5 (2015), pp. 535–543.

[223] George CM Siontis et al. ‘External validation of new risk prediction models is infrequent and reveals worse prognostic discrimination’. In: *Journal of clinical epidemiology* 68.1 (2015), pp. 25–34.

[224] Marek Śmieja et al. 'Generalized RBF kernel for incomplete data'. In: *Knowledge-Based Systems* 173 (2019), pp. 150–162.

[225] Marek Śmieja et al. 'Processing of missing data by neural networks'. In: *Advances in Neural Information Processing Systems*. 2018, pp. 2719–2729.

[226] Michael R Smith, Tony Martinez and Christophe Giraud-Carrier. 'An instance level analysis of data complexity'. In: *Machine learning* 95 (2014), pp. 225–256.

[227] Valerie Smith et al. 'SWAT 1: what effects do site visits by the principal investigator have on recruitment in a multicentre randomized trial?' In: *Journal of Evidence-Based Medicine* 6.3 (2013), pp. 136–137.

[228] Alex J Smola, SVN Vishwanathan and Thomas Hofmann. 'Kernel methods for missing variables'. In: *International Workshop on Artificial Intelligence and Statistics*. PMLR. 2005, pp. 325–332.

[229] Paria Soltanzadeh and Mahdi Hashemzadeh. 'RCSMOTE: Range-Controlled synthetic minority over-sampling technique for handling the class imbalance problem'. In: *Information Sciences* 542 (2021), pp. 92–111.

[230] Maria José Sousa et al. 'Decision-making based on big data analytics for people management in healthcare organizations'. In: *Journal of medical systems* 43 (2019), pp. 1–10.

[231] Harold C Sox and Steven N Goodman. 'The methods of comparative effectiveness research'. In: *Annual review of public health* 33 (2012), pp. 425–445.

[232] Joseph A Sparano et al. 'Adjuvant chemotherapy guided by a 21-gene expression assay in breast cancer'. In: *New England Journal of Medicine* 379.2 (2018), pp. 111–121.

[233] Joseph A Sparano et al. 'Clinical and genomic risk to guide the use of adjuvant therapy for breast cancer'. In: *New England journal of medicine* 380.25 (2019), pp. 2395–2405.

[234] Nitish Srivastava et al. ‘Dropout: a simple way to prevent neural networks from overfitting’. In: *The journal of machine learning research* 15.1 (2014), pp. 1929–1958.

[235] Ewout W Steyerberg and EW Steyerberg. *Applications of prediction models*. Springer, 2009.

[236] Ewout W Steyerberg et al. ‘Prognosis Research Strategy (PROGRESS) 3: prognostic model research’. In: *PLoS medicine* 10.2 (2013), e1001381.

[237] Heung-Il Suk et al. ‘Latent feature representation with stacked auto-encoder for AD/MCI diagnosis’. In: *Brain Structure and Function* 220 (2015), pp. 841–859.

[238] Srikanth Tammina. ‘Transfer learning using vgg-16 with deep convolutional neural network for classifying images’. In: *International Journal of Scientific and Research Publications (IJSRP)* 9.10 (2019), pp. 143–150.

[239] Xiaoqing Tan et al. ‘A tree-based model averaging approach for personalized treatment effect estimation from heterogeneous data sources’. In: *International Conference on Machine Learning*. PMLR. 2022, pp. 21013–21036.

[240] Fei Tang and Hemant Ishwaran. ‘Random forest missing data algorithms’. In: *Statistical Analysis and Data Mining: The ASA Data Science Journal* 10.6 (2017), pp. 363–377.

[241] Peng Tang et al. ‘Pcl: Proposal cluster learning for weakly supervised object detection’. In: *IEEE transactions on pattern analysis and machine intelligence* 42.1 (2018), pp. 176–191.

[242] Chris Thornton et al. ‘Auto-WEKA: Combined selection and hyperparameter optimization of classification algorithms’. In: *Proceedings of the 19th ACM SIGKDD international conference on Knowledge discovery and data mining*. 2013, pp. 847–855.

[243] Yingjie Tian et al. 'Non-tumorous facial pigmentation classification based on multi-view convolutional neural network with attention mechanism'. In: *Neurocomputing* 483 (2022), pp. 370–385.

[244] Ivan Tomek. 'Two modifications of CNN.' In: (1976).

[245] Eric J Topol. 'High-performance medicine: the convergence of human and artificial intelligence'. In: *Nature medicine* 25.1 (2019), pp. 44–56.

[246] Milena Trifunovic-Koenig et al. 'Correlation between Overconfidence and Learning Motivation in Postgraduate Infection Prevention and Control Training'. In: *International Journal of Environmental Research and Public Health* 19.9 (2022), p. 5763.

[247] Ilke Tunali et al. 'Novel clinical and radiomic predictors of rapid disease progression phenotypes among lung cancer patients treated with immunotherapy: An early report'. In: *Lung Cancer* 129 (2019), pp. 75–79.

[248] Gerhard Tutz and Shahla Ramzan. 'Improved methods for the imputation of missing data by nearest neighbor methods'. In: *Computational Statistics & Data Analysis* 90 (2015), pp. 84–99.

[249] Alvaro Ulloa et al. 'An unsupervised homogenization pipeline for clustering similar patients using electronic health record data'. In: *arXiv preprint arXiv:1801.00065* (2017).

[250] Milan R Vaghasiya et al. 'The Impact of Electronic Medication Management Systems on Medication Deviations on Admission and Discharge from Hospital'. In: *International Journal of Environmental Research and Public Health* 20.3 (2023), p. 1879.

[251] Vimal B Vaghela, Amit Ganatra and Amit Thakkar. 'Boost a weak learner to a strong learner using ensemble system approach'. In: *2009 ieee international advance computing conference*. IEEE. 2009, pp. 1432–1436.

[252] Stef Van Buuren and Karin Groothuis-Oudshoorn. ‘mice: Multivariate imputation by chained equations in R’. In: *Journal of statistical software* 45 (2011), pp. 1–67.

[253] Jesper E Van Engelen and Holger H Hoos. ‘A survey on semi-supervised learning’. In: *Machine learning* 109.2 (2020), pp. 373–440.

[254] Carl Van Walraven et al. ‘Derivation and validation of an index to predict early death or unplanned readmission after discharge from hospital to the community’. In: *Cmaj* 182.6 (2010), pp. 551–557.

[255] Ashish Vaswani et al. ‘Attention is all you need’. In: *Advances in neural information processing systems*. 2017, pp. 5998–6008.

[256] Jun Wang and Jean-Daniel Zucker. ‘Solving multiple-instance problem: A lazy learning approach’. In: (2000).

[257] Lipo Wang. *Support vector machines: theory and applications*. Vol. 177. Springer Science & Business Media, 2005.

[258] Xinggang Wang et al. ‘Max-margin multiple-instance dictionary learning’. In: *International conference on machine learning*. PMLR. 2013, pp. 846–854.

[259] Xinggang Wang et al. ‘Revisiting multiple instance neural networks’. In: *Pattern Recognition* 74 (2018), pp. 15–24.

[260] Zeyuan Wang, Josiah Poon and Simon Poon. ‘Ami-net+: A novel multi-instance neural network for medical diagnosis from incomplete and imbalanced data’. In: *arXiv preprint arXiv:1907.01734* (2019).

[261] Zeyuan Wang, Josiah Poon and Simon Poon. ‘Tcm translator: A sequence generation approach for prescribing herbal medicines’. In: *2019 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*. IEEE. 2019, pp. 2474–2480.

[262] Zeyuan Wang et al. ‘A novel method for clinical risk prediction with low-quality data’. In: *Artificial Intelligence in Medicine* 114 (2021), p. 102052.

[263] Zeyuan Wang et al. ‘Attention-based multi-instance neural network for medical diagnosis from incomplete and low quality data’. In: *2019 International joint conference on neural networks (IJCNN)*. IEEE. 2019, pp. 1–8.

[264] Zeyuan Wang et al. ‘Warfarin dose estimation on high-dimensional and incomplete data’. In: (2021).

[265] Zhi-Gang Wang, Zeng-Shun Zhao and Chang-Shui Zhang. ‘Online multiple instance regression’. In: *Chinese Physics B* 22.9 (2013), p. 098702.

[266] Zhijie Wang et al. ‘Artificial intelligence based on deep learning for automatic detection of early gastric cancer’. In: *Chinese Journal of Digestive Endoscopy* (2018), pp. 551–556.

[267] Colin Wei, Sham Kakade and Tengyu Ma. ‘The implicit and explicit regularization effects of dropout’. In: *International conference on machine learning*. PMLR. 2020, pp. 10181–10192.

[268] Stephen F Weng et al. ‘Can machine-learning improve cardiovascular risk prediction using routine clinical data?’ In: *PloS one* 12.4 (2017), e0174944.

[269] Dennis L Wilson. ‘Asymptotic properties of nearest neighbor rules using edited data’. In: *IEEE Transactions on Systems, Man, and Cybernetics* 3 (1972), pp. 408–421.

[270] Peter WF Wilson et al. ‘Prediction of coronary heart disease using risk factor categories’. In: *Circulation* 97.18 (1998), pp. 1837–1847.

[271] Lijun Wu et al. ‘R-drop: Regularized dropout for neural networks’. In: *Advances in Neural Information Processing Systems* 34 (2021), pp. 10890–10905.

[272] Lei Xu and Kalyan Veeramachaneni. ‘Synthesizing tabular data using generative adversarial networks’. In: *arXiv preprint arXiv:1811.11264* (2018).

[273] Lei Xu et al. ‘Modeling tabular data using conditional gan’. In: *Advances in Neural Information Processing Systems* 32 (2019).

[274] Wenbo Xu et al. ‘Differential analysis of disease risk assessment using binary logistic regression with different analysis strategies’. In: *Journal of International Medical Research* 46.9 (2018), pp. 3656–3664.

[275] Yan Xu et al. ‘Deep learning of feature representation with multiple instance learning for medical image analysis’. In: *2014 IEEE international conference on acoustics, speech and signal processing (ICASSP)*. IEEE. 2014, pp. 1626–1630.

[276] Alexandre Yahi et al. ‘Generative adversarial networks for electronic health records: A framework for exploring and evaluating methods for predicting drug-induced laboratory test trajectories’. In: *arXiv preprint arXiv:1712.00164* (2017).

[277] Yongluan Yan et al. ‘Deep Multi-instance Learning with Dynamic Pooling’. In: *Asian Conference on Machine Learning*. 2018, pp. 662–677.

[278] Chen Yang et al. ‘Prognosis and personalized treatment prediction in TP53-mutant hepatocellular carcinoma: an in silico strategy towards precision oncology’. In: *Briefings in bioinformatics* 22.3 (2021), bbaa164.

[279] Shan You et al. ‘Learning from multiple teacher networks’. In: *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. ACM. 2017, pp. 1285–1294.

[280] Ze Yu et al. ‘Predicting Lapatinib Dose Regimen Using Machine Learning and Deep Learning Techniques Based on a Real-World Study’. In: *Frontiers in Oncology* (2022), p. 2484.

[281] Amelia Zafra et al. ‘Multi-instance genetic programming for web index recommendation’. In: *Expert Systems with Applications* 36.9 (2009), pp. 11470–11479.

[282] Mohammed J Zaki and Wagner Meira. *Data mining and analysis: fundamental concepts and algorithms*. Cambridge University Press, 2014.

[283] Zheng-Jun Zha et al. 'Joint multi-label multi-instance learning for image classification'. In: *2008 ieee conference on computer vision and pattern recognition*. IEEE. 2008, pp. 1–8.

[284] Dan Zhang et al. 'Maximum margin multiple instance clustering with applications to image and text clustering'. In: *Ieee transactions on neural networks* 22.5 (2011), pp. 739–751.

[285] Qi Zhang and Sally Goldman. 'EM-DD: An improved multiple-instance learning technique'. In: *Advances in neural information processing systems* 14 (2001).

[286] Qi Zhang and Sally A Goldman. 'EM-DD: An improved multiple-instance learning technique'. In: *Advances in neural information processing systems*. 2002, pp. 1073–1080.

[287] Xinmeng Zhang et al. 'Predicting missing values in medical data via XGBoost regression'. In: *Journal of healthcare informatics research* 4 (2020), pp. 383–394.

[288] Yan-Ping Zhang, Li-Na Zhang and Yong-Cheng Wang. 'Cluster-based majority under-sampling approaches for class imbalance learning'. In: *2010 2nd IEEE International Conference on Information and Financial Engineering*. IEEE. 2010, pp. 400–404.

[289] YONGQING ZHANG et al. 'Evolutionary-Based Ensemble Under-Sampling for Imbalanced Data'. In: *2019 16th International Computer Conference on Wavelet Active Media Technology and Information Processing*. IEEE. 2019, pp. 212–216.

[290] Zhongheng Zhang. 'Missing data imputation: focusing on single imputation'. In: *Annals of translational medicine* 4.1 (2016).

[291] Zilong Zhao et al. 'Ctab-gan: Effective table data synthesizing'. In: *Asian Conference on Machine Learning*. PMLR. 2021, pp. 97–112.

[292] Ping Zheng et al. ‘Predicting blood concentration of tacrolimus in patients with autoimmune diseases using machine learning techniques based on real-world evidence’. In: *Frontiers in Pharmacology* 12 (2021), p. 727245.

[293] Huaqiong Zhou et al. ‘Utility of models to predict 28-day or 30-day unplanned hospital readmissions: an updated systematic review’. In: *BMJ open* 6.6 (2016), e011060.

[294] Zhi-Hua Zhou. ‘A brief introduction to weakly supervised learning’. In: *National science review* 5.1 (2018), pp. 44–53.

[295] Zhi-Hua Zhou, Kai Jiang and Ming Li. ‘Multi-instance learning based web mining’. In: *Applied intelligence* 22 (2005), pp. 135–147.

[296] Zhi-Hua Zhou, Yu-Yin Sun and Yu-Feng Li. ‘Multi-instance learning by treating instances as non-iid samples’. In: *Proceedings of the 26th annual international conference on machine learning*. ACM. 2009, pp. 1249–1256.

[297] Zhi-Hua Zhou and Min-Ling Zhang. ‘Ensembles of multi-instance learners’. In: *ECML*. Vol. 3. Springer. 2003, pp. 492–502.

[298] Zhi-Hua Zhou and Min-Ling Zhang. ‘Neural networks for multi-instance learning’. In: *Proceedings of the International Conference on Intelligent Information Technology, Beijing, China*. 2002, pp. 455–459.

[299] Zhi-Hua Zhou et al. ‘Multi-instance multi-label learning’. In: *Artificial Intelligence* 176.1 (2012), pp. 2291–2320.

[300] Yu-bin Zhu et al. ‘Development of a novel individualized warfarin dose algorithm based on a population pharmacokinetic model with improved prediction accuracy for Chinese patients after heart valve replacement’. In: *Acta Pharmacologica Sinica* 38.3 (2017), pp. 434–442.

[301] Bing Zhu et al. ‘A GAN-based hybrid sampling method for imbalanced customer classification’. In: *Information Sciences* 609 (2022), pp. 1397–1411.

[302] Xiaojin Jerry Zhu. ‘Semi-supervised learning literature survey’. In: (2005).

[303] Jack E Zimmerman et al. ‘Acute Physiology and Chronic Health Evaluation (APACHE) IV: hospital mortality assessment for today’s critically ill patients’. In: *Critical care medicine* 34.5 (2006), pp. 1297–1310.

[304] Hui Zou and Trevor Hastie. ‘Regularization and variable selection via the elastic net’. In: *Journal of the royal statistical society: series B (statistical methodology)* 67.2 (2005), pp. 301–320.

List of Ph.D. Publications

Here is a list of the selected publications during my PhD study.

Journal Papers:

- Ying Y, Jia L, Wang Z, Jiang W, Zhang J, Wang H, Yang N, Wang R, Ren Y, Gao F, Ma X, Tang Y, McDonald W. Electroconvulsive therapy is associated with lower readmission rates in patients with schizophrenia[J]. Brain Stimulation, 2021, 14(4): 913-921.

- Wang Z, Poon J, Wang S, Sun S, Poon S. A novel method for clinical risk prediction with low-quality data[J]. Artificial Intelligence in Medicine, 2021, 114: 102052.

- Li Z, Yuan L, Zhang C, Sun J, Wang Z, Wang Y, Hao X, Gao F, Jiang X. A novel prognostic scoring system of intrahepatic cholangiocarcinoma with machine learning basing on real-world data[J]. Frontiers in Oncology, 2021, 10: 576901.

- Tian Y, Sun S, Qi Z, Liu Y, Wang Z. Non-tumorous facial pigmentation classification based on multi-view convolutional neural network with attention mechanism[J]. Neurocomputing, 2022, 483: 370-385.

- Huang X, Yu Z, Wei X, Shi J, Wang Y, Wang Z, Chen J, Bu S, Li L, Gao F, Zhang J, Xu A. Prediction of vancomycin dose on high-dimensional data using

machine learning techniques[J]. Expert Review of Clinical Pharmacology, 2021, 14(6): 761-771.

- Guo W, Yu Z, Gao Y, Lan X, Zang, Y, Yu P, Wang Z, Sun W, Hao X, Gao F. A machine learning model to predict risperidone active moiety concentration based on initial therapeutic drug monitoring[J]. Frontiers in Psychiatry, 2021, 12: 711868.

- Zhang C, Sun S, Tian Y, Wang Z. Structured Output Prediction Using Privileged Information[J]. IEEE Access, 2019, 7: 106065-106074.

Conference Papers:

- Wang Z, Sun S, Poon J, Poon S. CNN based multi-instance multi-task learning for syndrome differentiation of diabetic patients[C]//2018 IEEE International Conference on Bioinformatics and Biomedicine (BIBM). IEEE, 2018: 1905-1911.

- Wang Z, Poon J, Sun S, Poon S. Attention-based multi-instance neural network for medical diagnosis from incomplete and low quality data[C]//2019 International joint conference on neural networks (IJCNN). IEEE, 2019: 1-8.

- Wang Z, Poon J, Poon S. Ami-net+: A novel multi-instance neural network for medical diagnosis from incomplete and imbalanced data[J]. Australian Journal of Intelligent Information Processing Systems, 8.

- Wang Z, Poon J, Poon S. Tcm translator: A sequence generation approach for prescribing herbal medicines[C]//2019 IEEE International Conference on Bioinformatics and Biomedicine (BIBM). IEEE, 2019: 2474-2480.

- Wang Z, Poon J, Yang J, Poon S. Warfarin dose estimation on high-dimensional and incomplete data[J]. 2021.