Doctoral Thesis

# Research on Two-Dimensionalization Algorithms for Improving Emotion Recognition Accuracy in Speech Data and its Evaluation of Generalized Deployment (感情認識精度向上のための音声データの二次元化アルゴリズムの研究およびその汎用展開の評価)

Zijun Yang

November 22, 2024

Abstract

With the development of society and the intensification of competition, people face increasing life pressure in their daily life, which has a significant impact on the mental health of individuals. This study is dedicated to exploring how this psychosocial health issue can be attended to and addressed through speech emotion recognition. Speech, as a natural and intuitive way of expressing emotions, has been found to contain up to 38% of emotional information. Through in-depth sentiment analysis of speech, we can better understand the emotional state of individuals and provide feedback accordingly, thus helping to alleviate the stress they face in life In this study, we innovatively start from speech and use a novel time series analysis method to transform speech time series into 2D images. In this process, we employed Hilbert curves to map the time series to the image space. In this way, we successfully capture the dynamic features of speech into static images, which lays the foundation for subsequent emotion recognition. In order to realize the accurate recognition of speech emotion, we designed a neural network suitable for this image representation. This neural network can effectively extract the key features in the image, thus realizing the recognition of different emotions. Through a large number of experiments, we have proved that our method has achieved remarkable results in speech emotion recognition, providing a solid foundation for further research and application. Not only that, this study also optimizes other time series imaging algorithms. We improved the Gram's Corner Field algorithm by using different downsampling techniques and designed a neural network model for Gram's Corner Field. This optimization makes our method more versatile and able to adapt to different time series data, providing a wider range of possibilities for future applications. In order to understand the individual's emotional state more comprehensively, this study introduces the CyTex method in the extension of the method and incorporates the concept of speech rate for the segmentation of time series. This innovative approach further improves the accuracy of speech emotion recognition and lays a solid foundation for future applications. In the segmentation processing of time series, we adopt the CyTex method, which effectively divides the time series while maintaining its continuity. This segmentation allows the neural network to learn the emotional information in each time period more precisely. Compared with the traditional holistic learning method, segmentation learning is more capable of capturing the subtle differences of emotional changes in speech, making the recognition results more accurate. At the same time, we introduce the concept of speech rate as a new analytical dimension to be incorporated into the time-series features. Speech rate is not only a surface feature of speech, but it also combines short-time features and rhythmic features to reflect the emotional information in speech more comprehensively. By considering speech rate in segmentation learning, we enable the neural network to be more sensitive to capturing emotional changes in speech,

thus improving the accuracy of recognition. This approach experimentally demonstrates that the segmental learning approach, which introduces CyTex and speech rate, performs well in the speech emotion recognition task compared to the traditional holistic learning approach. This provides a more refined and accurate processing means for future speech emotion recognition applications and lays a more solid foundation for practical applications. Therefore, by adopting the CyTex method and introducing the concept of speech rate, we analyze the time series more carefully, which makes our algorithm achieve more satisfactory results in the emotion recognition task. This innovative approach provides new perspectives and methods in the field of speech emotion recognition and brings wider possibilities for future research and applications. This research transcends the confines of speech emotion recognition, extending its applicability to the realm of brainwave analysis. The methodologies, initially designed for speech, prove to be versatile as they are successfully applied to brainwave time series, achieving remarkable results in the identification of distinct epileptic seizure types. This breakthrough not only signifies the adaptability and efficacy of the proposed methods but also opens new avenues for applications in neurology and clinical diagnostics. In achieving excellence in epileptic seizure type recognition, the study sets the stage for future endeavors aimed at identifying depressive states and discerning emotional nuances through brainwave analysis. The envisioned expansion of research activities in these directions reflects the commitment to pushing the boundaries of knowledge and practical applications in mental health research. This forward momentum not only enhances our understanding of neurological disorders but also holds promise for the development of novel diagnostic tools and therapeutic interventions. The exploration of brainwave signals emerges as a powerful avenue for gaining profound insights into an individual's mental state and emotional experiences. Through meticulous analysis of brainwave patterns, this study provides a nuanced understanding of cognitive processes, presenting itself as a valuable tool for researchers in psychology and neuroscience. The nuanced nature of brainwave data offers a rich tapestry of information, shedding light on the intricate interplay of emotions and mental states. In conclusion, this study, with a comprehensive scope spanning speech emotion recognition to brainwave analysis, has reached a pivotal milestone by excelling in epileptic seizure type identification. The transformative methodologies introduced in speech analysis seamlessly extend to the realm of brainwave time series, opening up new vistas for exploration. The fusion of innovative approaches with optimized time series imaging algorithms not only enables accurate emotional state recognition but also propels the research landscape into promising territories within neurology and mental health. With a commitment to ongoing research, the study serves as a beacon for future investigations, offering a wealth of tools and insights for understanding, mitigating, and addressing various aspects of individual life stress, mental health, and neurological disorders.

# Contents

# List of Figures

# List of Tables

# Chapter 1

# Introduction

## 1.1 background

Understanding the intricate relationship between social dynamics and mental well-being is essential in contemporary society. As individuals navigate through the complexities of modern life, they encounter various stressors originating from social interactions, societal expectations, and personal experiences. These stressors, collectively known as social stress, have a profound impact on mental health, shaping individuals' emotional and psychological states. This section delves into the multifaceted nature of social stress and its implications for individual well-being.

### 1.1.1 The impact of social stress on individual mental health

An increase in social pressure leads to psychological stress in individuals [1]. Accumulated mental stress adversely affects mental health and results in various diseases [2]. Mental health issues are pervasive globally. Fig. 1.1 displays the global prevalence of the top5 mental disorders in 2019, as estimated by the World Health Organization (WHO)[3]. The estimates are derived from representative surveys, medical data, and statistical modeling, reflecting the proportion of individuals affected by the most prevalent mental illnesses in 2019. As of 2019, the prevalence of the top 5 mental illnesses accounts for 8.2% of the global population. Specifically, these prevalent mental health disorders include:

1. **Anxiety disorders** (3.8%): Anxiety disorders encompass a range of conditions characterized by excessive worry, fear, or nervousness, which can interfere with daily life and functioning. Common subtypes include generalized anxiety disorder (GAD), panic disorder, and social anxiety disorder. Anxiety disorders involve more than temporary worry or fear, and the symptoms can interfere with daily activities such as job performance, schoolwork, and relationships [4].
2. **Depressive disorders** (3.4%): Depression is a mood disorder marked by persistent feelings of sadness, hopelessness, and a lack of interest or pleasure in activities.

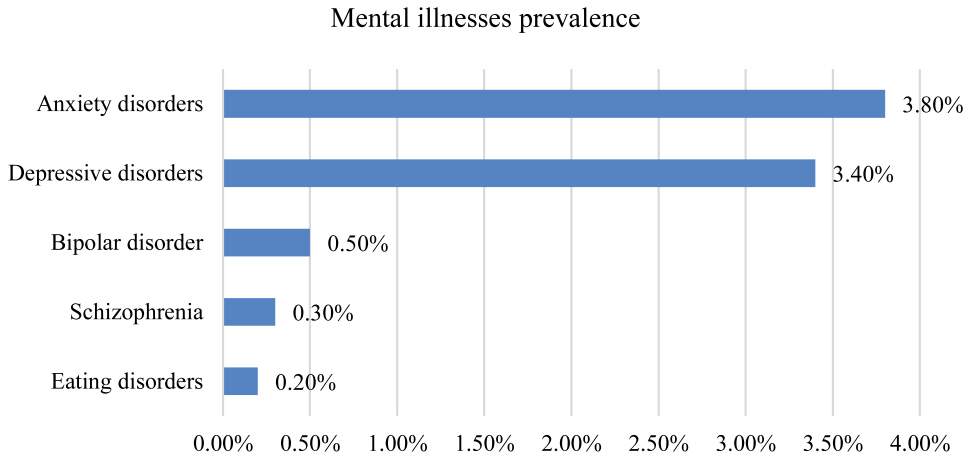



Fig. 1.1. Mental illnesses prevalence(World, 2019) [3].
The estimated share of people with each mental illness in a given year, whether or not they were diagnosed, based on representative surveys, medical data and statistical modeling.

It can affect one's ability to carry out daily tasks and maintain relationships [5]. Worldwide, depression is a major cause of disability and premature death. Depression is prevalent among all age groups and in almost all walks of life. It may range from a very mild condition to severe (psychotic) depression accompanied by hallucinations and delusions. The potential causes of depression can include societal expectations and pressure, physical health problems, transition to parenthood, social connectedness, personality and past psychological history, child health and temperament challenges, unmet care needs, unmet expectations for childbirth, and other life stressors [6].

3. **Bipolar disorder**(0.5%) : Bipolar disorder involves extreme mood swings, ranging from episodes of elevated mood and energy (mania or hypomania) to periods of depression. These mood shifts can significantly impact a person's daily life [7]. The mechanism by which mood episodes with completely opposite characteristics appear repeatedly in patients with bipolar disorder remains unknown, and true mood-stabilizing drugs effective for treating both manic and depressive episodes currently do not exist. Mood monitoring is widely used in the treatment and self-management of bipolar mood swings to help individuals achieve greater awareness and understanding of their affective states, which enables them to better prepare and account for problematic changes in their mood, preventing escalation to mood episodes and relapse [8].
4. **Schizophrenia** (0.2%) : Schizophrenia is a severe mental disorder characterized by distorted thinking, hallucinations, delusions, and impaired emotional responses. It often leads to disruptions in social and occupational functioning [9]. Negative

symptoms (e.g., anhedonia, amotivation, and expressive deficits) are refractory to current medications and are among the foremost causes of disability in schizophrenia. A study used a two-step approach to identify and empirically test a brain network model of schizophrenia symptoms. The results demonstrated that a connectivity breakdown between the cerebellum and the right dorsolateral prefrontal cortex directly corresponded to negative symptom severity. Restoration of network connectivity with transcranial magnetic stimulation corresponded to amelioration of negative symptoms, showing a statistically significant strong relationship of negative symptom change in response to functional connectivity change.

5. **Eating disorders** (0.1%) : eating disorders encompass conditions such as anorexia nervosa, bulimia nervosa, and binge-eating disorder. These disorders involve disturbances in eating habits, body image concerns, and often lead to severe physical and mental health consequences. Anorexia nervosa is characterized by restricted eating, intense fear of gaining weight, and a distorted body image. Bulimia nervosa involves recurrent episodes of binge eating followed by compensatory behaviors such as vomiting or excessive exercise. Binge-eating disorder is characterized by recurrent episodes of eating large quantities of food, often rapidly, to the point of discomfort, without the purging behaviors seen in bulimia nervosa. These disorders can have serious health consequences, including nutritional deficiencies, electrolyte imbalances, and damage to the digestive system. Treatment for eating disorders often involves a combination of medical care, nutritional counseling, and therapy to address the underlying psychological issues.

Unfortunately, many societies and healthcare systems frequently overlook this aspect, failing to provide the necessary medical care and support that individuals require [10]. Mental health problems have profound and extensive impacts on society. From an individual standpoint, these issues heighten risks for individuals, families, and communities. At a societal level, the prevalence of mental health problems amplifies risks for entire populations and may hinder global efforts to enhance people's well-being. Fig. 1.2 indicates the worldwide burden caused by the top 5 mental illnesses in 2019. The data in the figure represent the estimated number of disability-adjusted life years (DALYs) per 100,000 people for each category of mental illness.

In this regard, the main current threats from mental include:

1. **Economic downturns and social divisions**: Economic downturns and conditions of social inequality may contribute to an upsurge in mental health problems. Factors such as unemployment, poverty, and social exclusion have been identified as having a detrimental impact on an individual's mental well-being [11, 12]. The costs associated with lost productivity and other indirect social ramifications often far surpass medical expenses. From an economic standpoint, schizophrenia stands

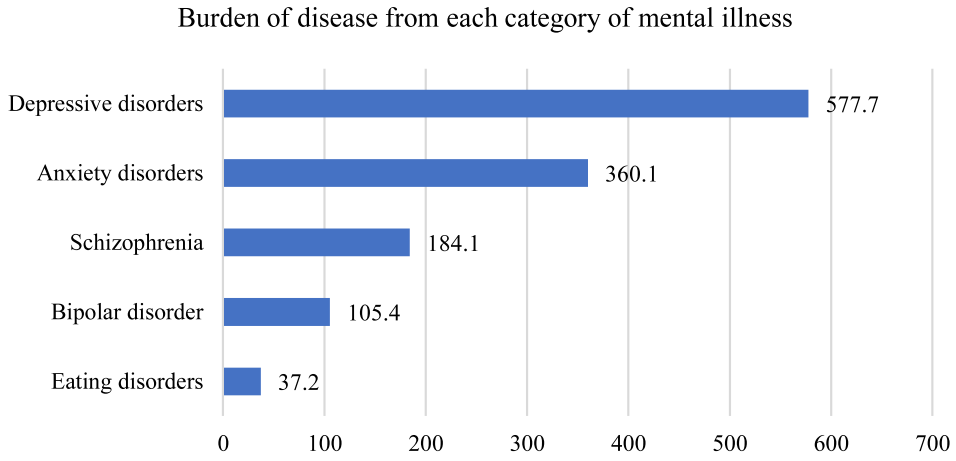



Fig. 1.2. Burden of disease from each category of mental illness (World, 2019) [3]. Estimated number of disability-adjusted life years (DALYs) per 100,000 people due to each category of mental illness.

out as the mental disorder incurring the highest per capita costs to society [13]. Although depression and anxiety disorders exhibit lower per capita costs, their higher prevalence means they make the most significant contribution to total national costs [14].

With the escalation of economic and social pressures, there is a parallel increase in workplace psychological stress, posing escalating risks to individuals' mental health. Workplace mental health risks encompass insufficient skills, excessive workload, prolonged working hours, precarious working conditions, limited co-worker support, incidents of violence, harassment, and discrimination [15]. It is noteworthy that over half of the global workforce engages in informal employment, often lacking health and safety regulations. Such workers frequently operate in unsafe environments, endure extended work hours, lack social and economic security, and confront discrimination, all of which can detrimentally impact their mental health.

2. **Public health emergencies**: Large-scale public health events, such as pandemics, not only pose a threat to physical health, but can also trigger mental health problems. Social isolation, fear and uncertainty can all have an impact on people's mental health [16]. Public health emergencies, such as the COVID-19 pandemic, have significant implications for mental health. These emergencies can lead to long-term emotional distress, particularly for historically medically underserved and socially marginalized populations, as well as frontline healthcare workers. The impact of public health emergencies on mental health can manifest in various ways, including pre-existing mental disorders, emergency-induced grief and acute stress reactions, and humanitarian response-induced anxiety [17]. The demand for mental health services has increased during public health emergencies, with a notable rise in mental

illness among adults in the United States [18]. Even as public health emergencies end, the mental health crisis continues, necessitating ongoing adaptation and provision of mental health care services to communities. Research indicates that the mental health impact of economic crises can be mitigated by countries with strong social safety nets, suggesting that social protection responses are crucial in addressing poor mental health during economic downturns [12].

3. **Humanitarian crises and displacement**: Wars, natural disasters and other humanitarian crises often lead to displacement and trauma. The risk of mental health problems increases significantly in these situations [19].
4. **Climate crisis**: Climate change-induced extreme weather events and environmental damage may lead to increased social instability and mental health problems. There is a complex interrelationship between climate change and mental health [20].

Economic downturns, public health emergencies, humanitarian crises, and climate crisis not only elevate the risk of unemployment and economic instability but also contribute to a reduction in job opportunities. Discrimination and inequality in the workplace, stemming from factors such as race, gender, gender identity, sexual orientation, disability, social origin, immigration status, religion, or age, can be exacerbated by the work environment. Individuals with severe mental health problems often face exclusion from employment and encounter disparities in the workplace. Unemployment itself is a significant factor that can increase the vulnerability to mental health issues.

Transitioning to the global adolescent population, approximately one-sixth falls within the age group of 10 to 19 years, known as adolescence. This phase is characterized by substantial physical, emotional, and social changes. Adolescents may also confront challenges like poverty, abuse, or violence, increasing the likelihood of mental health problems. Globally, one in seven adolescents experiences mental disorders, constituting 13% of the global disease burden in this age group. Leading causes of illness and disability among adolescents include depression, anxiety, and behavioral disorders, with suicide ranking as the fourth leading cause of death among 15- to 29-year-olds. The repercussions of neglecting adolescent mental health extend into adulthood, affecting physical and mental health, and limiting opportunities for fulfilling lives. Mental health problems are not only common among teenagers, but one in eight people worldwide suffer from a mental disorder. This is an issue that cannot be ignored.

In summary, mental health challenges extend beyond the individual level and have broader societal implications. About one-eighth of the global population grapples with mental disorders, and individuals with severe mental health problems face premature mortality—up to twenty years earlier [21]. Mental disorders account for one-sixth of all global deaths [22]. Suicide, affecting people from all countries, backgrounds, and ages, is a pervasive issue globally, with each suicide potentially accompanied by 20 suicide attempts

[23]. Notably, suicide has become the leading cause of death among young people. The urgency to address these challenges is paramount for both individual and societal well-being.

### 1.1.2 The importance of Speech Emotion Recognition

Emotions play a crucial role in human life [24] and are an integral part of psychological survival [25]. Emotion detection research is evolving through collaborative research in the fields of psychology, cognitive science, machine learning, and natural language processing (NLP)[26]. In early studies, scientists such as Darwin [27]), considered emotional expression to be the last behavioral pattern preserved in human evolution.

Emotional expression serves as a crucial component of human communication, allowing individuals to convey their internal states and connect with others. Understanding and interpreting emotions have been essential for human species' social cohesion and adaptability. Early in human evolution, the ability to recognize and respond to emotions likely played a pivotal role in survival, enabling individuals to navigate complex social dynamics and environmental challenges. In the contemporary context, the study of emotions has expanded beyond traditional disciplines, incorporating advanced technologies and interdisciplinary approaches. The intersection of psychology and computational sciences, particularly machine learning and NLP, has given rise to sophisticated tools for emotion detection and analysis. As technology continues to advance, the application of emotion detection has diversified across various domains, including human-computer interaction, sentiment analysis, mental health diagnostics, and virtual communication platforms. Researchers and practitioners in these fields collaborate to enhanc ehuman understanding of emotions, develop more accurate detection methods, and explore innovative applications that positively impact individuals' well-being. The exploration of emotions and the development of emotion detection techniques represent a rich interdisciplinary endeavor. From its roots in evolutionary psychology to contemporary collaborations between diverse fields, the study of emotions reflects the intricate relationship between human behavior, technology, and societal advancements.

Emotion detection techniques have evolved significantly with advancements in technology, drawing on interdisciplinary approaches to understand and interpret human emotions. Various methods and technologies contribute to the field of emotion detection, each with its strengths and applications:

1. **Facial Expression Analysis**: Facial Expression Analysis involves the utilization of computer vision algorithms to analyze facial features, including eyes, mouth, and brows, for the identification and categorization of emotions. This method plays a pivotal role in various applications such as enhancing human-computer interaction, conducting market research, and performing sentiment analysis by discerning emo-

tional states based on facial expressions. The precision of this technique contributes to its effectiveness in interpreting human emotions in diverse contexts. Common techniques for facial expression analysis include computer vision-based methods that analyze facial features such as eyes, mouth, and brows to determine emotional states. Utilizing neural networks, particularly convolutional neural networks (CNNs), helps learn and recognize patterns of facial expressions [28]. Heatmap analysis associates temperature variations in different facial regions with emotional states [29]. Three-dimensional facial modeling provides more accurate facial movement information for precise emotional state inference [30]. Deep learning methods, such as recurrent neural networks (RNNs) or long short-term memory networks (LSTMs), enable time-series analysis of facial expressions [31]. These techniques can be used individually or in combination, allowing flexibility in choosing the most suitable approach based on specific contexts and requirements. The continuous development and refinement of these methods contribute to the widespread application of facial expression analysis in various fields.

2. **Speech Analysis**:
   Speech analysis involves scrutinizing vocal cues, including tone, pitch, and speech patterns, to deduce emotional states. This process employs Natural Language Processing (NLP) algorithms that meticulously process spoken language, enabling the detection and interpretation of emotions. This technique finds application in various domains, such as call center analytics, where understanding customer emotions is crucial for providing effective service. Additionally, virtual assistants utilize speech analysis to comprehend user sentiments and respond accordingly. In the realm of emotion-aware technology, this method enhances human-computer interaction by allowing systems to adapt based on the emotional context expressed through speech.
3. **Biometric Sensors**: Biometric sensors are instrumental in emotion detection by harnessing physiological data, including metrics like heart rate, skin conductance, and body temperature, to discern emotional responses. These sensors, often integrated into wearable devices, meticulously capture physiological changes that correlate with distinct emotions. For instance, in stress management, these sensors can track elevated heart rates or increased skin conductance as indicators of heightened stress levels [32, 33]. In mental health monitoring, deviations in body temperature might signify emotional distress. Moreover, in the realm of personalized user experiences, biometric sensors can adapt interfaces based on detected emotions, providing tailored interactions for enhanced well-being [34].
4. **Text Analysis**: Text analysis, specifically the application of Natural Language Processing (NLP) algorithms, involves the examination of written text to discern emotional content. Through sophisticated linguistic analysis, these algorithms take

into account factors such as sentiment, tone, and contextual cues within the text [35]. This method finds diverse applications, including sentiment analysis in social media to gauge public reactions to events, customer feedback analysis for businesses to understand consumer sentiments, and content moderation on online platforms to filter out inappropriate or harmful content [36]. For instance, sentiment analysis tools can determine whether social media posts express positive or negative emotions, aiding in assessing public opinion about a product, service, or current affairs. Similarly, businesses can utilize customer feedback analysis to identify areas for improvement based on the emotional tone of customer reviews [37, 38]. In content moderation, NLP algorithms can flag and filter out content that may be offensive or violate platform guidelines by analyzing the emotional nuances in the text. Overall, text analysis with NLP is a powerful tool for understanding and managing emotional aspects of written communication in various contexts.

5. **Brain-Computer Interfaces (BCIs)**: Brain-Computer Interfaces (BCIs) involve the measurement of brain activity to comprehend emotional responses [39]. This is achieved through technologies such as Electroencephalogram (EEG) and Functional Magnetic Resonance Imaging (fMRI), which capture distinct brain signals associated with various emotions [40]. For instance, in neurofeedback therapy, BCIs use real-time brain activity feedback to help individuals regulate their emotional states, contributing to mental health interventions [41]. In the gaming industry, BCIs enhance player experiences by allowing direct interaction with the game environment based on their emotional states, creating a more immersive and personalized gaming experience [42]. Additionally, BCIs are instrumental in human-machine interfaces, where the interpretation of emotional signals can improve the interaction between individuals and machines, leading to more intuitive and responsive technology interfaces.
6. **Gait Analysis**: Gait analysis involves the examination of body movements and posture to infer emotional states [43]. This method utilizes computer vision algorithms to scrutinize the way individuals walk or move, extracting valuable information about their emotional well-being [44]. For example, in surveillance applications, gait analysis can be employed to identify suspicious or distressed behaviors, enhancing security measures [45]. Additionally, in human-computer interaction, understanding the user's emotional state through gait analysis allows for more personalized and responsive systems [46]. This technology finds practical application in security contexts, where abnormal gait patterns might indicate potential threats or distress, contributing to the development of proactive security measures.
7. **Deep Learning Models**: Deep Learning Models for emotion detection involve the utilization of neural networks to autonomously learn and extract features associated with emotional states [47]. By training these models on extensive datasets,

they become adept at recognizing intricate patterns indicative of various emotions. This approach finds applications in diverse fields such as image and video analysis, where it can accurately identify and interpret emotional cues within visual content. Additionally, deep learning models contribute to enhancing autonomous systems, enabling them to perceive and respond to human emotions, and they play a pivotal role in the development of immersive experiences in virtual reality by enabling systems to understand and adapt to users' emotional states.

8. **Multimodal Fusion**: Multimodal fusion in emotion detection involves the integration of data from diverse sources, such as facial expressions, speech, and physiological signals, to enhance the robustness of emotion analysis [48]. By combining information extracted from different modalities, this approach aims to achieve higher accuracy and reliability in understanding human emotions [49]. For instance, in human-computer interaction scenarios, a system employing multimodal fusion could analyze facial expressions, voice intonation, and physiological responses concurrently to better interpret the user's emotional state [50]. In immersive experiences like virtual reality, combining visual, auditory, and physiological data enables more nuanced emotional feedback, creating a more engaging and responsive environment [51]. Additionally, in affective computing applications, multimodal fusion contributes to a more comprehensive understanding of emotional responses, leading to improved adaptive systems and personalized user experiences.

Emotion detection techniques continue to advance, driven by ongoing research in artificial intelligence, machine learning, and human-computer interaction. The advantages and applications of each method and technique are shown in the Table 1.1. These techniques play a crucial role in creating emotionally intelligent systems and applications that better understand and respond to human emotions. Speech signaling is an important mode of emotional expression that accounts for 38% of emotional communication [52]. This function plays a crucial role in the recognition and communication of emotions [53]. Speech emotion recognition (SER) is a branch of emotion detection [54]. It has more than two decades of research history, and has accumulated numerous research results [55]. This subject focuses on the recognition of emotions in speech without considering semantic content [56, 57].

The significance of SER lies in its widespread applications across the fields of human-computer interaction, emotion-aware technology, and psychology. Here are several aspects that emphasize the importance of SER:

1. **Human-Computer Interaction**: In the realm of human-computer interaction, SER plays a crucial role in enhancing communication between computers and users. The ability to recognize emotions in speech contributes to the development of emotionally intelligent systems that can understand and adapt to the user's emotional

Table 1.1. The advantages and applications of each method and technique

| **Method** | **Definition** | **Applications** |
|---|---|---|
| Facial Expression Analysis | Analyzing facial expressions to identify and categorize emotions. | Human-computer interaction, market research, sentiment analysis |
| Speech Analysis | Examining vocal cues, tone, pitch, and speech patterns to infer emotional states. | Call center analytics, virtual assistants, emotion-aware technology |
| Biometric Sensors | Using physiological data to infer emotional responses. | Stress management, mental health monitoring, personalized user experiences |
| Text Analysis | Analyzing written text to identify emotional content. | Sentiment analysis in social media, customer feedback analysis, content moderation |
| Brain-Computer Interfaces (BCIs) | Measuring brain activity to understand emotional responses. | Neurofeedback therapy, gaming, human-machine interfaces |
| Gait Analysis | Studying body movements and posture to infer emotional states. | Surveillance, human-computer interaction, security |
| Deep Learning Models | Leveraging neural networks for emotion detection. | Image and video analysis, autonomous systems, virtual reality |
| Multimodal Fusion | Integrating data from multiple sources for more robust emotion detection. | Human-computer interaction, immersive experiences, affective computing |

state [58]. In the field of human-computer interaction, SER is a key technology that enhances the gap between computers and users and completely changes the way human interact with machines. If a virtual assistant with advanced SER capabilities to interact with users in natural language. As the user communicates, the SER system discerns the emotional nuances in their speech, detecting elements of joy, frustration, or urgency. In response, the virtual assistant not only comprehends the user's requests but also adapts its tone, language, and responses to align with the user's emotional state.For instance, if the SER system identifies signs of stress or frustration, the virtual assistant might offer calming responses or provide solutions with heightened empathy. This dynamic adaptation transforms the interaction into a more personalized and emotionally intelligent experience, making technology feel more attuned and responsive to the user's needs. This capability extends to other HCI applications, such as smart home devices that can adjust their behavior based on the user's emotional cues, creating a seamless and emotionally aware human-computer interaction.

2. **Customer Service and Market Analysis**: In the business domain, understanding the emotions conveyed during communication is essential for improving service quality. SER can be employed to analyze emotions in phone conversations or customer service dialogues, aiding businesses in gaining insights into customer needs, concerns, and satisfaction [59]. This helps in refining customer service, adjusting marketing strategies, and increasing customer loyalty. In the domain of Customer Service and Market Analysis, Speech Emotion Recognition (SER) proves to be instrumental in enhancing customer interactions and gaining valuable insights for businesses. For instance, call centers use SER to analyze the sentiment conveyed during customer-agent conversations. By accurately detecting customer sentiments, businesses can identify areas of improvement in their service quality, promptly address customer concerns, and tailor their responses to ensure a more positive customer experience. Additionally, in market analysis, SER can be applied to analyze public opinions expressed in recorded customer feedback, reviews, or interviews. This allows businesses to gauge the overall sentiment towards their products or services, providing actionable data for refining marketing strategies and maintaining a competitive edge in the market. Through the integration of SER in customer service and market analysis, businesses can not only enhance customer satisfaction but also make data-driven decisions that positively impact their overall performance and market positioning.
3. **Psychological Research**: In the fields of psychology and cognitive science, SER provides a tool for gaining in-depth insights into emotional expression and communication [60]. By analyzing emotions in speech, researchers can better understand individual emotional responses in different contexts and explore the relationship

between emotion and language. By analyzing emotions in speech, researchers can more fully understand individuals' emotional responses in different environments, as well as the complex relationship between emotion and language. For example, by analyzing the speech of individuals in different social situations, researchers can explore the manifestations of social anxiety or emotional disorders, thereby providing deeper insights into the diagnosis and treatment of related mental illnesses. In addition, by simulating changes in specific emotional states in a laboratory environment through SER, researchers can experimentally explore the connection between emotion and psychological processes such as cognition and memory, providing empirical support for the development of psychological theories. Therefore, the application of speech emotion recognition in psychological research not only helps to deepen the understanding of the mechanism of emotional expression, but also provides a new, non-invasive tool for various research in the field of psychology.

4. **Emotional Health Monitoring**: SER can be utilized for monitoring emotional health on an individual level. By analyzing speech features, systems can detect potential emotional issues such as depression or anxiety. This provides an opportunity for early intervention and support, contributing to the improvement of an individual's mental health [61]. By analyzing the emotional characteristics in speech, the system can identify an individual's emotional state and provide targeted emotional health monitoring. For example, a voice emotion recognition system could detect signs of depression or anxiety in speech, providing individuals with opportunities for early intervention and support. This is particularly important for individuals who may be experiencing emotional health challenges, as the system can help improve an individual's mental health by identifying problems early and prompting professional help to be sought. Through accurate monitoring of voice emotion recognition, human can pay more comprehensive attention to the emotional state of individuals and provide them with better mental health support.
5. **Education and Training**: In the education sector, SER can assess students' emotional states and engagement. This helps educators tailor teaching strategies, provide personalized learning experiences, and detect challenges students might be facing [62]. SER holds substantial importance by providing valuable insights into students' emotional states and engagement levels. For instance, during online learning sessions, SER can analyze students' speech patterns to discern emotions such as frustration, enthusiasm, or disinterest. This information enables educators to adapt their teaching methods accordingly, offering personalized support to students who may be struggling or tailoring challenges for those in need of additional stimulation. By integrating SER into educational technologies, institutions can create emotionally intelligent systems that enhance the overall learning experience, contributing to more effective and student-centric education. This not only facilitates a better un-

derstanding of students' individual needs but also fosters a positive and supportive learning environment.

In conclusion, SER's applications in various domains underscore its critical role in enhancing technological intelligence, improving services, and advancing research. By understanding and leveraging emotional information in speech, human can build more human-centric and intelligent systems that better meet the needs of individuals and society.

Hence, this research introduces a novel algorithm centered on speech emotion recognition, aiming to enhance the precision of emotional analysis in speech. Additionally, optimizations were applied to existing algorithms. Through empirical investigations, it was ascertained that the newly proposed algorithm not only excelled in speech emotion recognition but also exhibited adaptability for other time series recognition tasks. This underscores the algorithm's versatility and broadens its applicability beyond the realm of speech emotion recognition, showcasing its potential impact in diverse time series recognition domains.

## 1.2 Related work

The primary objective of this study is to delve into the classification of speech emotions. Within the existing body of research, nine algorithms have emerged as pivotal contributors to the field of time series classification and have made great contributions to speech emotion classification. These algorithms play a crucial role in deciphering and categorizing emotions embedded in speech patterns.

1. **1NN-RAW**: "1NN-RAW" stands for "One Nearest Neighbor - Raw", representing a straightforward machine learning algorithm based on nearest neighbor classification. It serves as one of the most elementary nearest neighbor algorithms. The term "1NN" signifies "1-Nearest Neighbor", indicating that, given a specific data point, the algorithm identifies the closest neighboring data point and utilizes its label to predict the label of the given data point. The inclusion of "RAW" emphasizes that the algorithm employs raw, unprocessed data during its predictive process. The functionality of the algorithm unfolds as follows: when presented with a labeled training dataset, each data point within it possesses a feature vector and an associated label. When confronted with a new, unlabeled data point requiring classification, the 1NN-RAW algorithm computes the distance (typically using the Euclidean distance or another distance metric) between this new data point and each data point in the training set. Subsequently, it singles out the closest data point in the training set, referred to as the Nearest Neighbor data point (1-Nearest Neighbor). Ultimately, the 1NN-RAW algorithm attributes the label of the selected Nearest Neighbor data point to the new data point as its predictive label. While

the 1NN-RAW algorithm is straightforward and easy to implement, it comes with certain limitations. It is highly sensitive to noisy data, as an inaccurate prediction may result if the nearest neighbor data point is a noisy point. Additionally, the efficiency of the 1NN-RAW algorithm may be compromised when dealing with high-dimensional data or large-scale datasets, given the need to calculate the distance to each training data point. In practice, more sophisticated variants of nearest neighbor algorithms, such as K-Nearest Neighbors (KNN), are often considered to address some of the drawbacks associated with 1NN-RAW. KNN incorporates the votes of multiple nearest neighbor data points for classification, offering better resilience to noisy data points and enhanced handling of high-dimensional data.

2. **1NN-DTW-BWW**: The fundamental approach of 1NN-DTW-BWW involves the integration of the 1NN method with DTW and BWW. DTW serves as a mechanism to gauge the similarity between two time series, particularly when there are variations in their speeds or lengths. It assesses the distance between data points in a time series by dynamically adjusting their alignment. The BWW method is employed to select the optimal alignment window in DTW, defining segments of time series that can be stretched or compressed during the DTW calculation. Optimal window selection enhances the performance of time series classification. In the 1NN-DTW-BWW algorithm, the process begins by computing the DTW distance between a new time series and all time series within a known category. Subsequently, it identifies the closest time series within the known category. Throughout this procedure, the BWW method plays a crucial role in selecting the most suitable alignment window, ensuring that the calculation of DTW considers the similarity of time series. This algorithm finds extensive application in various domains, including bioinformatics for gene sequence classification and protein structure identification, medicine for disease diagnosis and vital signs monitoring, finance for stock price prediction and trade analysis, engineering for fault detection and predictive maintenance, and natural language processing for tasks such as text classification and speech recognition.
3. **1NN-DTW-nWW**: 1NN-DTW-nWW is a time series classification algorithm that bears similarity to 1NN-DTW-BWW, but it adopts a distinct alignment window selection strategy known as nWW. Unlike the BWW strategy, nWW is employed to enhance the classification performance by considering the structural characteristics of time series more effectively. The algorithm initially computes the Dynamic Time Warping (DTW) distance between a new time series and all known-category time series. Subsequently, it selects the known-category time series with the closest distance. Throughout this process, the nWW method is instrumental in aligning the window, ensuring a more comprehensive consideration of the time series structure and consequently improving classification performance. Primarily employed for time

series classification, the 1NN-DTW-nWW algorithm shares its application domains with 1NN-DTW-BWW, finding utility in various fields.

4. **SAX BoP** [63]: SAX BoP (Symbolic Aggregate approXimation Bag of Patterns) represents an innovative approach to time series classification by amalgamating two distinct techniques, namely SAX (Symbolic Aggregate approXimation) and BoP (Bag of Patterns). SAX operates as a time series compression and dimensionality reduction method that transforms continuous time series data into discrete symbolic sequences. This compression is achieved through the segmentation of time series data, assigning a symbol to each segment, thereby reducing the dimensionality while retaining crucial features. On the other hand, BoP, a machine learning method widely employed in image and natural language processing, serves as a feature extraction and classification tool. In the context of time series, BoP is utilized to characterize the set of patterns within a time series that are pivotal for distinguishing various time series classes. The SAX BoP algorithm unfolds in several steps: initially, the SAX technique is applied to convert time series data into symbolic sequences. Subsequently, a Bag of Patterns is constructed, encompassing significant patterns or symbolic sequences within the time series. Finally, these patterns are utilized for representation and classification of the time series. The distinctive advantage of the SAX BoP algorithm lies in its ability to transform time series data into a more interpretable collection of patterns, enhancing the differentiation between various time series classes. It emerges as a valuable tool in the realm of time series analysis and classification, contributing to improved classification performance and pattern recognition.
5. **Fast Shapelet** [64]: The Fast Shapelet algorithm is designed for efficient mining of shape subsequences within time series datasets. Its primary objective is to identify the most representative shape subsequences, enhancing classification performance. In this context, a shape subsequence refers to a segment of a time series that exhibits a distinct shape or pattern, serving as a representative feature for the entire series. Typically short yet meaningful, shapelets contribute to distinguishing between different classes of time series data. The algorithm achieves this objective through a series of computational and filtering steps, rapidly pinpointing the most crucial shapelets. Its primary application lies in time series classification, with potential applications spanning various domains. The notable advantage of the Fast Shapelet algorithm is its ability to extract key features from time series data, ultimately improving classification accuracy. This makes it a valuable tool in the realm of time series analysis and classification, aiding in the extraction of meaningful information from intricate time series datasets.
6. **RPCD** [65]: RPCD (Random Projection for Classification of Time Series Data) is a method designed for time series classification that leverages random projection to

decrease the dimensionality of time series data. The key mechanism employed by RPCD is the random projection technique, which transforms high-dimensional time series data into a lower-dimensional representation. Random projection, as a dimensionality reduction method, effectively preserves the distance information between data points. By implementing random projection, RPCD achieves a reduction in the dimensionality of the original time series data while retaining crucial differential features. This reduction not only lessens the computational complexity of subsequent classification tasks but also holds the potential to enhance classification accuracy. Following the dimensionality reduction step, RPCD employs various classification algorithms such as k-Nearest Neighbors (k-NN), Support Vector Machines (SVM), decision trees, or other machine learning models. The dimensionality-reduced data serves as the input for the classification process. The trained classification model generated is then utilized to predict class labels or categories for new, previously unseen time series data. RPCD addresses the challenge of high-dimensional time series classification by effectively reducing data dimensionality while maintaining pertinent information. This approach is particularly advantageous when dealing with extensive and intricate time series datasets, as it aims to enhance the efficiency and effectiveness of classification tasks within this context.

7. **SAX VSM** [66]: The SAX VSM (Symbolic Aggregate approXimation Vector Space Model) is an innovative approach to time series classification and similarity computation, amalgamating SAX (Symbolic Aggregate approXimation) and VSM (Vector Space Model). The VSM, widely utilized in text retrieval and information retrieval, serves to represent words or features within textual documents. It transforms a text document into a vector, where each dimension corresponds to a word or feature, and the vector's value signifies the importance of the word or feature in the document. The SAX VSM algorithm seamlessly integrates SAX and VSM by initially converting time series data into a sequence of symbols using the SAX technique. Subsequently, it represents each time series as a vector, where each dimension correlates with a symbol. The significance of each symbol is determined by its frequency in the time series. Ultimately, these vectors can be employed for similarity computation, clustering, or classification of time series data. The algorithm's primary strength lies in its ability to represent time series data as vectors, enabling the application of vector space modeling techniques for similarity calculations and classifications. This makes it a valuable tool in the realm of time series analysis and similarity computation, addressing the intricacies associated with time series data.
8. **TSBF** [67]: TSBF, an acronym for "Time Series Bag-of-Features", is a method designed to represent time series data through a set of features for classification purposes. The essence of TSBF lies in the extraction of pertinent features from time series data, capturing diverse characteristics and patterns such as statistical

measures, frequency domain features, and other relevant attributes. Following feature extraction, TSBF transforms each time series into a feature "package." Unlike conventional approaches, TSBF disregards the order of features within a time series, focusing solely on their presence and values to represent the data. This feature packet representation serves as input for classification algorithms like k-Nearest Neighbors (k-NN), Support Vector Machines (SVM), Decision Trees, or other machine learning models. The primary application of TSBF revolves around classifying time series data, assigning it to predefined classes or categories based on its feature representation. This approach simplifies the classification task by condensing time series information into a feature package while retaining critical details from the original data.

9. **SMTS** [68]: SMTS (Shape-based Multi-instance Time Series Classification) is a method used for multi-instance time series classification, primarily relying on shape features. This approach utilizes the shape features inherent in time series data to effectively represent multi-instance time series and subsequently classify them. In certain applications, time series data may exhibit a multi-instance structure, where each time series instance comprises multiple subsequences, each representing a fragment of the overall time series. The objective in multi-instance time series classification is to determine the class of the entire time series, considering the entirety of the individual subsequences. The SMTS method tackles this challenge by leveraging shape features of the time series data to represent each multi-instance time series. Shape features, in this context, encompass characteristics such as the curve shape, profile information, and key attributes of the time series, including peaks, valleys, and slopes. These features play a crucial role in capturing the essential traits of the time series, facilitating the discrimination between different classes of time series. Utilizing the extracted shape features, the SMTS method employs classification algorithms like k-Nearest Neighbors (k-NN), Support Vector Machines (SVM), Decision Trees, among others, to classify multi-instance time series. The focus of the SMTS method is centered on utilizing the shape features of the time series to accomplish the classification task effectively.

Through meticulous examination, the study explores and understands the distinct contributions of each algorithm, shedding light on their effectiveness in accurately classifying diverse emotional states expressed through speech. The summary of these nine algorithms is shown in the Table 1.2. These investigations are pivotal in advancing human comprehension of the nuanced landscape of speech emotion recognition. As research advances, classic algorithms combine time series with images and use images to classify time series. Table 1.3 details several algorithms related to images or shape.

Advancements in research are pushing the boundaries of speech emotion recognition (SER), with a notable trend focusing on the transformation of one-dimensional speech

Table 1.2. Summary of Time Series Classification Algorithms

| Algorithm | Definition | Basic Method | Application Areas |
|---|---|---|---|
| 1NN-RAW | 1NN-RAW is a nearest neighbor classification method using raw time series data. | Find the nearest neighbor based on raw time series data. | Time series classification and similarity matching. |
| 1NN-DTW-BWW | 1NN-DTW-BWW combines nearest neighbor classification with Dynamic Time Warping (DTW) and "BWW" method. | Combines DTW and "BWW" method for nearest neighbor classification. | Time series classification and similarity matching. |
| 1NN-DTW-nWW | 1NN-DTW-nWW is similar to the previous method, combining nearest neighbor classification with DTW and "nWW" method. | Combines DTW and "nWW" method for nearest neighbor classification. | Time series classification and similarity matching. |
| Fast Shapelet | Fast Shapelet is a time series classification method designed to rapidly discover and utilize shape features in time series for classification. | Utilizes shape features to represent and classify time series. | Time series classification, anomaly detection, etc. |
| SAX BoP | Symbolic Aggregate Approximation Bag-of-Patterns is a time series classification method that represents time series as symbolic sequences for classification. | Converts time series into symbolic sequences and uses a "Bag of Patterns" approach for classification. | Time series classification and pattern recognition. |
| SAX VSM | Symbolic Aggregate Approximation Vector Space Model is a time series classification method that represents time series as vectors in a vector space for classification. | Transforms time series into vector representations and employs a vector space model for classification. | Time series classification and pattern recognition. |
| RPCD | RPCD is a time series classification method that uses random projection to reduce the dimensionality of time series data for classification. | Utilizes random projection to reduce dimensionality, followed by classification. | Time series classification and dimensionality reduction. |
| SMTS | SMTS is a multi-instance time series classification method based on shape features. | Uses shape features to represent multi-instance time series and performs classification. | Multi-instance time series classification. |
| TSBF | TSBF is a method to represent time series as a set of features for classification. | Transforms time series into a set of features and uses them for classification. | Time series classification and feature extraction. |

Table 1.3. Algorithms related to image or shape in time series classification.

| **Algorithm** | **Basic Method** |
|---|---|
| Fast Shapelet | Focus on efficient discovery of shape features in time series |
| SAX BoP | Convert time series to symbolic sequences, capturing overall shape patterns |
| SAX VSM | Utilize symbolic sequences within a vector space model, emphasizing overall shape characteristics |
| RPCD | Implement random projection to reduce dimensionality, potentially preserving shape-related information |

signals into two-dimensional images. This innovative approach aims to leverage comprehensive image processing technology for the classification of speech signals, thereby enhancing the accuracy and effectiveness of speech emotion recognition. By visualizing speech signals in this manner, researchers seek to capture nuanced emotional cues embedded in the speech patterns, enabling a more intricate analysis of various emotions expressed during verbal communication. Wang et al. [69] introduced the innovative use of the Gram Angle Field (GAF), a novel approach for transforming one-dimensional speech signals into two-dimensional images. This pioneering method provides a unique perspective for visualizing and analyzing speech patterns. Similarly, Bakhshi et al. [70] proposed an alternative technique by leveraging the periodicity inherent in speech signals to convert one-dimensional speech signals into two-dimensional images (CyTex). These advancements in signal processing contribute to the expansion of techniques for representing and interpreting speech data, offering new dimensions for research and applications in the field of speech signal analysis. This transformative shift from traditional one-dimensional analysis to image-based processing showcases the evolving landscape of SER, promising more sophisticated and nuanced insights into the emotional nuances of spoken language.

## 1.3 Challenges and difficulties

Within the current research framework, challenges and difficulties are intricately linked to the employed methods. We have delineated the prevalent issues and obstacles in research from three primary perspectives, namely: Challenges and difficulties of nine time series classification algorithms, Challenges and difficulties of time series imaging algorithms, Challenges and difficulties of datasets.

### 1.3.1 Challenges and difficulties of nine time series classification algorithms

Within the existing research framework, there are nine time series classification algorithms, each encountering specific challenges and difficulties, as shown in Table 1.4. Firstly, the

1NN-RAW algorithm may face issues of computational complexity and storage requirements when dealing with large-scale datasets. Secondly, the 1NN-DTW-BWW algorithm, with its dynamic time warping (DTW), might be impacted by differences in sequence lengths, leading to increased computational costs. While the 1NN-DTW-nWW algorithm introduces normalization strategies to address length disparities, selecting the appropriate normalization method remains a challenge for certain datasets. The Fast Shapelet algorithm's efficiency might diminish when dealing with a substantial number of sequences due to the exhaustive search for potential shapes. SAX BoP algorithm requires parameter tuning to adapt to different datasets, which may necessitate domain-specific knowledge. SAX VSM algorithm needs careful selection of parameters when constructing the vector space model, as this choice can significantly impact the algorithm's performance. RPCD algorithm, being based on random projection, could be influenced by the selection of the projection matrix. Operating within a multi-instance learning framework, the SMTS algorithm needs to handle relationships among multiple instances, presenting a complex challenge. Finally, the TSBF algorithm requires the selection of an appropriate feature extraction method based on the dataset's characteristics, which might depend on the specific domain features.

Addressing these challenges and difficulties is crucial for advancing further research and enhancing the performance of these algorithms in various contexts.

### 1.3.2 Challenges and diffculties of time series imaging algorithms

Wang et al. and Bakhshi et al. proposed two methods present unique challenges and complexities in the field ofSER. For Wang et al.'s approach, utilizing the Gram angle field for converting one-dimensional speech signals into two-dimensional images introduces challenges in terms of the accuracy and robustness of the transformation. The effectiveness of this method depends on how well the Gram angle field captures the essential features of emotional expression in speech. Ensuring that the two-dimensional representation preserves the relevant emotional cues while minimizing information loss is a critical challenge. On the other hand, Bakhshi et al.'s method, based on the periodicity of the speech signal, faces challenges related to the generalizability of the conversion technique. The periodicity of the speech signal may not universally capture the diverse patterns present in emotional speech across different individuals and cultural contexts. Adapting this approach to accommodate variations in speech characteristics and emotional expressions poses a significant challenge.

Both methods need to address issues of standardization and calibration to ensure that the transformed images effectively convey emotional information across various speakers and scenarios. Additionally, the interpretability and explainability of the generated two-dimensional representations remain important challenges in both approaches. Balancing the need for complexity in capturing emotional nuances with the requirement for simplic-

Table 1.4. Challenges in time series classification algorithms.

| **Algorithm** | **Challenges** |
|---|---|
| 1NN-RAW | Computational complexity and storage requirements in handling large-scale datasets. |
| 1NN-DTW-BWW | Influence of sequence length differences on dynamic time warping (DTW), leading to increased computational costs. |
| 1NN-DTW-nWW | Introducing normalization to address length differences, but selecting appropriate normalization methods remains a challenge for certain datasets. |
| Fast Shapelet | Efficiency may decrease with a large number of sequences due to the need to search all possible shapes. |
| SAX BoP | Requires parameter tuning for different datasets, which may demand domain-specific knowledge. |
| SAX VSM | Building a vector space model requires selecting suitable parameters, impacting algorithm performance. |
| RPCD | Performance may be influenced by the choice of the random projection matrix. |
| SMTS | Operates in a multi-instance learning context, requiring handling relationships between multiple instances—a complex problem. |
| TSBF | Selection of appropriate feature extraction methods depending on dataset characteristics, relying on domain-specific knowledge. |

ity and generalizability is a key difficulty in advancing these methods for practical SER applications.

### 1.3.3 Challenges and diffculties of dataset

The study encountered challenges related to the availability and adequacy of suitable databases. It proved difficult to identify a database that perfectly aligned with the specific requirements of the research. Additionally, the scarcity of data within the identified databases presented a limitation. These challenges underscore the broader issue of the need for comprehensive and diverse databases in the field, reflecting the variety of scenarios and domains where time series classification algorithms might be applied. Addressing these challenges can contribute to the robustness and applicability of the research findings.

## 1.4 Research Content

Based on the above difficulties and challenges, this study proposes an innovative method to convert one-dimensional time series into two-dimensional images. Specifically, the method involves mapping the speech signal onto an image following the trajectory of the Hilbert curve, utilizing the arrangement of the Hilbert curve path. Through a convolution operation, the feature values of the resulting image are extracted. These eigenvalues are then flattened based on the Hilbert curve arrangement method, yielding a one-dimensional vector. Subsequently, a classifier is employed to categorize the vector, effectively achieving the objective of speech emotion recognition. This methodology capitalizes on the structural properties of the Hilbert curve, providing a unique and efficient means to represent and analyze speech signals for emotion recognition purposes.

Methods using Hilbert curves have some advantages when dealing with time series classification problems and can solve or avoid some of the challenges and difficulties mentioned previously.

1. **Challenges and difficulties of nine time series classification algorithms**: First of all, the path of the Hilbert curve is a compact, continuous and orderly curve, which can map one-dimensional time series data to a two-dimensional image. By arranging the time series data along the path of the Hilbert curve, we can preserve the temporal nature of the sequence on the image while reducing the dimensionality. This mapping to images allows local features of the data to be preserved, helping to better capture patterns in sequences. Second, by performing convolution operations on the Hilbert curve path, we can extract the feature values of the image. This feature extraction method is local and can focus on important local structures in the sequence, thereby better reflecting the characteristics of the sequence. This process reduces computational complexity, especially relative to some algorithms that require searching all possible shapes, such as Fast Shapelet. Additionally, the process of flattening an image into a one-dimensional vector and passing it through a classifier is relatively efficient. This makes the Hilbert curve method more easily adaptable to large-scale data sets, reduces computing and storage requirements, and thus avoids the problems that the 1NN-RAW algorithm may face. In summary, when dealing with time series classification problems, the Hilbert curve method is expected to solve or alleviate the challenges faced by some common algorithms by effectively retaining the temporal nature of the sequence and extracting local features.
2. **Challenges and diffculties of time series imaging algorithms**: The Hilbert curve method effectively addresses challenges associated with standardization, calibration, interpretability, and generalizability within the context of emotion recogni-

tion from speech signals. By providing a structured trajectory for mapping speech signals to images, the Hilbert curve method ensures a standardized and calibrated approach, promoting consistent conveyance of emotional information across diverse speakers and scenarios. Moreover, the inherent properties of the curve contribute to enhanced interpretability, facilitating the understanding and explanation of emotional features in the generated two-dimensional representations. Additionally, the method strikes a balance between capturing intricate emotional nuances and maintaining simplicity and generalizability, making it well-suited for practical SER applications. In summary, the Hilbert curve method offers a comprehensive solution to the challenges inherent in emotion recognition methodologies, making it a promising approach in the field of SER.

This study also expanded the database through simple transformations. The database is augmented by simple transformations, a step that is crucial to improve the performance and robustness of the algorithm. The expanded database can better capture the diversity of time series and make the algorithm more generalizable. Simple transformations may include operations such as translation, rotation, or scaling. Through these transformations, the original data gets some variations, which helps the training algorithm to better adapt to various situations. Database expansion also helps solve the problem of data imbalance and ensures that the algorithm can be fully trained on samples of each category. This is critical to maintaining the fairness and reliability of the algorithm, especially when faced with imbalanced data sets in the real world. Overall, augmenting the database with simple transformations is a key step in research that helps improve the algorithm's performance, generalization ability, and adaptability to imbalanced data.

## 1.5 Article structure

This study consists of six chapters.

**Chapter 1: Introduction**

In the introduction, the study unfolds against the backdrop of the current landscape, delving into the intricacies of the field and the specific challenges it faces. It articulates the primary goals and objectives of the research, shedding light on the proposed methodology's nuances and underscoring its distinct advantages. A structural overview of the entire paper is provided, serving as a roadmap for readers to navigate the comprehensive exploration of speech emotion recognition.

**Chapter 2: Proposal 1: Speech Emotion Recognition Based on Gramian Angular Field**

Here, a nuanced optimization proposal for the Gramian Angular Field (GAF) method takes center stage. The optimization is bifurcated into refining GAF's dimensionality reduction process, pitting it against alternative sampling methods, and enhancing the

GAF network model through a groundbreaking feature extraction approach rooted in the fundamental structure of GAF images. The chapter meticulously substantiates the efficacy of this approach through a thorough examination of experimental results.

**Chapter 3: Proposal 2: Speech Emotion Recognition Based on CyTex and Speech Rate**

In this chapter, the focus shifts to Proposal 2, introducing a mutli-feature value - speech rate – amalgamating speech rhythm and short-time features. The methodological intricacies of feature extraction, particularly the segmentation of speech emotions using these composite features, are laid bare. The synergy between the CyTex approach and an LSTM network model is dissected, elucidating the steps taken for effective feature extraction and subsequent speech emotion recognition. Experimental outcomes conclusively demonstrate the efficiency of this innovative method.

**Chapter 4: Proposal 3: Speech Emotion Recognition Based on Hilbert Curve**

This chapter intricately unpacks the novel approach of leveraging the Hilbert curve to transmute one-dimensional speech data into two-dimensional images. The methodology's rationale, experimental datasets, and the meticulous steps of the experimental process are explicated. Rigorous attention is devoted to showcasing the tangible outcomes of the proposed technique through compelling experimental results, thus validating its efficacy in the realm of speech emotion recognition.

**Chapter 5: Applications of the Proposed Two-Dimensionalization Algorithm in EEG Field**

Expanding the horizon of the proposed methodologies, this chapter traverses into uncharted territory by applying the CyTex method to transform brainwave signals into two-dimensional images for epilepsy recognition. The chapter serves as a testament to the adaptability and robustness of the proposed techniques across diverse domains. Detailed insights into the transformation method and the commendable results achieved with RNN and LSTM network models underscore the versatility and real-world applicability of the experimental proposals.

**Chapter 6: Summary and Discussion**

The concluding chapter encapsulates the essence of the entire research journey, offering a reflective summary of accomplishments and milestones achieved. The discussion extends beyond the immediate findings, delving into the broader implications of the research. Future trajectories and potential avenues for further exploration are outlined, providing a forward-looking perspective that adds a layer of depth and anticipation to the study's culmination.

# Chapter 2

# Proposal 1: Speech Emotion Recognition Based on Gramian Angular Field

## 2.1 Introduction

In recent years, the interest in speech signals emotion recognition is increasing. The increase in social pressure causes people to experience psychological stress [71]. The accumulates mental stress impacts people's mental health and generates diseases [72]. These illnesses can affect the diseases and their families [73, 74]. Furthermore, threatens social stability [75, 76, 77]. However, despit the many negative consequences of mental illness, most people who need mental health services cannot receive treatment for various reasons [77]. Researchers have confirmed speech can convey the psychological state of the interlocutor [78, 79, 80]. Therefore, researchers propose research on speech emotion recognition [81, 82]. Researchers confirm the application of speech emotion recognition in human-computer interaction (HCI). HCI system identifies the user's emotions and provides comfort and encouragement. Likewise, it provides psychological counseling and therapy [77, 83, 84].

The recognition of speech emotion comprises two main steps: feature extraction and classifier construction [85]. The input speech signal is highly time-dependent and continuous. For this reason, the traditional speech feature extraction method is Short Time Fourier Transform (STFT). This method obtains the speech spectral features [86, 87]. The spectral features contain plenty and complex information. Such as pitch, speech rate, rhythm, and energy [88]. Previous research confirms the single features are not universal [89], mixes features are leading over-fitting [90]. Therefore, the researcher suggests using sentiment labels as supervision for machine learning. Using machine learning trains models and extracts features. This proposal simplifies the process of manual feature ex-

traction. Traditional machine learning methods can build classifiers based on artificial speech features or deep speech features, such as classical methods like Gaussian Mixture Models (GMM), Hidden Markov Models (HMM), and Support Vector Machines (SVM) [91]. Drawing upon successful experiences in computer vision [69], suggests a recognition method. This proposal converts speech signals into image signals. This method preserves the temporality of speech [70] and uses computer vision techniques identifies speech emotions [69].

In this study, a new method is proposed on this basis, which aims to convert one-dimensional time series data into the form of two-dimensional images containing more information. The algorithm is inspired by converting one-dimensional time-series data into a two-dimensional image with the structure and shape of a Hilbert curve. Subsequently, a convolutional neural network is utilized to extract the features of the image, and the extracted feature values are passed through a fully connected network for sentiment classification. It is worth mentioning that the study compared the performance of this method with that of previous studies ([69] and [70]). This comparison validated the feasibility and effectiveness of the proposed algorithm in the field of emotion recognition. The results of this study show that the proposed algorithm has potential advantages in emotion-recognition tasks. The innovative aspect of this study is that it not only provides a new method for processing one-dimensional time-series data but also demonstrates the potential application of this method in emotion recognition. By presenting time series data as two-dimensional images, open up more possibilities in the field of sentiment analysis, with promising improved accuracy and performance. The success of this approach demonstrates the potential value of deep learning and computer vision for sentiment recognition.

## 2.2 Gramian Angular Field method

Wang et al. [92] introduced a novel methodology for converting one-dimensional speech signals into two-dimensional images, leveraging Gram Angle Fields (GAF). The GAF-based approach offers a unique perspective on visualizing and analyzing speech patterns, providing valuable insights into emotional expression.

The GAF algorithm is a time–series data coding method that extends the framework proposed by Campanharo et al. [93] and aims to preserve time-domain information [92]. For a time series $X$ containing real–valued observations $(x_1, x_2, ..., x_n)$, the algorithm uses the rescaling method to normalize the time series $X$ to the interval [0,1] or [-1,1]. Subsequently, the algorithm uses polar coordinates to represent the rescaled time series $\widetilde{X}$, mapping values to cosine angles and timestamps to radii. This innovative representation provides a new perspective for understanding time series. As time progressed, the values between different angular points on the polar coordinates changed. Unlike the traditional Cartesian coordinate system, the polar coordinates retain absolute temporal relationships.

Different rescaled data points correspond to different angular boundaries. For example, [0, 1] corresponds to the cosine function of [0, $\pi/2$], whereas the cosine value of [-1, 1] lies on the [0, $\pi$] angular boundary.

After converting the time series to polar coordinates, the algorithm can easily recognize the temporal correlation of different time intervals and calculate the triangular sum/difference between points using Eq (2.1). One to generate a two–dimensional image from a one-dimensional time series, as shown in Fig. 2.1. The generalized autocorrelation matrix (GAF) of the GAF algorithm has several advantages: firstly, it preserves temporal dependence; secondly, it preserves temporal correlation; and finally, it provides advanced time-series features for deep neural network learning. This approach has a wide range of applications in time series analyses.

$$G(x_1, \cdots, x_n) = \begin{bmatrix} \langle x_1, x_1 \rangle & \langle x_1, x_2 \rangle & \cdots & \langle x_1, x_n \rangle \\ \langle x_2, x_1 \rangle & \langle x_2, x_2 \rangle & \cdots & \langle x_2, x_n \rangle \\ \vdots & \vdots & \ddots & \vdots \\ \langle x_n, x_1 \rangle & \langle x_n, x_2 \rangle & \cdots & \langle x_n, x_n \rangle \end{bmatrix} \tag{2.1}$$

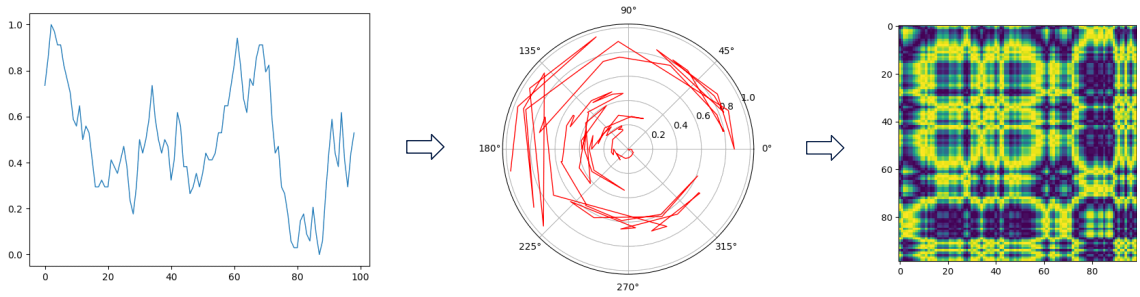



Fig. 2.1. Schematic diagram of the GAF algorithm. The time series are converted from a right-angle coordinate system to a polar coordinate system by transformation, and the GAF image is then generated by Eq (2.1).

Despite its innovation, Wang et al.'s approach faces challenges in emotion recognition. The accuracy and robustness of the conversion process, particularly the dependence on GAF for capturing essential emotional features, present hurdles. The challenge lies in ensuring that the two-dimensional representation retains relevant emotional cues while minimizing information loss.

This chapter enhances the GAF method through a dual approach. Firstly, in terms of data sampling, the research incorporates diverse downsampling algorithms to effectively downsample the speech signal. This methodology not only ensures the preservation of data integrity but also contributes to a more robust representation. Secondly, concerning network design, a novel model is introduced that aligns more closely with the distribution of GAF image features. This strategic enhancement leads to improved accuracy in feature extraction, thereby enhancing the overall performance of the system.

## 2.3 Optimization methods

The original GAF (Gramian Angular Field) algorithm used the PAA algorithm for downsampling the data in order to convert the time series data into an image representation. However, this study explores and improves on this further. The study introduces new downsampling methods that go beyond the original ppa algorithm and improve the accuracy and efficiency of the algorithm. In this study, these new downsampling methods were shown to better capture important features in the time series while reducing the loss of data information. Compared to the original ppa algorithm, the new methods are able to retain meaningful data details more accurately, providing richer inputs for subsequent neural network training. Based on this optimized data representation, study furthers improve the neural network model. The study propose a novel diagonal matrix-based network structure that aims to better utilize the data features provided by the GAF algorithm. This diagonal matrix-based network model is able to learn and utilize the intrinsic structure of the time series data more efficiently, allowing the neural network to capture the data features more accurately during the learning process.

In the next part, study discuss the principles and implementation details of the new downsampling method and the diagonal matrix network model in detail. Study presents the experimental design and results to show how these improvements have improved the performance of the algorithm. At the same time, study conduct a comparative analysis to verify the superiority of the new method over the traditional method, as well as further explore the potential impact of these improvements and future application directions.

### 2.3.1 Introduction of downsampling methods

Time series data are usually characterized by high dimensionality and contain numerous time points. Therefore, dimensionality reduction is performed when necessary. The purpose of this step is to reduce computational complexity and eliminate redundant information from the data. Commonly used dimensionality reduction algorithms include:

1. **Piecewise Aggregate Approximation Algorithm**: In 2000, Eamonn Keogh et al. [94] proposed the Piecewise Aggregate Approximation Algorithm. The length of time series $X = \{x_1, x_2, x_3, \cdots, x_n\}$ is $n$. Equally divided $X$ into $m$ part, and each part express as length $N$ vector $\overline{X}$. The $ith$ element of $\overline{X}$ expresses as $\overline{X} = \frac{N}{n} \sum_{j=\frac{n}{N}(i-1)+1}^{\frac{n}{N}i} x_j$.
2. **Ramer Douglas Peucker Algorithm**: In 1973, D. Douglas and T. Peueker proposed the Ramer Douglas Peucker Algorithm(RDP). RDP is a classic algorithm for linear feature extraction. While conducting retain the geometric shape, the amount of data is simultaneously reducing [95]. As illustrated in Fig. 2.2, the schematic

diagram represents the underlying principles of the Douglas Peukcer algorithm.

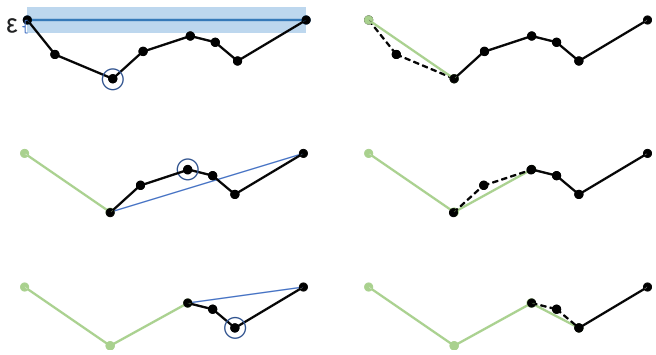



Fig. 2.2. The underlying principles of the Douglas Peukcer algorithm. Connects the start($S$) and end($E$) point of the sequence, obtains the line $SE$. Calculates the vertical distance $d_i$ from each point in the sequence to the line $SE$. Compares the distance $d_i$ with the threshold value$\epsilon$. If $d_i < \epsilon$, delete the point $d_i$, connect $S$ and $d_{(i+1)}$. $d_{(i+1)}$ replaces the $d_i$ become the new $d_i$; If $d_i > \epsilon$, keep the point $d_i$, take $d_i$ as start connects with $E$. Repeat the previous steps.

3. **The Visvalingam-Whyatt Algorithm**: Visvalingam-Whyatt(VW) Algorithm, alternativelyreferred to as Visvalingam Algorithm. The algorithm extracts thekey points to generate curves with similar shapes [96]. Resemblingthe RDP Algorithm, VW Algorithm employs a threshold value ofarea to execute downsampling. This algorithm iteratively removes vertices froma line based on their significance or importance. The operation principle is illustrated as shown in the Fig. 2.3. Utilizes this algorithm can delete over 95% of the data points whileretaining enough feature points. This algorithm's parameters define indifferent ways. It can use the percentage of points, the number of points, or thearea threshold to keep the downsampling [97]. This research uses thenumber of points and an area threshold for two functions.
4. **The Largest Triangle Three Bucket Algorithm**: The Largest Triangle Three Bucket Algorithm (LTTB) combines WhyttAlgorithm and Intuitionistic Algorithm [98]. It combines the bucketingidea in the intuitional algorithm, quantifies the importance of the data. Reduces the number of data while preserves the essential shape and featuresof the data. The basic principle of the LTTB algorithm as illustrated in Fig. 2.4.

The advantage and disadvantage of these methods is shown in Table 2.1.

### 2.3.2 Novel Neural Network Modeling

The GAF matrix is symmetric about the main diagonal [99, 100]. And the convolution disallows rotationalinvariance. Therefore, utilizes convolution causes repeated extraction data feature. In order to solve the above problem, this research proposes a method forex-

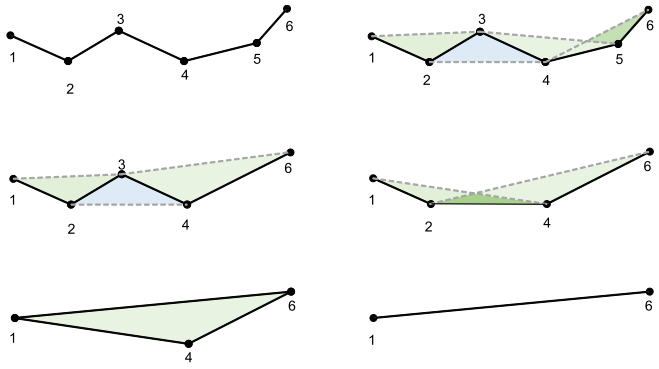



Fig. 2.3. The principles of the Visvalingam Whyatt Algorithm. The sequence $X = x_1, x_2, x_3, \cdots, x_n$ length is n. For any point $x_i$ except the start and end, composites atriangle with the two adjacent points $x_{(i-1)}$ and $x_{(i+1)}$. When thearea of the triangle is less than the threshold value, the point $x_i$ will bedeleted.

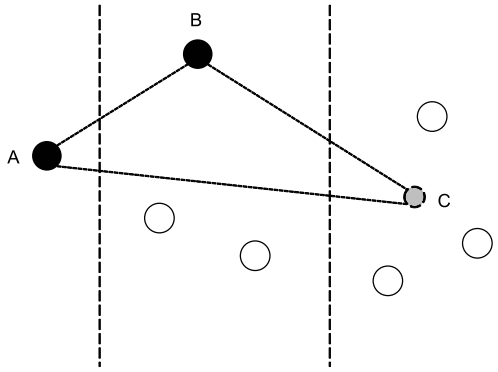



Fig. 2.4. The specific steps of the Largest Triangle Three Bucket Algorithm. Retains the first and the last points. Divides themiddle points into $n$ parts an equally in order. Each part with $m$ points. Point $A$ as the first point, and point $C$ as the temporary point. Point $C$ is the average of this part. Point $B$ is the point which makes the largesttriangle $ABC$ area. Iteration thesesteps, until the last point.

tracting the eigenvalues of GAF images. The symmetric matrix is a matrix which main diagonal is the axis of symmetry. Each of its elements corresponds to the same. For a real symmetric matrix $\boldsymbol{A}$ (the elements are all real numbers), there exists an orthogonal matrix $\boldsymbol{P}$ and a diagonal matrix $\boldsymbol{\Lambda}$ such that the following equation holds $\boldsymbol{P}^{-1}\boldsymbol{A}\boldsymbol{P} = \boldsymbol{\Lambda}$. Where $\boldsymbol{A}$ is the real symmetric matrix, $\boldsymbol{P}$ is the orthogonal matrix, $\boldsymbol{\Lambda}$ is the diagonal matrix, and $\boldsymbol{P}^{\boldsymbol{T}}$ denotes the transpose of $\boldsymbol{P}$. Consequently, this research proposes a Symmetric Diagonal Matrix (SDM) network structure. The structure of SDM as shown in Fig. 2.5. First, calculates the symmetric matrix obtains a diagonal matrix. Second, calculates the eigenvalue vector of the diagonal matrix. The eigenvalue vector serves as input data for the full connected network. Finally, utilizes the fully connected network for the classification of speech emotion.

Table 2.1. Methods for Time Series Downsampling: Advantages and Disadvantages.

| Method | Advantages | Disadvantages |
|---|---|---|
| Piecewise Aggregate Approximation Algorithm | • Preserves data integrity through equal division.<br>• Provides a straightforward expression for each part of the time series. | • May lose detailed information due to averaging. |
| Ramer Douglas Peucker Algorithm | • Retains geometric shape while reducing data.<br>• Classic algorithm for linear feature extraction. | • Complexity may increase for intricate shapes. |
| Visvalingam-Whyatt Algorithm | • Preserves key points for similar-shaped curves.<br>• Efficiently removes vertices based on significance. | • Requires setting appropriate threshold values.<br>• The algorithm's parameters may be defined in different ways. |
| Largest Triangle Three Bucket Algorithm | • Combines the benefits of Whytt and Intuitionistic Algorithms.<br>• Quantifies data importance through bucketing. | • Reduction in data points may impact certain features. |

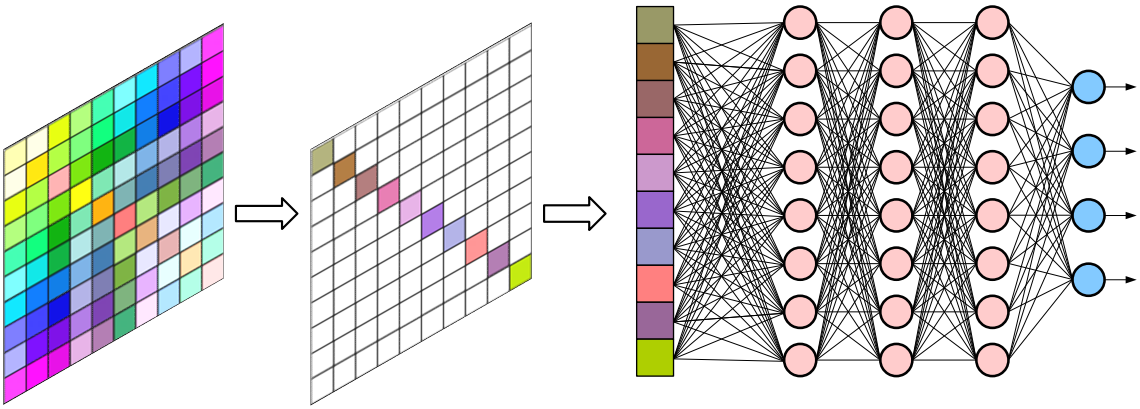


Fig. 2.5. The structure of symmetric diagonal matrix network.

## 2.4 Experimentation and methodology

In this section, the paper provides a comprehensive overview of the experimentation process and the methodology employed in the research. I detail the experimental setup, including the selection and preparation of datasets, the configuration of experimental parameters, and the execution of experiments. Additionally, I elucidate the underlying principles guiding the approach and outline the rationale behind the chosen methodologies. Through a systematic exploration of experimental design and methodology, the research aims to ensure robustness, reliability, and reproducibility in our research findings.

### 2.4.1 Dataset

This research uses the speech emotion signal corpus CASIA. This corpus is recorded by the Automation of the Chinese Academy of Sciences. According to the experimental requirements, the research pair randomly partitioned the database to form the data set shown in Fig. 2.6.

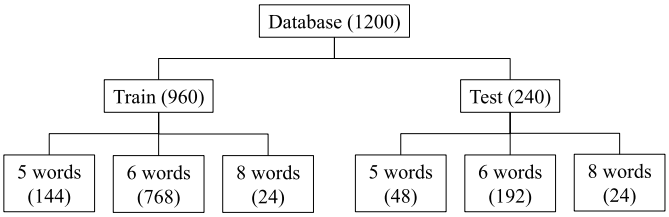



Fig. 2.6. The data set composition.

### 2.4.2 Experiment

The research set the parameters of the experiment as shown in Table. 2.2. The experiment mainly uses Adam, Adagrad, and SGD optimizer to tune the parameters of deep learning. In addition, other optimizers are also used. The advantages and disadvantages of the optimizers a shown in the table. 2.3. The illustration of experiment as shown in Fig. 2.7, In this experimental phase, the procedures are meticulously executed under predefined conditions. The process commences with the acquisition of speech data, subsequently transformed into waveform representations. To ensure accuracy, white spaces at the beginning and end of the data are eliminated using endpoint detection. Employing four distinct downsampling algorithms, namely Piecewise Aggregate Approximation (PAA), Ramer Douglas Peucker (RDP), Visvalingam-Whyatt (VW), and Largest Triangle Three Bucket (LTTB), the speech signals undergo varied preprocessing. The downsampling outcomes are then imaged using the Gramian Angular Field (GAF). Three emotion classification neural network models are introduced, each depicted with an illustrative

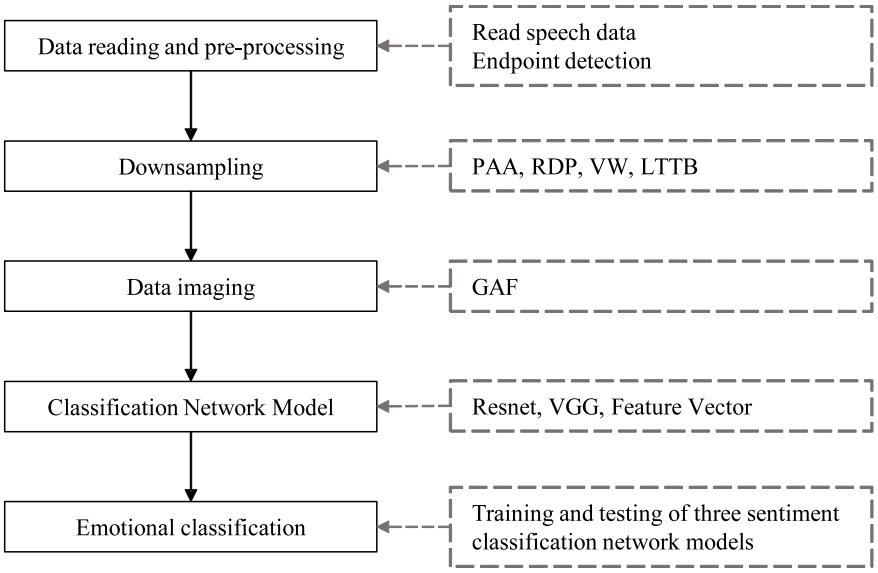



Fig. 2.7. Illustration of experiment. For data reading and preprocessing, reads the speech data and converts it into a waveform graph. Eliminate blank data at the beginning and end of the speech data using endpoint detection technology to ensure the accuracy and consistency of the data. The application of downsampling algorithms: four downsampling algorithms are used to process speech signals: the Piecewise Aggregate Approximation (PAA) algorithm, the Ramer Douglas Peucker (RDP) algorithm, the Visvalingam-Whyatt (VW) algorithm, and the Largest Triangle Three Bucket (LTTB) algorithm. With these algorithms, multiple sets of downsampled speech data were obtained, with the VW algorithm using two different evaluation methods, resulting in a total of five sets of data. The downsampled data were imaged according to Gramian Angular Field (GAF) to obtain the corresponding image data. Three network models were designed for sentiment classification. It includes input layer, hidden layer and output layer. Training and test the generated image data as input.

structure as sown in Fig 2.8. Additionally, a feature vector network model is designed to calculate feature vectors from the GAF-generated feature matrix. Finally, the emotion classification experiments are conducted, utilizing the image data as input for training and evaluating the three neural network models, with a meticulous record of classification accuracy and other performance metrics. This detailed experimental approach aims to comprehensively explore the impact of different downsampling algorithms on speech emotion classification and assess the effectiveness of combining feature extraction and neural network models in emotion classification tasks.

Table 2.2. The parameter information of the experiment

| Parameter | Experimental parameter setting |
|---|---|
| Dataset partitioning | train set 80%, test set 20%. |
| Optimizer | Adam, Adagrad, SGD, etc |
| Learning rate | 1e-5, 5e-5 |
| Loss fuction | NLLloss |
| Batch size | 12 |
| Epoch | 500 |

Table 2.3. The advantages and disadvantages of common optimizers.

| **Optimizer** | **advantage** | **Disadvantage** |
|---|---|---|
| SGD | Highly computationally efficient, more scalable for large data and complex models. | Less convergent and the update direction is unstable. |
| Adam | Combines the benefits of momentum optimization and adaptive learning rates. | Over-adjusted the learning rate. |
| AdaGrad | Adaptive learning rate, different parameters have different learning rates. | The learning rate may drop excessively, resulting in slower convergence. |
| RMSProp | Adaptive learning rate,performs smaller updates for parameters with larger gradients. | In some cases, learning rates may still decline too quickly. |
| Adamx | Adaptive learning rate, supports sparse gradient, fast convergence. | High memory consumption, sensitive to learning rate, difficult to adjust parameters. |
| AdamW | Better regularization, robustness, high adaptive learning rate. | High memory occupation and difficult hyperparameter adjustment. |
| Nadam | Fast convergence, adaptive learning rate, introduction of Nesterov momentum to accelerate convergence and improve model performance. | High memory consumption, high sensitivity to learning rate, difficult parameter tuning. |
| Radam | Adaptive learning rate, robustness, solving offset problem. | High difficulty of hyperparameter adjustment, large memory consumption. |

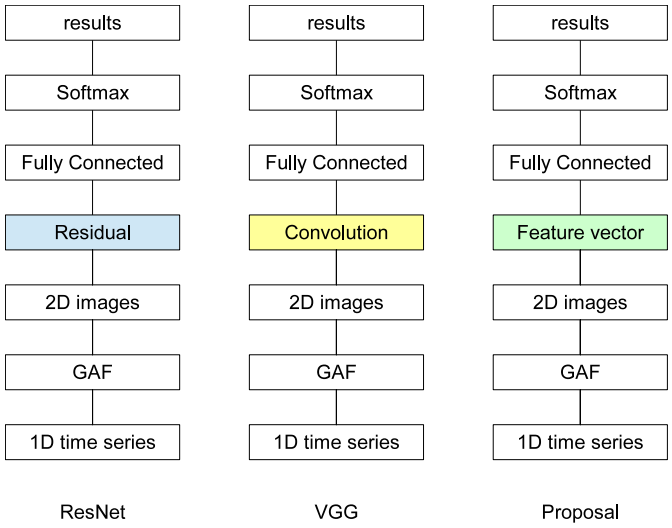



Fig. 2.8. Network structures.

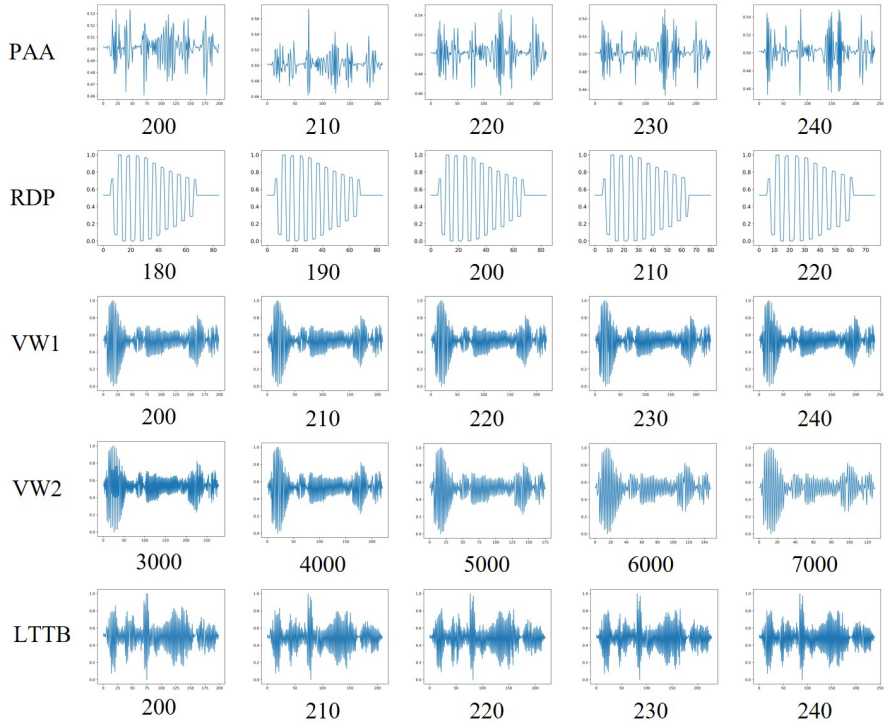
## 2.5 Results and discussion

The experiment performs experimental operations according to the set conditions and records the results. The experimental operations include reading the data and converting the speech data into a waveform graph. Four downsampling algorithms were used to process the speech signal. Since the VW algorithm uses two evaluation methods, the experiment obtained five sets of data. The experiment is based on the GAF algorithm to complete the time series imaging, as shown in Figure reffig:3-9. The final image data will be classified by three network models for sentiment classification. The study uses accuracy as the evaluation metric, and the results are shown in Table 2.4

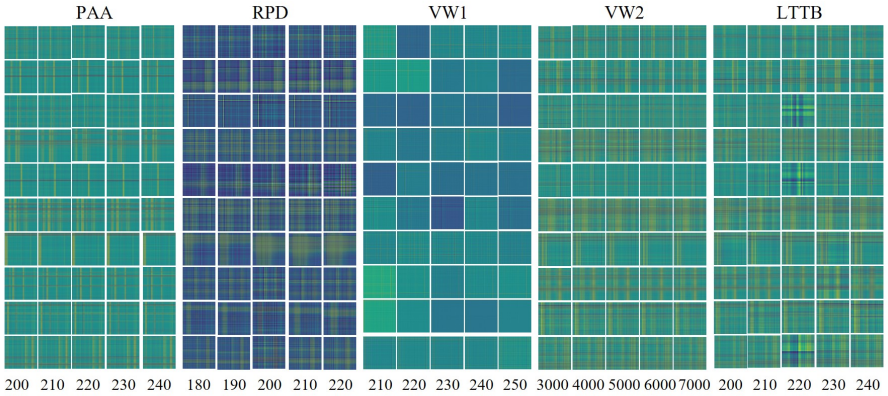
## 2.6 Conclusions

The main goal of this section is to explore how to improve the GAF methodology to provide a clearer presentation of the data and a more efficient means of analysis. In this process, we face challenges such as high memory requirements and computational complexity, and need to consider the efficiency of the algorithm and the accuracy of the data visualization. In order to better preserve the integrity of the data, we used a variety of downsampling methods to process the input data.

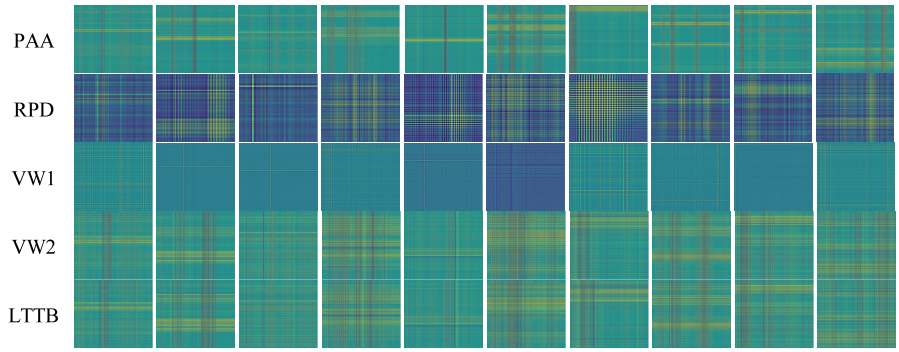
For speech emotion recognition, we chose to use SDM as the feature extractor and fully connected classifier. Compared with the traditional convolutional neural network, SDM outperforms the VGG network model for recognition. However, it should be noted that the SDM algorithm occupies more memory and takes more time compared to the other two models. This suggests that the contribution of the SDM model to speech emotion




(a) Comparison of waveforms after downsampling by various algorithms.




(b) Comparison of GAF images after downsampling with various algorithms.




(c) Comparison of GAF optimal parameter maps after downsampling with various algorithms.

Table 2.4. Accuracy on the three network.

| Network | Optimizer | lr | PAA | LTTB | VW_NUM | VW_TH | RDP |
|---|---|---|---|---|---|---|---|
| VGG16 | Adagrad | 1E-05 | 18.33 | 18.75 | 17.08 | 22.08 | 17.91 |
| | | 5E-05 | 18.75 | 17.50 | 20.41 | 28.33 | 00.00 |
| | Adam | 1E-05 | 16.67 | 16.67 | 16.67 | 16.67 | 16.67 |
| | | 5E-05 | 16.67 | 16.67 | 16.67 | 16.67 | 16.67 |
| | SGD | 1E-05 | - | - | 16.67 | - | 16.67 |
| | | 5E-05 | - | - | 16.67 | - | 16.67 |
| ResNet34 | Adagrad | 1E-05 | 24.16 | 26.67 | 28.75 | 29.16 | 26.66 |
| | | 5E-05 | 25.83 | 31.25 | 26.60 | 28.33 | 25.83 |
| | Adam | 1E-05 | 18.37 | 20.20 | 19.05 | 20.38 | 23.31 |
| | | 5E-05 | 26.38 | 21.25 | 25.06 | 23.35 | 25.03 |
| | SGD | 1E-05 | - | - | 16.67 | - | 16.67 |
| | | 5E-05 | - | - | 16.67 | - | 16.67 |
| ResNet50 | Adagrad | 1E-05 | 25.83 | 25.00 | 28.75 | 28.30 | 27.91 |
| SDM | Adagrad | 1E-05 | 21.33 | 24.75 | 28.08 | 27.08 | 21.91 |
| | | 5E-05 | 22.35 | 24.26 | 26.43 | 28.35 | 16.67 |
| | Adam | 1E-05 | - | - | - | 22.08 | - |
| | SGD | 1E-05 | - | - | - | 16.67 | - |
| | | 5E-05 | 16.67 | - | 16.67 | 16.67 | - |
| | Adamax | 1E-05 | - | - | - | 24.58 | - |
| | AdamW | 1E-05 | - | - | - | 21.25 | - |
| | Nadam | 1E-05 | - | - | - | 22.91 | - |
| | Radam | 1E-05 | - | - | - | 23.75 | - |
| | RMSprop | 1E-05 | - | - | - | 22.91 | - |

- Indicates the accuracy rate value is unsatisfactory, is not indicated.

recognition is significant.

Future research directions include further optimizing the parameters in the diagonal matrix model to improve the performance of the SDM model in speech emotion recognition. This will help to better understand and utilize the potential of the SDM model in practical applications.

# Chapter 3

# Proposal 2: Speech Emotion Recognition Based on CyTex and Speech Rate

## 3.1 Introduction

Speech is one of the most important ways for humans to communicate. It is a common carrier of information and emotions. Researchers have studied speech recognition extensively since the late 1950s. Speech recognition involves converting human speech into word sequences [101]. However, in this process, the machine cannot understand the emotional state of the speaker during human–computer interaction (HCI). Speech emotion recognition— a relatively new area of research—focuses on identifying the intention of speech delivery [102, 103]. It involves extracting the emotional state of the speaker from speech. Researchers have confirmed that speech emotion recognition can extract useful semantics from speech and improve the performance of speech recognition systems [104].

Speech emotion recognition has important applications in various fields. In HCI systems, it can improve the user experience. An HCI system identifies the emotional state of the user to provide personalized and emotional responses. For example, when the user discontentment reaches a predetermined level [105], the system transfers control to manual customer service. This provides better assistance and service to users [106]. Speech emotion analysis is widely used by businesses, governments, and other organizations [107]. To improve the quality of goods and services, institutions collect and analyze people's opinions and impressions of various topics, products, themes, and services. The objective is to enhance people's happiness and satisfaction [108].

Researchers have confirmed the associations between speech emotion and mental health and mental state. Emotion recognition of speech assists in the diagnosis of depression and suicide risk assessment. In addition, speech emotion recognition can be used to detect

and predict mental states. In 1921, Kraepelin [109] corroborated and described depressed patients' articulatory features and speech characteristics before suicide [110]. The results of [109, 110] revealed the ability of speech emotion to track the severity of depression objectively and its usefulness for evaluating the effects of treatment [111]. [112] discussed speech emotion features associated with autism. [113] focused on Alzheimer's disease detection using speech. [114] confirmed the effectiveness of speech for predicting mild cognitive impairment. Moreover, for car driving, monitoring the emotional state of the driver through speech recognition can prevent traffic accidents [115].

Feature extraction is an important step in speech emotion recognition. There are various types of speech features, such as rhythmic features (e.g., pitch, tone, and prosodic contour), qualitative features (e.g., resonant peak frequencies and spectral features), and derived features (e.g., Mel-frequency cepstral coefficients and linear predictive coding coefficients) [116]. Research on speech and emotional states has indicated that some features are interrelated; e.g., the information of fundamental frequency and speech quality can be extracted from the continuous acoustic variables [104]. Feature selection is one of the central issues in speech emotion recognition [117]. In previous research, speech emotion recognition has largely been based on the rhythmic features [118]. However, it is unclear which features have the most significant influence on emotion classification [119]. Such as, volume is higher for anger or excitement and lower for shyness or frustration; pitch is higher for excitement or nervousness and lower for frustration or sadness. Volume and pitch both affect excitement or frustration, but it's unclear which is more important. Therefore, single-parameter extraction is non-universal [120]. Therefore, researchers have explored feature sets with a mixture of features [120, 121]. Nevertheless, with the increasing number of parameters, the dimensionality of features is increasing. This makes the recognition of speech emotions more complex and leads to overfitting [122]. Classifiers are another important part of speech emotion recognition. For single-parameter classification, shallow classifiers, such as SVM classifiers [123], ELM classifiers [122] and BayesNet classifiers [124], can achieve satisfactory classification performance. Classifiers suitable for a mixture of features, include the Gaussian Mixture Model classifier [125], ANN classifier [126], and RNN classifier [127].

To address the aforementioned research gaps, we propose a feature value extraction method that combines multiple features of speech. It achieves the mixture of features, reduces the dimensionality of speech feature values, and avoids the overfitting phenomenon caused by complex computation. First, in accordance with previous research, the time–frequency features of speech are extracted. Then, the time–frequency features are combined with the phonetic features. We propose a new speed feature based on the voiced consonant, to reduce the number of features. Finally, recognition is performed using the selected speech emotion dataset. Through several experiments and comparisons, we verified the effectiveness of the proposed method with the Chinese speech database from

the Institute of Automation of the Chinese Academy of Sciences (CASIA) [128]. In the experiments, the SVMs, ELMs and decision trees are used as classifiers to evaluate the recognition accuracies for single and multiple features. The recognition accuracy for the multiple features is higher than that for the single features, and the same results are achieved in the less time. The results indicate that using multiple features can increase the recognition accuracy. Moreover, it solves the problem of time redundancy. In addition to this, LSTM [129] was used as a network model to complete learning and recognition. The results show that the multi-feature method has excellent recognition results, reaching an average accuracy of up to 98.15%.

The remainder of this chapter is organized as follows. Section 3.2 briefly describes the relevant algorithms and neural networks covered in this chapter. Section 3.3 outlines the methodology and provides details about the specific experimental procedures. Section 3.4 presents the results obtained from the experiments and engages in a discussion of these results. Finally, Section 3.5 concludes the study.

## 3.2 Speech rate

In this study, I propose a composite speech feature called the speech rate composite feature. This feature combines the short duration feature of speech and the rhythmic feature of speech. The rhythmic feature allows us to find how turbid sounds are represented in speech. The short-time feature, on the other hand, captures the location of turbid sounds in speech. The neighboring intervals between the positions of the turbid sounds reflect the rate of speech. Using this composite feature, we can segment the speech signal and perform emotion recognition.

### 3.2.1 Short-time feature

An important part of speech emotion recognition is the extraction of feature values. The speech signal is a non-stationary signal [130]; it contains rich time-series information, and the statistical properties change over time. However, speech signals are typically processed with a high sampling rate, which causes a flat change in the speech signal over a short period. This is referred to as short-time smoothness [131]. The speech short-time features are among the most responsive to the speaker's emotion [132]. In this section, we introduce three short-time features: the short-time energy, short-time zero-crossing rate, and short-time average amplitude difference (Amdf) [132]. These features reflect the emotion of speech.

1. **Short-time energy**: The short-time energy of speech is the energy change of the speech signal in short time frames. It indicates the amount of energy in each short time frame of the speech signal. In a speech signal, the stronger parts have a higher

short-time energy. The silent or noisy parts have a lower short-time energy. For a speech signal frame of length $N$, the short-time energy can be calculated using Equation (3.1):

$$E(n) = \sum_{i=0}^{N-1} x^2(n-i), \tag{3.1}$$

where $E(n)$ represents the short-time energy of the $n^{th}$ frame and x(n-i) is the $i^{th}$ sample point value of the speech signal. $\sum$ in the equation is the summation operation, which ranges from the start sample point to the end sample point in current frame—typically from 0 to $N-1$.

2. **Short-time zero-crossing rate**: The short-time zero-crossing rate is used to describe the rate at which the signal changes its sign within short-time frames. A higher zero-crossing rate indicates that the speech signal waveform changes faster. Specifically, for a speech signal frame of length $N$, the short-time zero-crossing rate is given by Equation (3.2):

$$Z(n) = \frac{1}{2N} \sum_{i=0}^{N-1} \left|\text{sign}[x(n-i) - x(n-i-1)]\right|, \tag{3.2}$$

Where $Z(n)$ represents the short-time zero-crossing rate in units of *crossing / frame*, and $N$ represents the number of samples within the frame. $x[n]$ denotes the $n^{th}$ sample within the frame. $sign()$ denotes the sign function, which returns 1 when $x[n]$ is positive, $-1$ when $x[n]$ is negative, and 0 when $x[n]$ is zero.

3. **Short-time average amplitude difference**: The Amdf is utilized to explain the properties of the changes of the speech signal. It reflects the changes in the amplitude in the speech signal in a short period. A larger average amplitude difference indicates that the speech signal has a larger amplitude variation. The length of the speech signal $x[n]$ is denoted as $N$. Accordingly, the Amdf is expressed by Equation (3.3):

$$Amdf(n) = \frac{1}{N} \sum_{i=0}^{N-1} \left|x(n-i) - x(n-i-1)\right|, \tag{3.3}$$

where *Amdf(n)* represents the Amdf value of the $n^{th}$ frame, $x(n-i)$ and $x(n-i-1)$ denote the sample values of adjacent points within the frame, and the $\sum$ is taken over $N$ samples. By calculating the absolute magnitude difference between adjacent samples within one frame, the magnitude variation of the signal within a localized time period is quantified.

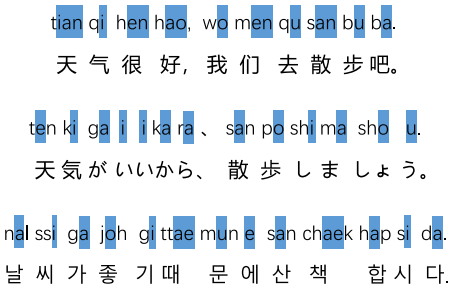



Fig. 3.1. Voiced sound in Chinese Pinyin, Japanese Kana, and Korean Pinyin.

### 3.2.2 Phonogram

Speech emotion is closely related to speech rate. Previous research has indicated that faster speech is typically associated with positive and active emotions. Conversely, slower speech corresponds to both negative and inactive emotions [133, 134]. Speech rate is usually expressed as the number of *syllables / duration*. There are various methods for counting the syllables, including counting based on the number of samples, counting based on the number of phonemes, and counting based on the number of frames. According to the concept of speech rate, this study proposes a multi-feature method that combines rhythmic and short-time features. This feature describes the phonogram speech rate.

The phonogram is a type of script in which speech is recorded using only a few letters. Common phonograms include the International Phonetic Alphabet [135], Chinese Pinyin [136], Japanese Kana [137], and Korean Hangul [138]. As shown in Fig. 3.1, in a phonogram, each syllable contains at least one vowel ( blue part of the diagram). Vowel pronunciation is accompanied by vibration of the vocal folds. This sound causes the vocal folds to vibrate when pronounced, which is called a voiced sound [135]. Fig. 3.2 shows the waveform of a sentence with six Chinese characters. In this image, the horizontal axis represents time in seconds (s). Each time point corresponded to the location of the waveform sampling point. The vertical axis represents the amplitude of the sound, which is dimensionless. This study divides the waveform into six parts by text and marks the vowel location by circles. The image below is the corresponding Chinese phonetic pinyin of the waveform. Voiced sound sections were labeled with squares.

Therefore, this study proposes a multi-feature method based on the phonogram. This approach defines speech rate features by calculating the time intervals between voiced sounds in the phonogram.

### 3.2.3 Speech rate

Previous research indicates that the vocalization of voiced sounds has excellent performance in short-time speech features. Vocal-fold vibration leads to speech signal energy

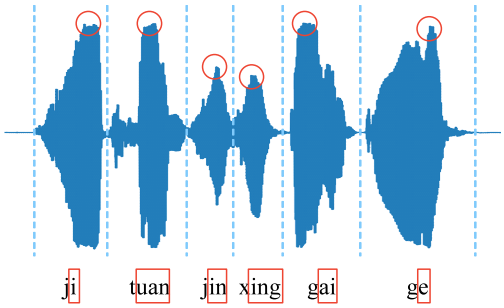



Fig. 3.2. Voiced sound for one sentence with six Chinese characters.

having a specific spectral distribution [139]. Compared with voiceless sound, voiced sound exhibits higher energy peaks and wider spectral bandwidths in short-time features [140]. Table 3.1 lists the common short-time features and the performance of the voiced sound [140, 138, 141, 142]. Voiced sounds exhibit unique characteristics within these short-term features, further demonstrating the potential for utilizing the identification of voiced sounds to define speech rate. The speech rate feature of the speech signal can be quantitatively obtained by extracting the frames where a voiced sound occurs within the short-time features and calculating the number of intervals between adjacent frames.

## 3.3 Experimentation and methodology

This section details the design of the experiment, execution process, and methods and techniques employed.

### 3.3.1 Dataset

The CASIA emotion corpus was used in this study. It consists of four professional speakers reciting 50 different sentences. Each sentence has one of six emotions: angry, happy, afraid, sad, surprised, and neutral. The entire dataset has a total of 1200 sentences. This corpus can be used to compare and analyze the acoustic and prosodic characteristics of different emotional states. It contains sentences of three lengths: 5 characters, 6 characters, and 8 characters. We randomly divided the corpus into a training set (80%) and a test set (20%), as shown in Table 3.2. The speech was recorded in a recording studio, with no external noise. The sampling frequency was 16000 kHz, and the storage format was PCM, 16-bit. The recording unit marked the emotion label.

### 3.3.2 Networks

1. **Support vector machine decision tree**: The SVM decision tree is based on the SVM algorithm [151]. It combines the advantages of the SVM algorithm and the decision-tree algorithm, improving the classification performance and interpretabil-

Table 3.1. Common short-time features and the performance for voiced sounds

| **Feature** | **Description** | **Performance for Voiced Sounds** |
|---|---|---|
| Short-Time Energy [143] | Overall signal energy in a short time frame | Higher energy due to vocal-fold vibrations |
| Short-Time Zero-Crossing Rate [144] | Rate at which the signal changes its sign in a short time frame | Voiced sound exhibits a lower zero-crossing rate |
| Short-Time Average Amplitude Difference [145] | Average difference between consecutive samples in a short time frame | Relatively small differences due to the smooth nature of the voiced sound |
| Short-Time Autocorrelation [146] | Indicates the periodicity of the signal | Clearly defined peaks indicating the periodic nature of the voiced sound |
| Short-Time Pitch [147] | Estimates the fundamental frequency of the signal and periodicity | Accurate pitch estimation reflecting the voiced sound |
| Short-Time Harmonic Ratio [148] | Indicates the ratio of harmonic to non-harmonic components | Higher ratio indicates the voiced sound |
| Short-Time Cepstral Coefficients [149] | Captures the resonant properties and spectral envelope of the signal | Voiced sound has a low-frequency component |
| Short-Time Peak Amplitude [150] | Represent the maximum amplitude of the signal | Peak amplitudes indicate the strengths and intensities of the voiced sound |

Table 3.2. Composition of the corpus.

| Number of words | 5 words | 6 words | 8 words | Total |
|---|---|---|---|---|
| Sentences numbers | 192 | 960 | 48 | 1200 |
| Training set numbers | 144 | 768 | 24 | 960 |
| Test set number s | 48 | 192 | 24 | 240 |

ity. The SVM algorithm is used for each leaf node subset, and the optimal partitioning is used to generate a tree structure. The SVM algorithm computes an optimal hyperplane at each node. The hyperplane segments the data into two child nodes. Using this approach, the next level of nodes is segmented. When the node data cannot be further partitioned, the node becomes a leaf node. Thus, the decision tree is constructed.

2. **Extreme learning machine decision tree**: The ELM decision tree is a hybrid machine learning algorithm that combines the ELM algorithm with a decision-tree structure for classification or regression tasks. The ELM model is used for each decision-tree node to make predictions and decisions [122].
3. **LSTM**: Recurrent Neural Networks (RNN) are artificial neural networks specialized for processing sequential data [152]. Unlike traditional neural networks, RNN excel at utilizing the internal memory states when dealing with sequential data. Their standout feature lies in their internal recursive structure, which allows the continuous looping of information within the network. This internal memory mechanism empowers the RNN to use prior information to influence output.

   Long short-term memory networks (LSTM) are specialized structures within the realm of RNN that are designed for processing and learning time-series data [153]. Unlike standard RNN, LSTM primarily focuses on overcoming common issues, such as gradient vanishing and explosion, particularly in managing lengthy sequential data. LSTM has gained widespread adoption because of its internal architecture, which allows the network to capture and retain long-term dependencies, which are crucial for tasks that require sustained memory. Information transfer within LSTM occurs through units equipped with gating mechanisms, which control the information flow by managing forgetting and adding memory, thus enabling more adaptable long-term memory storage.

   Structurally, LSTM differs from an RNN, as shown in Fig. 3.3. RNN consist of simple input, forget, and output gates that directly transfer information. By contrast,

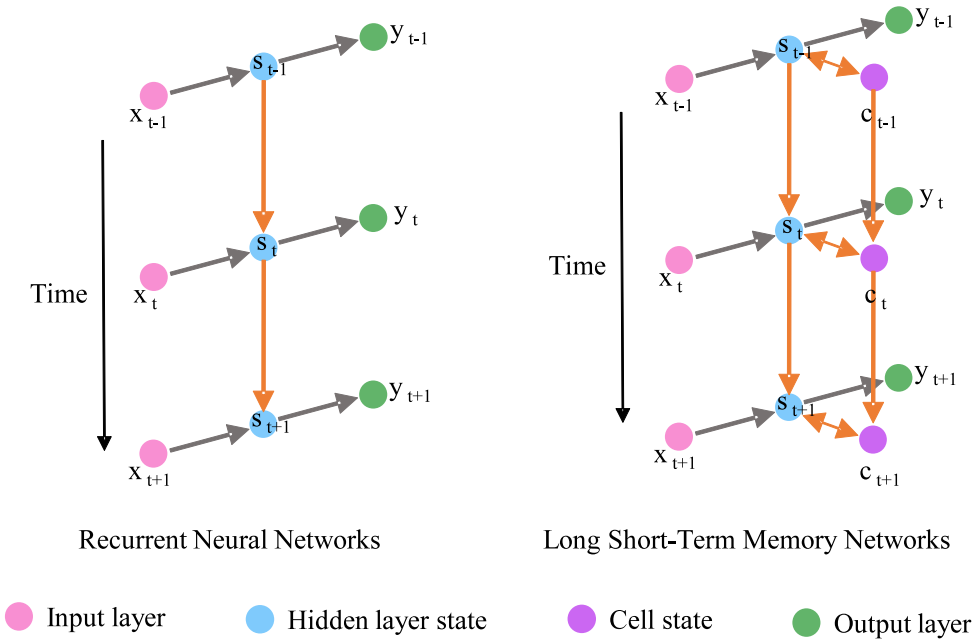



Fig. 3.3. Schematic of RNN and LSTM networks. The figure on the left shows a schematic of the basic RNN. The horizontal lines connect the network of time steps. Different rows represent multiple time steps organized in chronological order from top to bottom. The right-hand figure shows a schematic of the LSTM network with the cellular structures added to the RNN structure.

LSTM has a more intricate cellular structure. The cell state enables the network to finely control the information flow, prevent gradient vanishing, and effectively manage the long-term memory. Consequently, this enhancement in sequence handling allows the LSTM to capture and utilize long-term information more efficiently than a standard RNN.

The advantages and disadvantages of the traditional decision tree, SVM decision tree, and ELM decision tree are presented in Table 3.3.

### 3.3.3 Experiment

The experiment is divided into two parts, namely shallow classifier ensemble for speech emotion recognition and enhanced speech emotion recognition using LSTM. These two parts use different deep learning perspectives for the speech emotion recognition task. The shallow classifier ensemble in the first part is mainly based on traditional machine learning methods such as Support Vector Machine (SVM) and Decision Tree. These methods usually use hand-designed features or shallow network-based feature extraction methods, which are then fed into the classifier for training and classification. This approach is simple and straightforward, easy to understand and implement, and suitable for small-scale

Table 3.3. Comparison of the traditional decision tree, SVM decision tree, and ELM decision tree.

| **Decision-Tree Type** | **Advantages** | **Disadvantages** |
|---|---|---|
| Traditional Decision Tree | - Easy to understand and interpret<br>- Can handle both discrete and continuous features<br>- Robustness | - Prone to overfitting<br>- Instability<br>- Ignores feature correlations |
| SVM Decision Tree | - Effective for high-dimensional data<br>- Strong generalization ability<br>- Can handle nonlinear classification problems | - High computational complexity<br>- Parameter tuning required<br>- Not suitable for large-scale datasets |
| ELM Decision Tree | - High computation speed<br>- Good scalability | - Insensitive to initial weights |

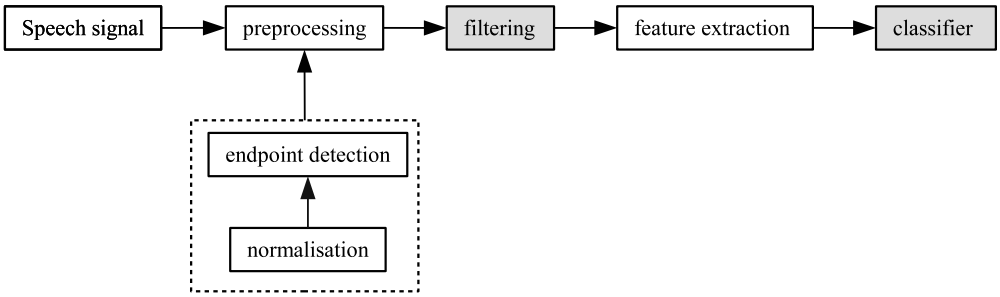



Fig. 3.4. Flowchart of the experiments.

speech emotion recognition tasks. While in the second part, Long Short-Term Memory (LSTM) network is used to enhance the performance of speech emotion recognition.LSTM is a deep learning model for sequential data processing, which is able to capture long-term dependencies in sequential data and has advantages for processing time-series data such as speech signals. By using the LSTM model, we can better model the temporal information in speech signals and improve the accuracy and robustness of speech emotion recognition.

These two parts of the experiment explore and compare the effect and performance of different deep learning methods in speech emotion recognition tasks from different perspectives, providing reference and guidance for practical applications.

#### Ensemble of Shallow Classifiers for Speech Emotion Recognition

In these experiments, speech emotion recognition was performed via the process shown in Fig. 3.4. The diagram outlines the progression from raw input data to the classification of emotions. Within this experiment, the initial speech data underwent preprocessing, followed by the extraction of short-time features. Subsequently, the speech rate was computed based on these extracted features. Based on the speech rate, speech data were supplied to the 5 shallow classifiers for training. The completed trained model recognizes the emotion of the speech data, thus completing the speech emotion recognition based on the speech rate. The configuration involved in the experiment is presented in Table 3.4.

The input speech signal was normalized to the range of $[0-255]$ and endpoint detection was performed. The purpose of the normalization is to unify the evaluation metrics and eliminate the amplitude differences [154]. The purpose of endpoint detection is to remove blank data at the beginning and end of the speech signal to reduce interference in the experiments and computational cost [155]. This experiment obtained short-time features using Python *wave* library. The common short-time features are presented in Table 3.1. Short-time features were selected according to three criteria: complexity, expression of speech time-domain features, and ability to identify voiced sound locations. As shown in Table 3.5, the short-time autocorrelation, short-time pitch, and short-time peak amplitude do not provide a clear indication of the location of the voiced sound. The short-time harmonic ratio and short-time cepstral coefficients are used to evaluate speech frequency

Table 3.4. Experiment settings

| **Project** | **Configuration** |
|---|---|
| Operating system | Ubuntu 20.04, Windows 10 |
| Programming language | Python 3.7 and Pytorch |
| Libraries | librosa, numpy, sklearn, torch, wave |
| Filter | moving average filter, savgol filter |
| Learning rate | 0.001 |
| Optimizer | Adam |
| Hiddern size | 256 |
| Number of layers | 3 |
| Epochs | 500 |

features. Consequently, this experiment selected three short-time features to describe the speech rate: short-time energy, short-time zero-crossing rate, and short-time average amplitude difference. Three short-time features were filtered using the filter. A Savgol filter was used for the short-time energy and short-time zero-crossing rate, and a moving average filter was used for the short-time average amplitude difference. Filtering makes the data smoother and makes it easier to calculate voiced sounds. In this experiment, Table 3.1 lists voiced sounds with short-term features. Based on this, the frames of the voiced sounds of the three short-time features were calculated individually. The experiment calculated the frames of the short-time energy peaks, short-time zero-crossing rate valleys, and short-time average amplitude difference variations. The results are shown in Fig. 3.5.

We replace the number of syllables with the number of frames between adjacent voiced sounds. Because of the voiced sound features of vowels, the focus of this research is finding the location of the voiced sound in the sentence. A sentence with frame length $N$ contains $n$ words and the multiple features $T$, as given by Equation (3.4):

$$T = \frac{\frac{1}{n-1}\sum_{1}^{n} f_i - f_{i-1}}{N}, \tag{3.4}$$

Here, $f_i$ denotes the $i^{th}$ frame of the voiced sound. The average of frame number between adjacent voiced sound is used as the speech rate of this sentence. We selected three short-time features to describe the voiced sound and used the SVM, ELM, and decision trees to evaluate the features.

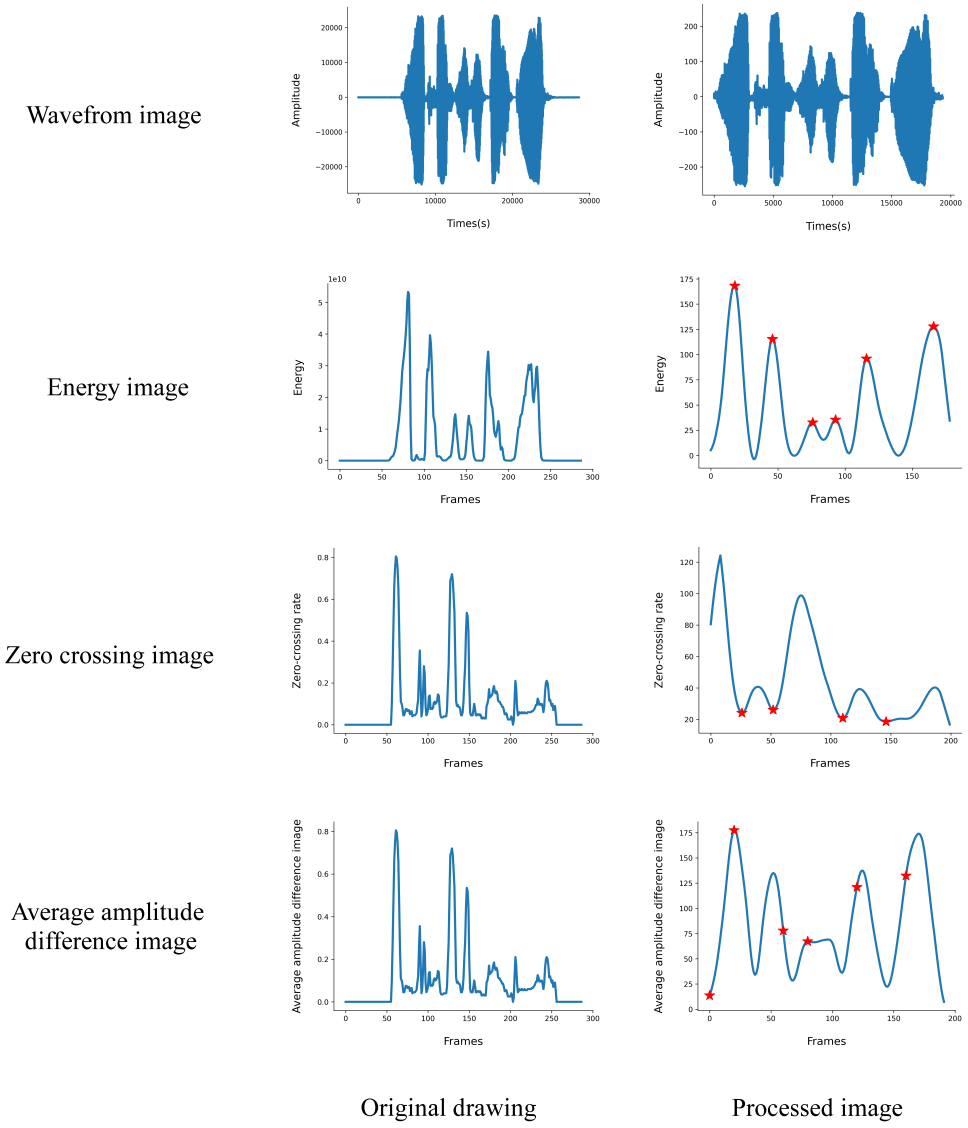



Fig. 3.5. A comparison between the original images and the image was filtered and labeled with the position of the voiced sound images. The first column of the figure represents the original images, from first row to last row: the original speech signal waveform image, short-time energy image, short-time zero-crossing rate waveform image, and short-time average amplitude difference waveform image. The second column shows the images regularized to the range of [-255, 255] from first row to last ros: the preprocessed speech signal waveform image, the voiced sound position in the preprocessed and filtered short-term energy image, the voiced sound position in the preprocessed and filtered short-time zero-crossing rate, and the voiced sound position in the preprocessed and filtered short-time average amplitude difference.

Table 3.5. Selection criteria for short-time parameters

| **Feature** | **C**[1] | **T**[2] | **V**[3] |
|---|---|---|---|
| Short-Time Autocorrelation | ✖ | ✔ | ✖ |
| Short-Time Average Amplitude Difference | ✔ | ✔ | ✔ |
| Short-Time Cepstral Coefficients | ✖ | ✖ | ✔ |
| Short-Time Energy | ✔ | ✔ | ✔ |
| Short-Time Harmonic Ratio | ✔ | ✖ | ✔ |
| Short-Time Peak Amplitude | ✔ | ✔ | ✖ |
| Short-Time Pitch | ✖ | ✖ | ✖ |
| Short-Time Zero-Crossing Rate | ✔ | ✔ | ✔ |

[1] **C** indicates the computational complexity of computing the feature: ✔implies indicates computational simplicity, and ✖implies indicates computational complexity.

[2] **T** indicates whether the feature can describe the speech signal in the time domain: ✔indicates that it can describe the speech signal in the time domain, and ✖indicates that it can describe the speech signal s in the frequency domain.

[3] **V** indicates whether the feature can indicate the position of the voiced sound clearly: ✔indicates that it can indicate the position, and ✖indicates that it cannot.

### Enhanced Speech Emotion Recognition using LSTM

The experimental process is illustrated in Fig. 3.6. The figure illustrates the process from the input data to emotion classification. In this experiment, the input speech data were preprocessed, and short-time features were extracted. The speech rate was calculated from the extracted features. Based on the speech rate, speech data were reconstructed to fit the LSTM network. Reconstructed data were supplied to the LSTM network for training. The completed trained model recognizes the emotion of the speech data, thus completing the speech emotion recognition based on the speech rate.

The experiments were based on computed voiced sound frames, and the speech data were reconstructed. The reconstructed data were used as inputs to the LSTM network for training. To satisfy the uniformity of the data size requirement supported by the LSTM network, the longest voiced sound interval in each short-time feature was calculated as the step size of the input data to the network, and the maximum number of voiced sound dots as the number of steps. Zero values were used to compensate for the shortfalls. As

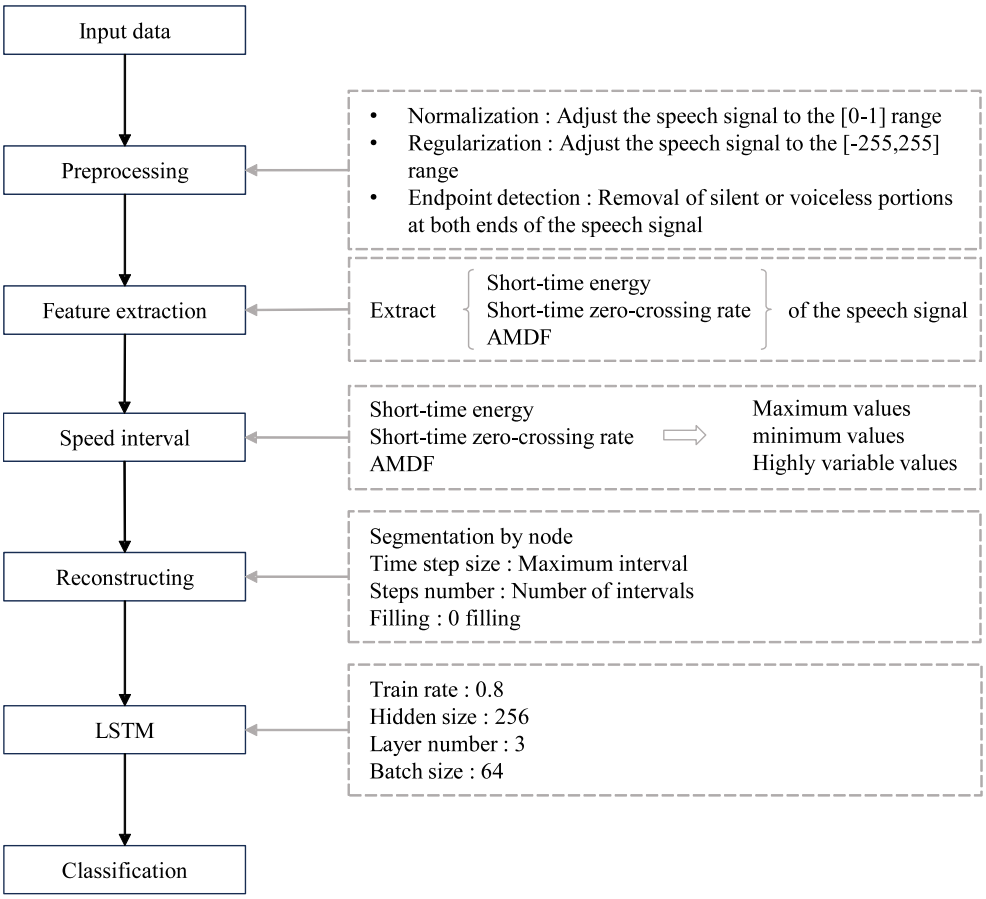



Fig. 3.6. Flowchart of the experiments.

shown in Fig. 3.7, the reconstructed data are trained on the network to obtain the best model. This model realizes emotion recognition of speech signals.

There were some problems in the experiments, mainly owing to the small size of the dataset, which led to insufficient training and thus had an impact on the results. To solve this problem, the experiment expanded the database [156]. By accelerating (1.5x speed) and decelerating (0.5x speed) the original speech signal, the experimental data size was expanded to three times the original data size [156]. The experiment and methodology employed are thoroughly detailed up to this point. The next section focuses on a comprehensive discussion of the obtained results, conducting an in-depth analysis, and exploration of the data and observations.

## 3.4 Results and discussion

In this experiment, multi-feature speech rates are proposed. Speech rate is obtained by calculating the voiced position of three short-term eigenvalues. This feature uses five shallow classifiers and LSTM network model respectively for emotion recognition.

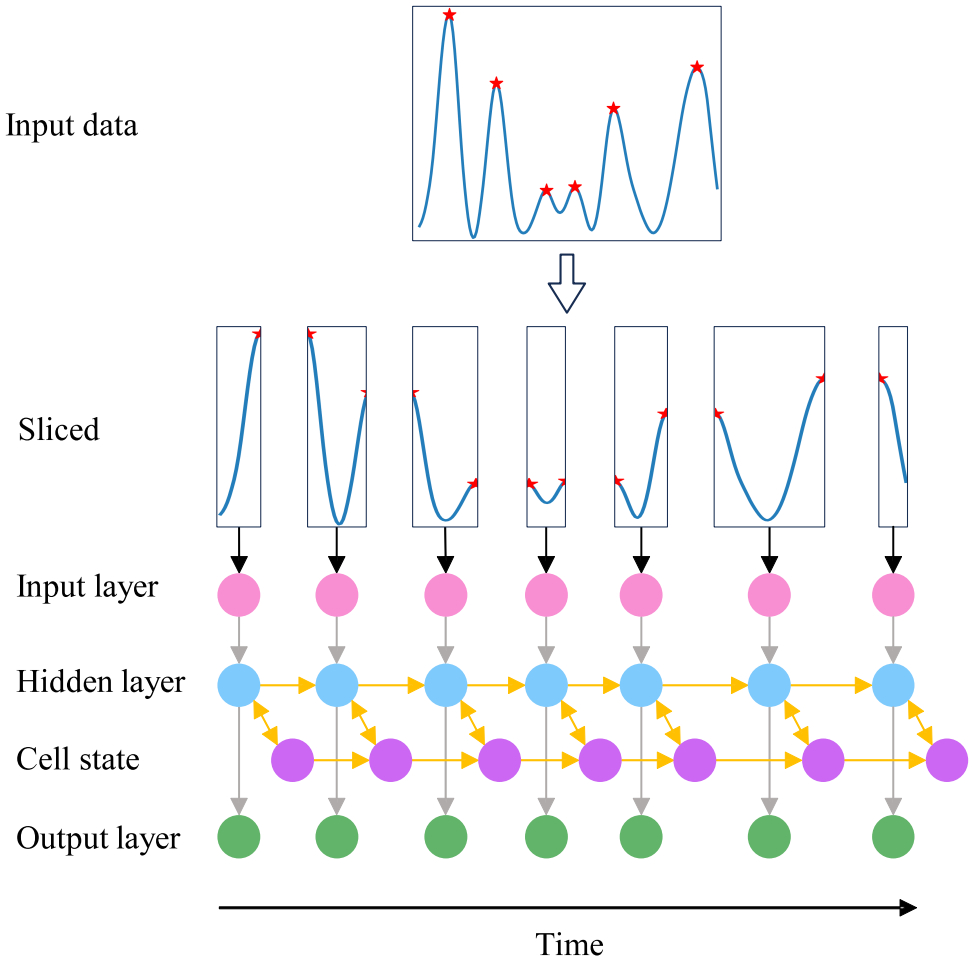



Fig. 3.7. Reconstruction the speech data for LSTM. The input speech data were segmented according to the frame nodes at which the voiced sounds were located. The speech signal in the example had six voiced nodes; thus, it was divided into seven segments. Each piece of speech was inputted into the LSTM network model as a single time step. The zero-filled speech segments were sequentially input into the LSTM network model in chronological order for learning to achieve speech emotion classification.

### 3.4.1 Experimental results based on shallow classifiers

We used five shallow classifiers to identify the speech emotions: the SVM, ELM, decision tree, SVM decision tree, and ELM decision tree. The accuracy and time cost for recognition are presented in Tables 3.6 and 3.7. We use Fig. *En* to represent the short-time energy, *Zcr* to represent the short-time zero-crossing rate, and *Amdf* to represent the short-time average amplitude difference. As shown in Fig. 3.8, we compared the corresponding multiple features with the single short-time feature. From 3.8(a), the accuracy of speech emotion recognition with multiple features was higher than that for the single

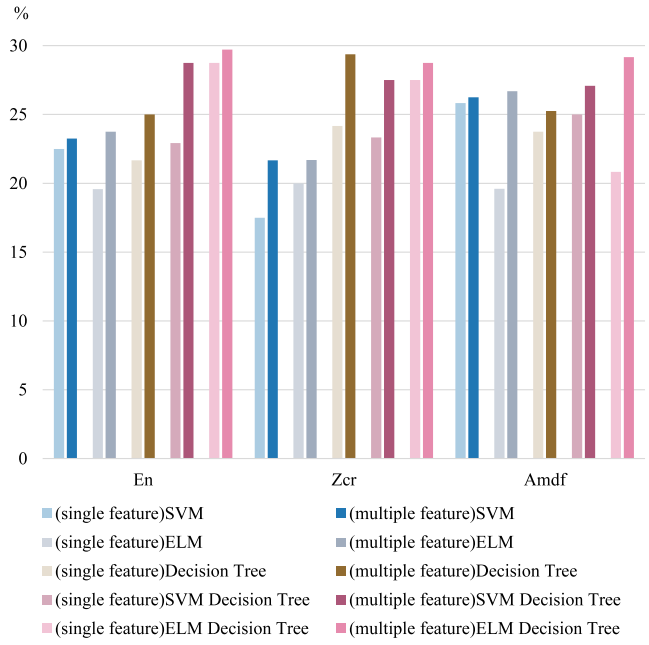
Table 3.6. Accuracy results for speech emotion recognition(%).

| | Short-time features | | | Multiple features | | |
|---|---|---|---|---|---|---|
| | En | Zcr | Amdf | En | Zcr | Amdf |
| SVM | 22.50% | 17.50% | 25.83% | 23.25% | 21.67% | 26.25% |
| ELM | 19.58% | 20.00% | 19.60% | 23.75% | 21.69% | 26.70% |
| Decision Tree | 21.67% | 24.17% | 23.75% | 25.00% | 29.37% | 25.25% |
| SVM Decision Tree | 22.92% | 23.33% | 25.00% | 28.75% | 27.50% | 27.08% |
| ELM Decision Tree | 28.75% | 27.50% | 20.83% | 29.72% | 28.75% | 29.17% |

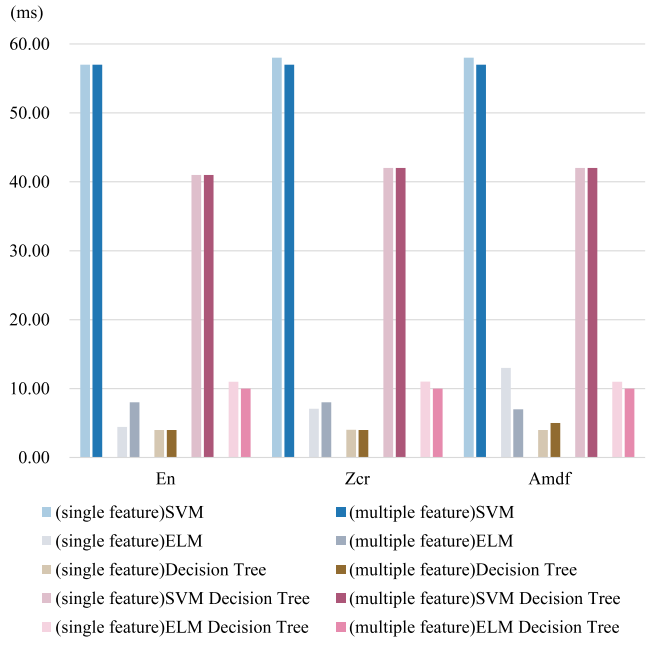
one. The highest identification accuracy was achieved by the ELM classifier with multiple short-time energy features . Fig. 3.8(b) indicates that the features had a similar time cost for speech emotion recognition.

We subtracted the accuracy and time cost of the corresponding single feature with the accuracy and time cost of each composite feature separately, as shown in Fig. 3.9. Fig. 3.9(a) indicates that compared with the single features, the multiple features yielded a higher accuracy. The short-time average amplitude difference based on the ELM decision tree exhibited the largest improvement (8.34%), and the short-time energy based on the ELM decision tree exhibited the smallest improvement (0.97%). The use of multiple features increased the accuracy by 3.40% on average. Fig. 3.9(b) shows the improvement of the computation time in the case of multiple features. The left side with red words indicates the time saved by using multiple features, and the right side with black words indicates the extra time used in the case of multiple features. The figure indicates that using multiple features took less time on average than using single features. Using multiple features took more time than using single features in 3 of the 15 cases, and it took less time in 7 of the 15 cases.

The results of the experiment indicate that the multiple features yielded a higher accuracy of speech emotion recognition than single short-time features. Additionally, in the case of multiple features, it took less time to achieve recognition. The results confirmed that the proposed feature outperformed short-time features alone. Thus, the proposed speech rate feature can solve the problems of a single feature and data redundancy.




(a) Accuracy (%)




(b) Time (ms)

Fig. 3.8. Results of speech emotion recognition.

Table 3.7. Time result for speech emotion recognition(ms).

| | Short-time features | | | Multiple features | | |
|---|---|---|---|---|---|---|
| | En | Zcr | Amdf | En | Zcr | Amdf |
| SVM | 57.01 | 58.01 | 58.02 | 57.01 | 57.01 | 57.01 |
| ELM | 4.43 | 7.06 | 13.00 | 8.00 | 8.00 | 7.00 |
| Decision Tree | 4.00 | 4.01 | 4.00 | 4.00 | 4.00 | 5.00 |
| SVMDecision Tree | 41.01 | 42.01 | 42.01 | 41.01 | 42.01 | 42.01 |
| ELM Decision Tree | 11.00 | 11.01 | 11.00 | 10.00 | 10.00 | 10.00 |

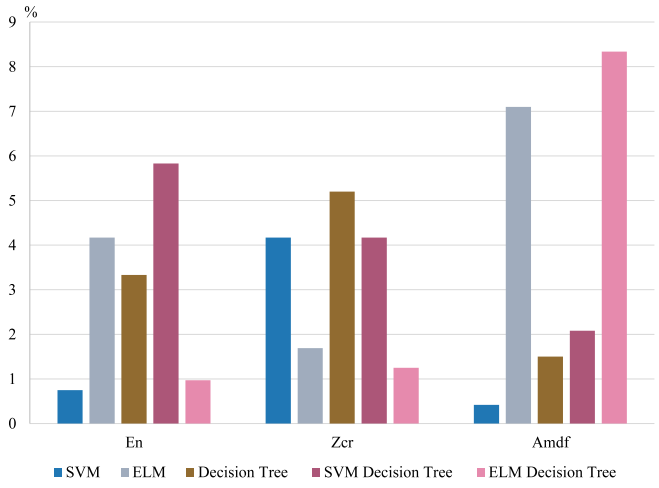
### 3.4.2 Experimental results based on LSTM

In this experiment, Accuracy was used as the main evaluation index, while the confusion matrix and ROC curve were used to demonstrate the classification results more intuitively.

1. **Accuracy** : Accuracy was calculated as the ratio of correctly predicted samples to the total number of samples in the classification problem. The formula for accuracy is expressed as a percentage, as shown in Equation (3.5). This signifies the proportion of all the samples that the model correctly classifies.

$$\text{Accuracy} = \frac{\text{Number of Correct Predictions}}{\text{Total Number of Predictions}} \tag{3.5}$$

2. **Confusion matrix** : A confusion matrix is used to evaluate the performance of a classification model, especially for classification problems in supervised learning. This shows the correspondence between the model's actual predictions and the actual labels in the test dataset. A typical confusion matrix is an $(N \times N)$ matrix where ($N$ represents the number of categories. The rows of this matrix represent the actual categories and the columns represent the predicted categories. In the confusion matrix, the diagonal elements represent the number of samples correctly predicted by the model, whereas the off-diagonal elements represent the number of samples incorrectly predicted by the model.
3. **ROC curve** : The receiver operating characteristic curve (ROC curve) is a tool for evaluating the performance of a binary classifier. In multi-classification problems,




(a) Accuracy (%)


(b) Time (ms)

Fig. 3.9. Comparison between multiple features and single features.

one-to-one and one-to-one rest methods can be used to draw and evaluate multi-category ROC curves. The one-to-one method plots a one-to-one ROC curve for each category by calculating each category separately in combination with the other categories. A pair of methods combines each category separately and draws the corresponding ROC curve. The area under the ROC curve (AUC) is an important indicator for evaluating the performance of a multiclass classifier. The closer the value is to one, the better the performance of the classifier. For the evaluation of multi-class classifiers, ROC curves and AUC provide a comprehensive measure of performance.

To verify the generality and robustness of the proposed method, experiments were validated using other datasets. Table 3.8 shows the experimental results for different datasets.

By testing our experimental results on speech data in multiple languages and emotion categories, we demonstrate that the proposed method exhibits good performance in emotion recognition tasks in most cases. Specifically, we observed that our model achieved high accuracy on most databases using short-time energy, short-time zero-crossing rate, and short-time average amplitude difference as features. For example, on the Emo-DB database, our model achieved 100% accuracy in speech emotion recognition, whereas close to or over 90% accuracy was obtained on other datasets. This demonstrates the strong generalization ability of our proposed speech-rate-based approach on the speech data of different languages and emotion categories, providing an effective solution for emotion recognition tasks.

Table 3.8. Presentation of experimental accuracy for different datasets (%)

| **Database Name** | **Language** | **Emotion Classes** | **En** | **Zcr** | **Amdf** |
|---|---|---|---|---|---|
| ANAD | Arabic | 3 classes | 96.83 | 94.72 | 95.66 |
| Emo-DB | German | 7 classes | 100 | 99.58 | 99.90 |
| CASIA | Chinese | 6 classes | 97.64 | 98.47 | 98.33 |
| EMOVO | Italian | 7 classes | 98.87 | 96.43 | 86.12 |
| RESD | Russian | 7 classes | 90.91 | 89.94 | 90.77 |
| TESS | Toronto | 7 classes | 96.20 | 96.89 | 97.58 |

This study performed a comparative analysis of three methods utilizing the same CASIA dataset for Chinese speech emotion recognition, as shown in Table 3.9. The DBN & SVM was proposed in [131]. This study explored methods for improving the accuracy of Chinese speech emotion recognition. The study extracted five features: MFCC, pitch, resonance peaks, short-time over-zero rate, and short-time energy through deep learning models. This study used a Deep Belief Network (DBN) combined with a Support Vector Machine (SVM) for the experiment, and the classification method achieved 95.8% accuracy, which exceeded the results of the DBN or SVM alone. The ELM decision tree was proposed in [122]. This study utilizes correlation analysis and Fisher's feature selection method to remove redundant features that are closely correlated. A sentiment recognition classifier based on an Extreme Learning Machine (ELM) decision tree was proposed. The proposed method achieved an average recognition rate of 89.6%.

By comparing these three methods, the proposed multi-feature speech rate achieved the highest recognition accuracy on the CASIA dataset. This shows that the use of speech-speed features to identify speech emotions is an effective method.

Table 3.10 shows the comparative results of the speech emotion recognition methods

Table 3.9. Comparison of speech emotion recognition methods for the CASIA dataset

| Method | Features | Network | Accuracy |
|---|---|---|---|
| DBN & SVM [131] | MFCC, pitch, resonance peaks, short-time over-zero rate, short-time energy | combine DBN and SVM | 95.8% |
| ELM decision tree [122] | correlation analysis, Fisher criterion | ELM decision tree, SVM decision tree | 89.6% |
| Our work | short-time energy, short-time zero-crossing rate, and short-time average amplitude difference. | LSTM | 97.64%, 98.47%, 98.33% |

on different datasets. The processing of audio clips using deep 1D and 2D CNN LSTM networks [157] in the Berlin EmoDB database achieved an accuracy of 92.34%. In contrast, the method using spectrogram features and combining them with CNN networks [158] achieved an accuracy of 53.11%. Using an approach based on the Bidirectional LSTM architecture and a deep confidence network [159], a BiLSTM network with RBF features achieved 85.57% accuracy.

In our study, we achieved 100%, 99.58%, and 99.90% accuracy on the Berlin EmoDB database using LSTM networks based on features such as energy (En), over-zero rate (Zcr), and average amplitude difference (AMDF). For the EMOVO database, our model achieved 98.87%, 96.43%, and 86.12% accuracies with the same features. Compared with previous methods, our method exhibits higher accuracy rates, proving its effectiveness and superiority in speech emotion recognition tasks.

## 3.5 Conclusions

The aim of this study is to explore and validate a novel approach to speech emotion recognition that utilizes a combination of short-and rhythmic features, with speech rate as a key parameter. By analyzing the speech rate information extracted from short-time features, our method avoids the overfitting problem that may be caused by complex computation, thus improving the accuracy and robustness of sentiment classification.

The experimental process involved several steps, starting with data preprocessing, including normalization and endpoint detection, to ensure the quality and consistency of the

Table 3.10. Comparison of speech emotion recognition methods on different datasets

| **Database Name** | **Paper** | **Features** | **Network** | **Accuracy (%)** |
|---|---|---|---|---|
| Berlin EmoDB | deep 1D & 2D CNN LSTM networks [157] | audio clips | 1D-CNN-LSTM | 92.34% |
| | Deep Learning Techniques for Speech Emotion Recognition [158] | Spectrogram | CNN | 53.11% |
| | Speech emotion recognition based on Bi-directional LSTM architecture and deep belief networks [159] | RBF | BiLSTM | 85.57% |
| | Our work | En<br>Zcr<br>Amdf | LSTM | 100%<br>99.58%<br>99.90% |
| EMOVO | Improved speech emotion recognition with Mel frequency magnitude coefficient [160] | MFMC 12<br>MFMC 24<br>MFMC 30 | SVM | 64.12%<br>70.92%<br>73.30% |
| | Our work | En<br>Zcr<br>Amdf | LSTM | 98.87%<br>96.43%<br>86.12% |

data. We then extracted short-time features using Python's wave library and calculated the speech rate. Next, we reconstructed the speech data using the extracted features and used them as the training input for the LSTM neural network. During model training, we used a series of hyperparameters and configurations such as the learning rate, optimizer, hidden layer size, and number of network layers. Through repeated experiments and tuning, we determined the optimal parameter settings for achieving the best emotion recognition performance.

We conducted experimental validation on multiple speech databases, and the results showed that our method achieved a significant accuracy improvement for each database. Specifically, the accuracies of the different databases were 97.64%, 98.47%, and 98.33%, respectively, demonstrating the wide applicability and effectiveness of speech rate as an important parameter for speech emotion recognition.

However, this study also has some limitations and challenges. First, the dataset size was relatively small, which may limit the training and generalization ability of the model. To mitigate this problem, we adopted the strategy of expanding the dataset to improve the model performance by increasing the amount of data. Second, our method must be further optimized to improve its robustness and generalization in complex speech environments.

In summary, this study provides a new method for speech emotion recognition using speech rate analysis. Future research can explore more data enhancement techniques and model optimization strategies to further improve the accuracy and robustness of sentiment recognition. This will promote the development of speech emotion recognition technology and provide more accurate and reliable emotion recognition capabilities for speech intelligence applications.

# Chapter 4

# Proposal 3: Speech Emotion Recognition Based on Hilbert Curve

In the field of speech emotion recognition, researchers have been exploring improved representation methods to enhance the capture of emotional information. Traditional one-dimensional time series classification exhibits limitations in expressing complex emotional patterns, particularly due to the nuanced and intricate nature of emotional information in speech signals, posing challenges in terms of accuracy and robustness in emotion classification tasks. This chapter proposes an innovative algorithm aiming to effectively capture and express emotional information by refining the representation method for time series. The method utilizes Hilbert curves to transform one-dimensional speech data into two-dimensional data, preserving the integrity of the data for more accurate feature extraction. Additionally, a tiling module based on fully connected layers is designed to fully leverage the arrangement of Hilbert curves, unfolding multidimensional feature maps extracted by convolutional neural networks based on temporal correlations for better capturing emotional information. The experimental results clearly demonstrate the outstanding performance of this approach. Firstly, in terms of spatial utilization efficiency, this method can save up to 23,195 times the pixel units compared to other time series imaging methods, effectively improving data storage efficiency. Secondly, in terms of accuracy, this method significantly outperforms other methods on the same dataset, particularly when using the Hilbert unfolding data representation method, achieving a remarkable accuracy of 98.73%. Most importantly, compared to traditional classification methods, this approach achieves higher accuracy on the same dataset, highlighting its superior performance in emotion classification. These data-driven findings effectively showcase the effectiveness and superiority of this proposed method, providing a solid empirical foundation for this research work.

## 4.1 Introduction

Emotions play an important role in human life and profoundly influence the psychology and behavior of individuals [161]. They are not only essential for individual survival and adaptation[162], but also play a crucial role in interpersonal relationships, social interactions [163], and decision-making [164]. Emotions are essential for expressing and understanding human experience, and they serve as a link between individuals and the world around them [165]. In social settings, emotions convey information, intentions, and relationships [166]. People express emotions to make connections, convey empathy, and generate deeper understanding in their interactions [167]. Emotions also play a key role in decision making [168]. Individuals' emotional states may influence their decision preferences and risk tolerance levels [169, 170]. Positive emotions may prompt more optimistic and risky decisions, while negative emotions may lead to more conservative and cautious choices [171, 172]. Therefore, understanding and identifying emotions is crucial to better understand individual behavior and decision-making processes.

Speech Emotion Recognition (SER), a branch of emotion detection, is an important form of emotional expression and accounts for 38% of emotional communication [173]. This subject focuses on the recognition of emotions in speech without considering semantic content [174, 175]. Speech signals contain many acoustic features that can reflect the emotional state of the speaker as well as information related to the speaker and speech. Therefore, the central concept of emotion detection is to study the acoustic differences produced when speech is vocalized in different emotional contexts [176].

SER has a wide range of applications in the field of emotion-related information. For example, speech emotion recognition provides important assistance in areas such as healthcare, education, and social interactions [177, 178, 179]. It helps monitor stress-induced changes in mental health [180], conduct mental health assessments [181], and quickly predict depression severity [182]. In social media analytics, speech emotion recognition plays an active role in improving decision–making in the tourism industry [183, 184]. In addition, it helps to collect and analyze people's opinions and impressions of various topics, products, themes, and services, improving their sense of well-being and satisfaction, and thus contributing to social stability and harmony [184] [185].

The convergence of SER with real–life applications enhances the close connection between human and computer interaction. By enabling computer systems to recognize and respond to human emotions, SER offers the possibility of more intuitive and emotionally rich human-computer interactions in various domains [178]. For example, in 2017, researchers [186] proposed a flexible emotion recognition system that utilizes visual and auditory input analyses. In 2020 [187], three models or experts were integrated using integrated learning, focusing on different feature extraction and classification strategies.

Other studies [188] have employed rectangular filters and improved pooling strategies to create lightweight SER models. They used a CNN approach to learn deep frequency features and trained CNN models for sentiment assessment using frequency features extracted from speech data. In addition, a new SER approach is based on the bidirectional attentional long short-term memory (BLSTMwA) model and deep convolutional neural networks (DCNN) [189]. Integrated classifiers have also been created using deep convolutional recurrent neural networks, specifically for SER [190]. These studies enrich knowledge in the field of SER and provide more options for practical applications.

However, one of the main challenges in the field of SER is to extract relevant features from speech signals to recognize emotional states [191, 192] and to develop appropriate classifiers [193, 175]. Speech feature extraction is considered to be a key issue in speech emotion recognition systems. Many studies have proposed a variety of speech features, such as pitch, energy, frequency, linear predictive resonance frequency coefficients (LPCC), MFCC, and modulation spectrum features that reflect the speaker's emotional information [194, 195, 196]. Therefore, many studies have combined multiple types of affective features to adequately characterize the available speech signals. However, combining multiple affective features increases the dimensionality and redundancy of speech data, thereby increasing the learning difficulty of most machine learning algorithms and the risk of overfitting [197]. Against this backdrop, two studies explored new approaches to combat this problem. Wang et al. proposed a compelling approach for transforming a one-dimensional speech time series into more informative two-dimensional image data [92]. This study utilizes the comprehensive and versatile nature of computer vision to improve the efficiency and accuracy of speech emotion recognition. Its innovation lies in presenting speech signals in a novel manner, providing more possibilities for sentiment analysis. Another study utilized the quasi-periodic nature of speech signals to transform one-dimensional speech data into more information-rich two-dimensional data by periodically segmenting speech signals [198]. The uniqueness of this approach is that it exploits the periodic nature of speech signals to improve emotion recognition performance. These two research results introduce new ideas and methods in the field of speech sentiment recognition that are expected to improve the accuracy and practicality of sentiment analysis. However, Wang et al. and Bakhshi et al. proposed two methods present unique challenges and complexities in the field of SER. For Wang et al.'s approach, utilizing the Gram angle field for converting one-dimensional speech signals into two-dimensional images introduces challenges in terms of the accuracy and robustness of the transformation. On the other hand, Bakhshi et al.'s method, based on the periodicity of the speech signal, faces challenges related to the generalizability of the conversion technique.

Building on these foundations, this study proposes a new method that adeptly tackles challenges associated with standardization, calibration, interpretability, and generalizability in emotion recognition from speech signals. This method endeavors to convert one-

dimensional time series data into more informative two-dimensional images. Inspired by converting one-dimensional time-series data into a two-dimensional image with the structure and shape of a Hilbert curve. Subsequently, a convolutional neural network is utilized to extract the features of the image, and the extracted feature values are passed through a fully connected network for sentiment classification. It is worth mentioning that compared the performance of this method with that of previous studies ([92] and [198]). This comparison validated the feasibility and effectiveness of the proposed algorithm in the field of emotion recognition. The results of this study show that the proposed algorithm has potential advantages in emotion-recognition tasks. The innovative aspect of this study is that it not only provides a new method for processing one-dimensional time-series data but also demonstrates the potential application of this method in emotion recognition. By presenting time series data as two-dimensional images, open up more possibilities in the field of sentiment analysis, with promising improved accuracy and performance. The success of this approach demonstrates the potential value of deep learning and computer vision for sentiment recognition.

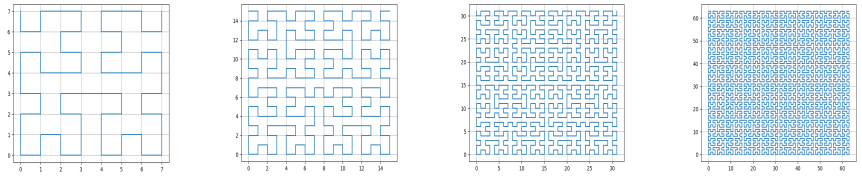
The remainder of this paper is organized as follows. Section 4.2 briefly describes the proposed method in this study. Section 4.3 provides details about the specific experimental procedures and dataset. Section 4.4 presents the results obtained from the experiments and engages in a discussion of these results. Finally, Section 4.5 concludes the study.

## 4.2 Hilbert Curve Path Arrangement method

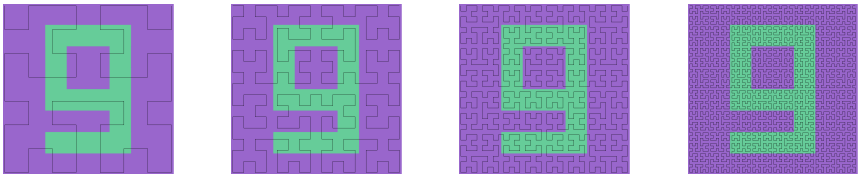
This study introduces an innovative method for transforming one-dimensional time series into two-dimensional images. Specifically, the approach involves arranging the signal along the path of the Hilbert curve, mapping the speech signal onto an image that follows the trajectory of the Hilbert curve. The purpose of using Hilbert curves is to reduce information loss and thus improve the accuracy of the conversion from 1D time series to 2D images. By mapping the speech signal onto the trajectory of the Hilbert curve, the temporal correlation is preserved, making the new representation more capable of capturing the temporal correlation and dynamic features in the speech signal.

The Hilbert curve, proposed by German mathematician David Hilbert in 1891 [199], is a special kind of space-filling curve with a wide range of applications. This curve can fill any bounded two-dimensional space, and is unique in that it maintains the locality of neighboring points, making neighboring points also neighboring on the curve, which facilitates localized access to spatial data. As shown in Fig. 4.1, the image represents the morphology of Hilbert curves of different dimensions in two dimensions. In addition, Hilbert curves are able to map multidimensional data onto one-dimensional curves, improving the efficiency of indexing and searching the data, as well as having a nested structure that allows for resolution adjustments as needed. The image representation utilizes Hilbert curves to

achieve a mapping of a two-dimensional image onto a one-dimensional curve as shown in Fig. 4.1(a). Such curves are also used in image compression and coding, where pixels are arranged in the order of the curve to achieve a compact representation of the data. Thus, Hilbert curves play a key role in the fields of geographic information systems, image processing, database query optimization, and distributed computing, and are particularly suitable for processing spatial and multidimensional data.




(a) Original shapes of Hilbert curves of dimensions 3, 4, 5, and 6.



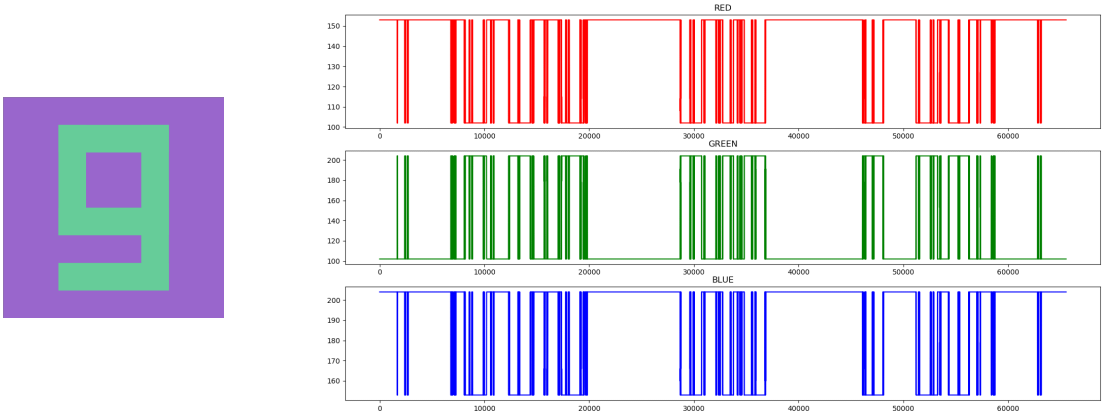

(b) Hilbert curves represent the localization of neighboring points on a two-dimensional image 1.


(c) Hilbert curves represent the localization of neighboring points on a two-dimensional image 2.

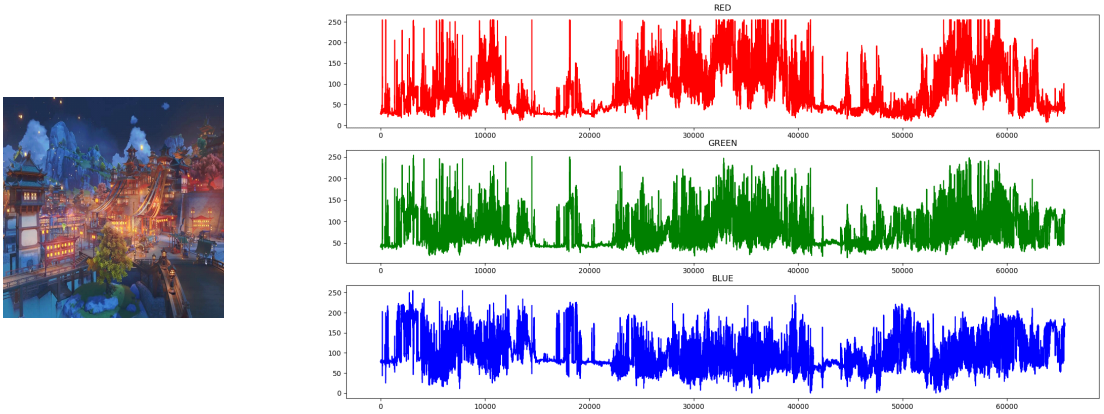
Fig. 4.1. The two-dimensional shape of the Hilbert curve. Fig. 4.1(a) shows the original shape of the Hilbert curve in dimensions 3, 4, 5, 6. Fig. 4.1(b) shows how the Hilbert curve represents the localization of neighboring points on a two-dimensional image. It can be observed from the figure that the higher the dimension of the dimension of the Hilbert curve, the richer the data between neighboring points, and the better the curve can describe the image.

In this study, two innovative proposals based on mapping and inverse mapping of Hilbert curves are presented to achieve an efficient conversion of 1D time series to 2D images.

1. **Conversion from one to two dimensions**: In this study, an innovative one-dimensional to two-dimensional upscaling method is proposed through the mapping of Hilbert curves. The overall pipeline of conversion from one to two dimensions is




(a) Waveforms of RGB image 1 of Hilbert curve projections.




(b) Waveforms of RGB image 2 of Hilbert curve projections.

Fig. 4.2. The Hilbert curve represents multidimensional data on a one-dimensional curve. Fig. 4.2(a) corresponds to Fig. 4.1(b) and Fig. 4.2(b) corresponds to Fig. 4.1(c), showing how the Hilbert curves can be transformed into each other in one or two dimensions. For ease of understanding, the experiment split the two-dimensional image according to the RGB three-color channel. Using the Hilbert curve, each channel was separately mapped from the higher-dimensional space to the one-dimensional space. The waveform in the projection shows the R, G, and B colors of the two-dimensional image according to the colors. The horizontal coordinate represents the number of sampling points and the vertical coordinate represents the size of the color pixel value. The number of sampling points determined the dimensions of the Hilbert curve. In this figure, every 20 pixel points was sampled to obtain the final waveform image. Fig. 4.2(b) shows a Hilbert waveform graph corresponding to a normal image.

illustrated in Fig. 4.3. This experiment involves treating speech data as grayscale pixel values in an image and plotting them according to the order of the Hilbert curve, creating an ordered image, as shown in Fig. 4.4. The image dimensions are set to (512, 512), determined by the minimum size required to encompass all speech data with a dimension value of 9. For cases where the data is insufficient to fill the entire image, two padding methods are employed: zero-padding and repetitive data read for padding, with excess portions discarded. Ultimately, this process provides an innovative approach, leveraging the ordered nature of the Hilbert curve, to represent speech information in the form of an image. The flexible handling of padding accommodates different image sizes without compromising information integrity.

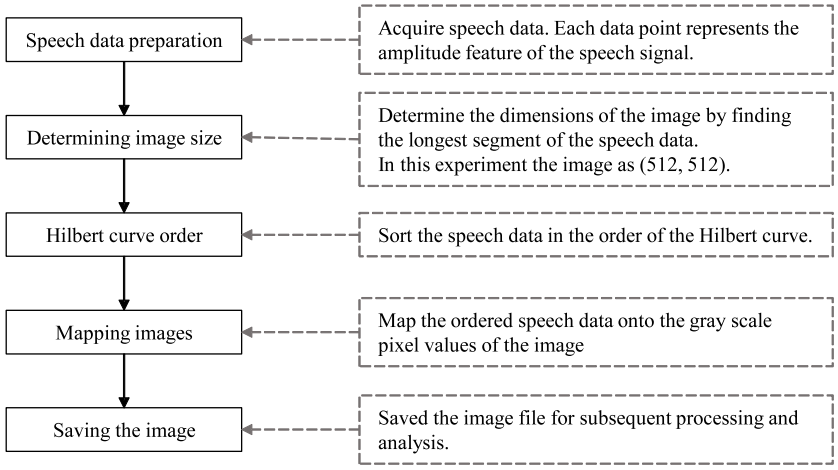
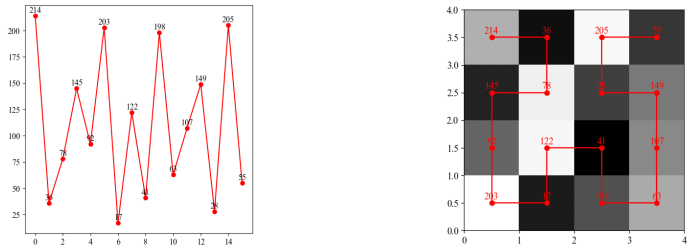



Fig. 4.3. Illustration of conversion from one to two dimensions. This experiment transforms speech data into an ordered image by mapping grayscale pixel values according to the Hilbert curve, with a fixed image size of (512, 512) determined by the longest segment of speech data, utilizing different padding techniques to handle incomplete regions.

2. **Conversion from two to one dimension**: In order to restore the information effectively, this study also proposes an innovative 2D to 1D dimensionality reduction method based on the Hilbert curve. The dimensionality reduction process from 2D to 1D is realized by inverse mapping, i.e., rearranging the 2D image onto the trajectory of the Hilbert curve. The images processed by the convolutional neural network consist of a set of feature maps with multiple channels, exemplified here with three channels as shown in Fig. 4.5(a). These feature maps are unfolded along the path of the Hilbert curve and concatenated in channel order, forming a one-dimensional array as illustrated in Fig. 4.5(b). This unfolding method better preserves the temporal correlation of speech data, ensuring that the order of features in the one-dimensional array aligns with their temporal positions in the original image. This Hilbert curve-based transformation from two-dimensional to one-dimensional



(a) One-dimensional time series before conversion.

(b) 2-dimensional image after conversion.

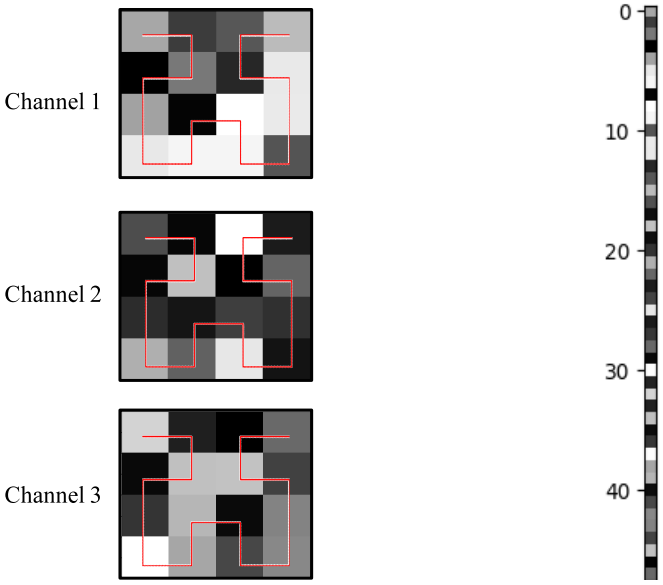
Fig. 4.4. Illustrate the conversion of one-dimensional speech data into two-dimensional images. This example illustrates a time series of length 16 as shown in Fig. 4.4(a), where the sequence values are arranged as grayscale pixel values in accordance with the Hilbert curve's order, resulting in a two-dimensional image as shown in Fig. 4.4(b). To provide a clearer representation of the image, study annotates the pixel values at corresponding positions and depict the path of the Hilbert curve. This process aims to articulate the mapping of speech data to an image in a more academically nuanced and fluent manner.

effectively maintains the temporal structure while taking into account the temporal correlation of the multi-channel feature maps produced by the convolutional neural network. It provides a more accurate and ordered representation for further analysis of speech data.

These two proposals aim to optimize the data representation and reduction process to meet the high demands for accuracy and effectiveness in the field of emotion recognition. By incorporating the unique mathematical properties of Hilbert curves, these proposals aim to provide a more informative representation of speech signals, thus providing a more reliable and powerful tool for emotion recognition tasks.

## 4.3 Experimentation and methodology

This subsection describes the specific experimental process of how to use Hilbert curves for conversion. First, the research collected a batch of speech samples and pre-processed them, including steps such as denoising, segmentation, and feature extraction. Then, converted each speech sample into its corresponding Hilbert curve representation. This step involves subjecting the speech signal to the Hilbert transform to obtain its Hilbert representation in complex form. Next, the research perform further processing on the obtained Hilbert curve, including smoothing, noise reduction, etc. This step aims to extract the key information in the speech signal and remove unnecessary noise and interference. After the Hilbert curve processing, input it as features into our model for training and testing.




(a) One-dimensional time series before conversion.

(b) 2-dimensional image after conversion.

Fig. 4.5. Illustrate the conversion of two-dimensional speech data into one-dimensional images.

The research employ various machine learning or deep learning models for the emotion recognition.

### 4.3.1 Dataset

This chapter utilized the CASIA Emotion Corpus, which features recordings from four professional speakers reciting 50 unique sentences. Each sentence was associated with one of six emotions: anger, happiness, fear, sadness, surprise, and neutrality. The corpus contains 1200 sentences, offering a valuable resource for analyzing the acoustic and prosodic attributes associated with various emotional states. These sentences vary in length, spanning five, six, and eight characters, respectively as shown in Table 4.1. To facilitate this study, the corpus was randomly divided into a training set (80%) and a test set (20%), as outlined in Table 4.2. Speech recordings were conducted within a controlled studio environment, free from external noise, with a sampling frequency of 16000 Hz and a storage format of PCM, 16-bit. The emotion labels were provided by the recording unit [200].

### 4.3.2 Experiment

The dataset and device information used in the experiment as shown in Table 4.3.

The entire experimental flow is shown in the Fig. 4.6, which helps us better understand the method and steps of the experiment. This method fully utilizes the potential of

Table 4.1. Attribute of the corpus.

| | |
|---|---|
| Corpus | CASIA |
| Language | Chinese |
| Size | 50 utterances × 4 actors × 6 emotions |
| Subject | 4 professional actors |
| Emotions | angle, fear, happy, neutral, sad, surprise |

Table 4.2. Composition of the corpus.

| Number of words | 5 words | 6 words | 8 words | Total |
|---|---|---|---|---|
| Sentences numbers | 192 | 960 | 48 | 1200 |
| Train set numbers | 144 | 768 | 24 | 960 |
| Test set numbers | 48 | 192 | 24 | 240 |

Table 4.3. Experimental environment and parameter configuration.

| **configuration item** | **Numeric/Descriptive** |
|---|---|
| Programming Language and Version | Python 3.11 |
| Deep Learning Framework | PyTorch |
| Libraries * | cv2, hilbertcurve, matplotlib, numpy, Scikit-image, torch, wave |
| Epochs | 100 |
| Training Rate | 0.8 |
| Optimizer | Adam |

* The libraries listed in the table follow Python's naming conventions, are sorted alphabetically, and can be installed using the *-pip* command.

computer vision and deep learning techniques in one-dimensional time-series data analysis.

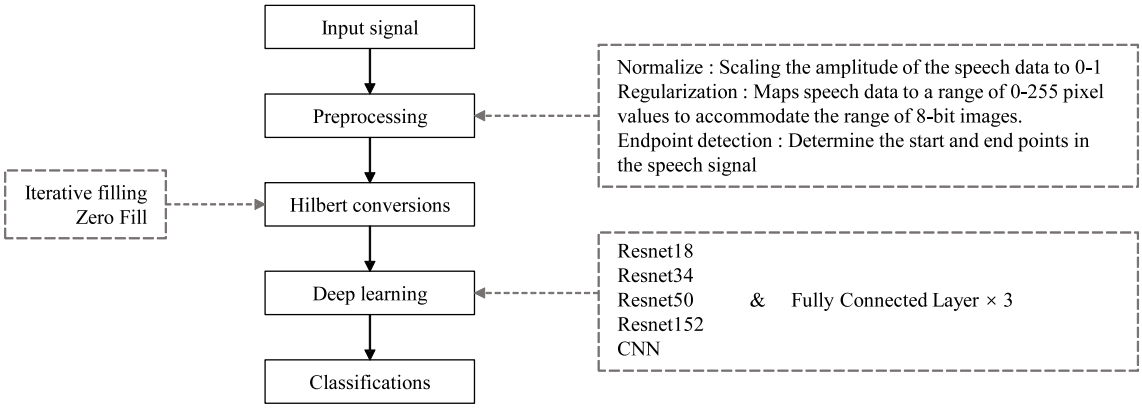



Fig. 4.6. The flowchart of the experiment.

In the preprocessing stage, the experiment first normalizes the speech signal between [0-1], and then regularizes it to make its value range between [0-255] to meet the requirement of an 8-bit image. Next, the one–dimensional data were converted to two–dimensional data according to the properties of the Hilbert curve. The relationship between the dimensions of the Hilbert curve and the length of the time series $X$ is presented in Table 4.4. For speech data of length $2^{2n}$, the dimension of the Hilbert curve is typically $(2^n, 2^n)$, which is converted into a Hilbert curve $2^n \times 2^n$. For speech data with lengths between $2^{2(n-1)}$ and $2^{2n}$, two different methods of data supplementation are used to ensure that the Hilbert curve dimension requirement is satisfied. Method one was to use for 0–padding, while method two was to iterate the data repeatedly to satisfy the dimensionality requirement. The method of complementing zeros was chosen primarily because of the translation invariance characteristic of the convolution operation [201].

The experimental results are shown in Fig. 4.7. Fig.4.7(a) represents the portion filled using 0, whereas Fig. 4.7(b) represents the speech signal replicated iteratively to satisfy the Hilbert curve dimensionality requirement. These generated images were feature-extracted using a convolutional neural network and then classified using a fully connected layer.

For comparison, different neural networks were used in the experiments to extract the feature values. One is the ResNet network model and the other is the experimentally proposed Hilbert-CNN network model, as shown in Fig. 4.8. Fig. 4.8(a) shows a flowchart of the network, using the Hilbert curve as the unfolding layer. The network was first convolved thrice to obtain a feature map. The feature map is then unfolded using the Hilbert curve. The flattened feature map passes through three fully connected layers to extract feature values. Finally, classification results were obtained. The data related to the convolution are shown in Fig. 4.8(b). The size of the convolution kernel was (4, 4), and the step size was 4. The size of the input image was (1, 512, 512), and the size of

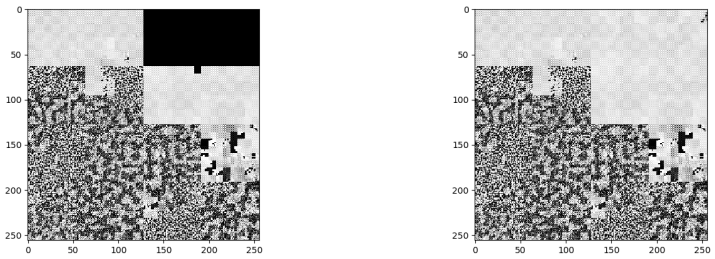
Table 4.4. Hilbert Curve Dimension and Time Series Length Relationship.

| **Dimension** | **Length of Time Series**[1] | **Image Size ((h, w))**[2] | **Speech Duration (sec)**[3] |
|---|---|---|---|
| 5 | 1024 | (32, 32) | 0.064 |
| 6 | 4096 | (64, 64) | 0.256 |
| 7 | 16384 | (128, 128) | 1.024 |
| 8 | 65536 | (256, 256) | 4.096 |
| 9 | 262144 | (512, 512) | 16.384 |

[1] Speech Duration (sec) = $\frac{Number\ of\ Data\ Points}{Sampling\ Frequency}$, the dataset was sampled at 16,000Hz.

[2] Image size =$(2^n, 2^n)$,$n$ is the dimension. $h$ denotes the height of the image, $w$ denotes the width of the image

[3] Legnth = $H \times W = (2^n)^2$ Length of time series indicates the maximum number of time series points that can be included.



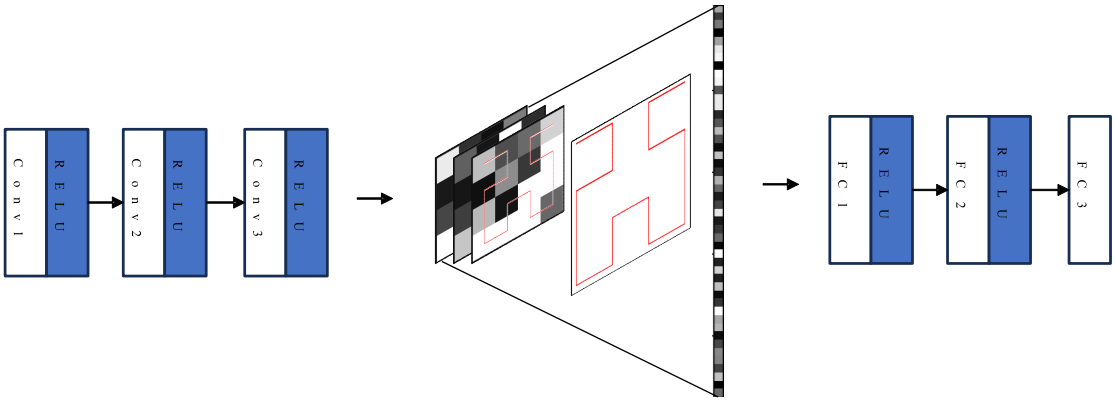
(a) Hilbert imaging algorithm with 0-filling.

(b) Hilbert imaging algorithm with iterative padding.

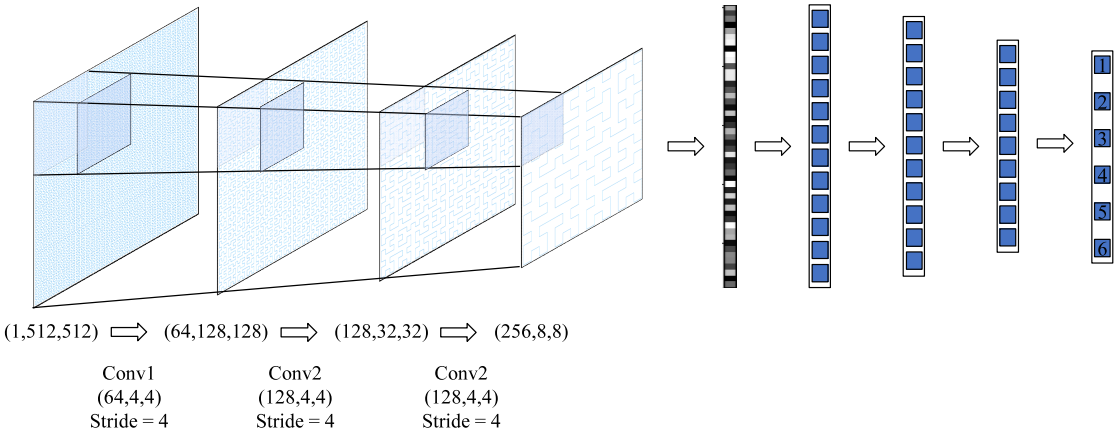
Fig. 4.7. The result of Hilbert imaging algorithm.

the convolved image was (256, 8, 8). Classification is performed by unfolding the Hilbert curve and inputting the fully connected layer. Because of the specificity of Hilbert-CNN, it is only used as a partial parameter for Hilbert imaging.

The entire experimental flow as shown in Fig. 4.6, which helps better understand the methods and steps of the experiment. This method fully utilizes the potential of computer vision and deep learning techniques for one-dimensional time-series data analysis.




(a) Hilbert-CNN network flow chart.




(b) Hilbert-CNN network flow chart schematic.

Fig. 4.8. Hilbert-CNN network model structure.

## 4.4 Results and discussion

In this section, this study delve into different methods for converting one-dimensional data into two-dimensional images, and the performance of these methods is compared and analyzed in detail. The method used in this study is primarily based on the properties of the Hilbert curve, which can convert one-dimensional time-series data into a two-dimensional image form with more information. By augmenting the training dataset with a 1.5x acceleration and 0.5x deceleration, the aim is to enhance the model's robustness and generalization to diverse speech features, encompassing varying speeds and speech rates. This expansion seeks to more comprehensively capture and process the nuanced characteristics present in real-world speech. To assess the effectiveness of this method, this study used accuracy as the evaluation index, which is the ratio of the number of samples correctly predicted by the model to the total number of samples in the classification problem, as shown in Eq. (4.1). Usually expressed as a percentage, it indicates the proportion of all samples that the model correctly classified. Study conducted a comparative study using related

methods. As shown in Table 4.5, the H-CNN means the network combined convolutional neural network and Hilbert curve. Bold indicates the data with the highest accuracy for each method. The tabulated results unmistakably reveal a significant enhancement in accuracy attributed to the utilization of the images generated by the Hilbert curve.

$$\text{Accuracy} = \frac{\text{Number of Correctly Classified Samples}}{\text{Total Number of Samples}} \times 100\% \tag{4.1}$$

Table 4.5. Accuracy comparison of different 1D data to 2D image conversion methods(%).

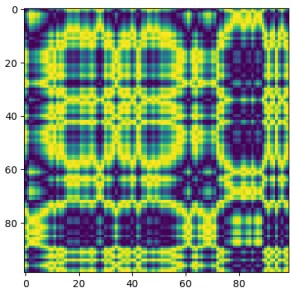
| **Algorithms** | **Network** | | | | | |
|---|---|---|---|---|---|---|
| | **Lr** | **Resnet18** | **Resnet34** | **Resnet50** | **Resnet153** | **H-CNN** |
| GAF | 1e-03 | 67.47 | 73.02 | 63.25 | 73.10 | - |
| | 1e-04 | **90.32** | 40.83 | 80.01 | 87.51 | |
| | 1e-05 | 90.00 | 34.58 | 81.32 | 90.01 | |
| | 1e-06 | 83.30 | 87.31 | 81.03 | 87.31 | |
| | 1e-07 | 74.37 | 92.38 | 81.73 | 82.13 | |
| CyTex | 1e-03 | 64.11 | 65.74 | 67.37 | 68.37 | - |
| | 1e-04 | 80.31 | 82.31 | 67.66 | 81.76 | |
| | 1e-05 | 82.94 | 84.53 | 85.34 | 86.31 | |
| | 1e-06 | 87.03 | 87.24 | **87.42** | 79.82 | |
| | 1e-07 | 75.22 | 73.01 | 70.28 | 73.50 | |
| Hilbert | 1e-03 | 73.58 | 76.29 | 80.14 | 80.01 | 81.91 |
| | 1e-04 | 80.09 | 81.14 | 81.79 | 80.13 | 84.08 |
| | 1e-05 | 86.17 | 88.06 | 89.37 | 87.16 | 90.13 |
| | 1e-06 | 94.16 | 95.21 | 95.19 | 92.13 | **95.95** |
| | 1e-07 | 82.93 | 84.19 | 88.16 | 87.97 | 89.01 |
| Hilbert(0-filled) | 1e-03 | 76.35 | 82.43 | 83.39 | 73.12 | 88.01 |
| | 1e-04 | 73.21 | 81.52 | 82.09 | 73.14 | 88.79 |
| | 1e-05 | 85.12 | 92.31 | 93.35 | 94.04 | **98.73** |
| | 1e-06 | 92.15 | 94.46 | 95.09 | 96.09 | 97.50 |
| | 1e-07 | 83.24 | 82.76 | 85.31 | 83.19 | 83.27 |

As shown in Table 4.6, compared to emotion speech recognition methods on the same dataset, the proposed approach in this study exhibits significant advantages. Bold indicates the most accurate data. Based on the comparative analysis, this study proposed emotional speech recognition method performs well on the same dataset and shows significant advantages. The higher accuracy level reflects the robustness and effectiveness of proposed method in capturing complex emotional nuances in speech signals.

Table 4.6. Accuracy comparison of same dataset methods(%).

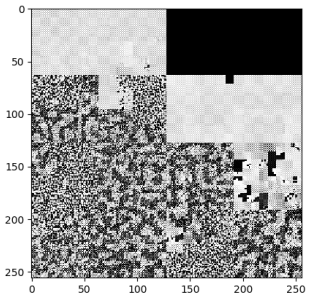
| Method | Description | Accuracy |
|---|---|---|
| DBN & SVM [131] | Deep Belief Network combined with Support Vector Machine for feature extraction and classification. | 95.80% |
| ELM decision tree [202] | Extreme Learning Machine decision tree approach for emotion recognition in speech signals. | 89.60% |
| GAF | GAF method transforming one-dimensional speech signals into two-dimensional images. | 95.95% |
| CyTex | CyTex method utilizing the inherent periodicity of speech signals for conversion to two-dimensional images. | 87.42% |
| Hilbert Curve | Innovative Hilbert Curve-based method mapping one-dimensional time series to two-dimensional images for emotion recognition. | **98.73%** |

First, this study introduced methods similar to those in this study, including GAF and CyTex, which also transform one-dimensional data into two-dimensional images. The 2D images generated by these methods exhibited different characteristics, as shown in Fig. 4.9(a), which represents a part of the image of the GAF method. Fig. 4.9(b) shows the image generated by the CyTex algorithm. Fig. 4.9(c), represents the image obtained using the Hilbert curve method, and the image is zero-filled, while Fig. 4.9(d) is transformed using the Hilbert curve but not zero-filled.The advantages and disadvantages of the three algorithms are shown in the Table 4.7. The GAF algorithm required significant memory resources. For a time series of length $L$, the GAF algorithm generates an image of size $(L, L)$, whereas CyTex generates an image of a size greater than *(C, L/C)*, where $C$ is the period. Hilbert generates an image of size $(2^n, 2^n)$, where $n$ is the smallest integer that satisfing the condition that $2^{2n}$ is greater than or equal to $L$. The following is an example of an integer that satisfies this requirement: The image generated using the GAF algorithm retains the temporal correlation; hence, inverse mapping can be realized. Whereas the image generated by the Cytex algorithm retains a portion of the temporal correlation within each cycle, the temporal correlation between cycles is lost due to 0-filling; hence, inverse mapping cannot be fully realized either. The image generated by the Hilbert algorithm also retains the temporal correlation and performs better in terms of accuracy than the other two algorithms. Thus, it can be concluded that the Hilbert curve is theoretically feasible.

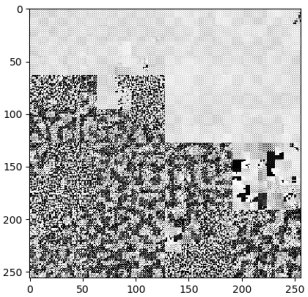


(a) GAF algorithm imaging.

(b) CytTex algorithm imaging.



(c) Hilbert imaging algorithm with 0 padding.



(d) Hilbert imaging algorithm with iterative padding.

Fig. 4.9. The result of Hilbert imaging algorithm.

Table 4.7. Comparison of GAF, Cytex, and Hilbert.

| Method | Size[1] | Temporal Relationship | Inverse Mapping | Accuracy[2] |
|---|---|---|---|---|
| GAF | *** | Preserve | Feasible | 2 |
| Cytex | ** | Partially Preserve | Partially Feasible | 3 |
| Hilbert | * | Preserve | Feasible | 1 |

[1] Size: Use * to indicate the size, the more * the larger the size.

[2] Accuracy: Positive ranking of accuracy expressed as a number, with smaller numbers indicating higher accuracy.

By comparing these methods, some important conclusions can be drawn. In this study, accuracy was used as the main evaluation metric and was compared with the results of other studies. The results show that the algorithm proposed in this study performs better in the same environment, particularly when the Hilbert method is used. In addition, from the perspective of computational cost, this study is more efficient than the two methods mentioned above, which provides more advantages for practical application.

Overall, the method proposed in this study provides a more promising way to process one-dimensional time series data. By presenting these data as 2D images, this study not only increased the richness of the information but also improved the accuracy and performance. The successful application of this method highlights the potential value of deep learning and computer vision in emotion recognition and other fields, thereby providing new directions for future research and applications.

## 4.5 Conclusions

Our research aims to explore how one-dimensional time-series data can be transformed into a visual form to provide a clearer presentation of the data and more effective analysis methods. In this process, the study face challenges such as high memory requirements and computational complexity, and the study need to consider the efficiency of the algorithm and the accuracy of data visualization. The Hilbert curve approach excels in this context because it can transform 1D time-series data into rich 2D images. Our results show that sentiment recognition using the Hilbert curve method is significantly better than that of other methods, emphasizing the importance and practical application of this method in the field of time-series visualization. This study not only provides new ideas for the visualization of time-series data but also offers the possibility of performance improvement of deep learning models. At the same time, the study emphasize the complexity of the field of sentiment analysis as it must deal with a diversity of sound signals, which makes sentiment recognition more challenging. Finally, the study argue that this research also has a wide range of applications that can be extended to other domains, such as throat condition detection for speech signals and basic brain detection based on brainwave signals, thereby providing new opportunities for interdisciplinary research and applications.

# Chapter 5

# Applications of the proposed two-dimensionalization algorithm in other fields

Epilepsy is a neurological disorder that seriously affects patients' lives and health. The accurate identification of epilepsy species is essential for developing effective treatment and management plans. This study aimed to enhance the efficient recognition of epilepsy types. Our methodology combines the CyTex algorithm with multichannel parallel convolution and RNNs neural networks for comprehensive analysis and classification. This integrated approach yielded notable results, allowing for the accurate differentiation of diverse epileptic events and ultimately achieving a recognition accuracy of 76.84%. Although these results are promising, it is acknowledged that there is still potential for further improvement in accuracy. This study provides valuable insights into epilepsy recognition and lays the foundation for future research in medical diagnosis and disease classification, although it does not represent a significant breakthrough in accuracy.

## 5.1 Introduction

Epilepsy is a neurological disorder caused by an abnormal discharge in the brain [203, 204]. It is one of the most common neurological diseases in the world. By 2023, the World Health Organization (WHO) reported that epilepsy will affect approximately 50 million individuals worldwide [205]. Seizures face potential dangers, such as falls, drowning, traffic accidents, pregnancy complications, and mental health issues [206]. However, these dangers should not be disregarded. If the patients fall during the seizure, they can injure their head or spinal; if they have a seizure while swimming or bathing, they are at risk of accidental drowning; seizures can lead to loss of control over the body or even loss of consciousness, making it impossible for the person to control the vehicle and causing traffic

accidents; seizures during pregnancy can endanger the pregnant woman and the fetus, and anti-seizure medications increase the risk of physical defects in the fetus. The risk of people with epilepsy is more likely to experience depression, anxiety, or other mental health problems. The medications used to treat epilepsy can cause these mental health problems. Therefore, prompt and accurate diagnosis of epilepsy is crucial for patients [207]. Currently, the diagnosis of epilepsy entails a comprehensive acquisition of medical history and meticulous neurological assessments. Additionally, supplementary examinations such as neuroimaging and electroencephalography (EEG) are employed in the diagnostic process. The EEG identifies interictal (between seizures) and ictal (during seizures) epileptiform abnormalities. A neurologist investigated epileptiform abnormalities by visual examination of EEG data, providing valuable information about the type of epilepsy and its cause [208].

However, there are several problems with interpreting EEG signals through visual assessment. First, different EEG readers can generate different types of EEG data. Second, epileptic discharges are not constant; however, EEG captures only a snapshot of the brain activity. Therefore, interpreting EEG signals by visual assessment is time consuming and may even require a manual review of the hours or days of EEG data [209]. In addition, epileptic EEG signals are more chaotic and variable than are normal EEG signals. Although epileptiform abnormalities are always present, they are not observed on the scalp surface electrodes. All of these factors can make diagnosis very difficult. Hence, the use of computer-aided diagnosis (CAD) systems is crucial for achieving accurate, rapid, and objective diagnosis [208].

In 1908, Neminsky discovered action currents in the central nervous system of frogs and electrical fluctuations in the brain of dogs [210]. He was also the first to print an electroencephalogram. In 1934, Penfield founded the Institute of Neurology in Montreal to study brain function and to treat epilepsy [211]. In 1936, Harvard et al.. reported the occurrence of slow waves in patients with seizures. This has led to progress in the study of epileptic electroencephalograms (EEG) [212]. Adeli used wavelet transforms to analyze epileptic EEGs in 2003 [213]. Rajendra Acharya summarized EEG signal-based automated epilepsy detection techniques in 2013 [214]. In 2015, Oliver Faust summarized epilepsy diagnosis based on wavelets, nonlinear dynamics, and the chaos theory [215]. In the same year, Rajendra Acharya reviewed the application of entropy in the EEG-based automatic diagnosis of epilepsy [216].

EEG signals are nonlinear and nonsmooth [217]. This complexity makes EEG signals too intricate for an intuitive interpretation. We implemented the CyTex transform and recurrent neural networks (RNNs) to detect various types of epilepsy and overcome this challenge. The CyTex transform was used to rationalize EEG signals into an image representation, thus preserving both short- and long-term information in the time series [218]. This transformation allows for a more thorough capture of EEG signal features,

thereby providing more meaningful inputs for subsequent classification tasks. Based on this premise, we implemented RNNs to process the altered EEG images [219]. RNNs were used to process the EEG signals. RNNs can capture the temporal dependencies in sequential data and have inherent memory functions. Because of the quasi-periodic nature of EEG signals, we calculated the fundamental period and organized the data as time steps while extracting relevant input features. This processing helps classify epilepsy. By combining the CyTex transform and RNN neural network, we were able to better capture complex EEG signal features for the effective classification of different epilepsies. This approach combines signal processing techniques and deep learning to provide a powerful tool for research and medical diagnosis in the field of neuroscience. This method primarily aids healthcare providers in understanding and interpreting patient data and ultimately relies on the judgment and expertise of medical professionals for diagnosis.

The remainder of this paper is organized as follows. Section 5.2 briefly describes the dataset and CyTex algorithm. Section 5.3 describes the extraction of EEG features. Section 5.4 presents the experimental results and analysis. Finally, Section 5.5 concludes this section.

## 5.2 Related work

This section describes the basic concepts of electroencephalography (EEG), the manifestations of epilepsy on EEG, the EEG databases used and the network models employed. EEG is a technique for recording the electrical activity of the brain by placing electrodes on the scalp to detect the electrical activity of the cerebral cortex, thus reflecting the state and changes in brain function. Epilepsy is a common neurological disorder, and its manifestation on EEG usually includes different types of EEG waveforms, such as sharp and slow waves, which reflect abnormalities in the electrical activity of the brain during seizures. In this paper, a specific network model is used for seizure identification and classification. By learning and extracting features from EEG data and performing automatic recognition and classification of epileptic seizures. By introducing the basic concepts of EEG, epilepsy manifestation on EEG, data models and network models, this paper provides the necessary background and theoretical foundation for seizure detection and recognition methods in the subsequent chapters.

### 5.2.1 EEG

Brain-computer interfaces (BCIs) are also known as brain-machine interfaces (BMIs). This machine enables the human brain to communicate directly with external computers and machines. In general, BCIs include four main areas: invasive brain-computer interfaces, partially invasive brain-computer interfaces, non-invasive brain-computer interfaces, and synthetic telepathy [220, 221, 222]. the advantages and disadvantages as shown in Table

5.1.

Table 5.1. Transposed BCI Interface Advantages and Disadvantages.

| **Interface** | **Advantages** | **Disadvantages** |
|---|---|---|
| Invasive BCIs | Precise signal acquisition. | Invasive surgery risks. |
| | High information transfer rate. | Potential tissue damage. |
| | Direct access to neural activity. | Ethical concerns. |
| Partially Invasive BCIs | Enhanced signal quality. | Still some surgical risks. |
| | Moderate information transfer rate. | Limited access to neural data. |
| | Reduced surgical risks. | Ethical considerations. |
| Non-Invasive BCIs | No surgery or physical intrusion. | Lower signal quality. |
| | Minimal health risks. | Slower information transfer. |
| | Ethical acceptance. | Limited access to neural activity. |
| Synthetic Telepathy | Potential for direct communication. | Highly speculative. |
| | Reduced physical involvement. | Technical challenges. |
| | Ethical and social implications. | Privacy concerns. |

Although there are many BCI systems and techniques, non-invasive BCI via EEG has been the most widely investigated. It is relatively inexpensive, easy to carry, and use. In addition, it has an elaborate temporal resolution [223]. In this study, we use the 10–20 system principle to read electrical signals from the scalp [224]. In this study, we explored and implemented a brain–computer interface (BCI)-based recognition of epilepsy types. By combining neuroscience and machine learning techniques, the information embedded in the EEG data can be utilized to develop an effective system that can accurately identify

different epileptic seizures.

### 5.2.2 Epilepsy

Seizures differ significantly on the EEG as shown in the Table 5.2, which describes several common patterns and their characteristics in the EEG. These patterns include sharp waves, sharp waves, sharp-slow wave complexes, sharp-slow wave complexes, polysharp wave complexes, polysharp-slow wave complexes, and sharp wave rhythms. Each pattern has unique features such as duration, frequency, and amplitude that can be used to accurately identify and diagnose seizures. These features show significant changes on the EEG during seizures, providing an important diagnostic basis for medical professionals, which helps them to develop individualized treatment plans and to monitor and manage patients effectively.

### 5.2.3 Dataset

This study used the Temple University Hospital (TUH) open-source database [225]. This database contains the basic statistics for the TUH EEG seizure corpus (TUSZ). The annotation file in TUSZ contained 13 different types of labels, as shown in Table 5.3, eleven specific seizure labels are used in the multiclass annotations. In bi-class annotations, the specific seizure is not annotated, only whether a seizure has occurred. This is referred to as a seizure (SEIZ). Only non-seizure annotation within the TUSZ background (BCKG) was used to identify the background. Therefore, the annotation files available within the TUSZ contained 13 different labels that consisted of seizure events and background annotations. In the two-category annotation, no specific seizures were annotated, only whether they occurred. This was labeled as seizure (SEIZ). BCKG was the only non-epileptic annotation in the TUSZ. In this study, we use the *train* section of the TUSZ dataset. It contains 4,664 brainwave data files stored in *.edf* format. The EEG data in each file were collected through 33 channels, with a sampling frequency of 250 Hz. The details of the data are presented in Table 5.4. In these data files, the channel numbers were labeled according to the standard ACNS TCP montage definition for channel numbering. For example, Channel 1 was obtained by calculating the voltage difference between electrodes F7 and T3, denoted as (F7-REF)–(T3-REF), as shown in Fig. 5.1 [225]. In this manner, we generated a data file of 22 channels for subsequent analysis. In this study, we accurately identified different epilepsy species by utilizing these data files, data preprocessing, feature extraction, and appropriate model construction for the pairs. In this study, 80% of the data were used as the training set, and 20% of the were used as the validation set.

Table 5.2. Description of EEG images during seizures

| **Waveform characteristics** | **Description** |
|---|---|
| Spike | The spikes are the most basic paroxysmal EEG activity, with a duration of 20 70 ms. Amplitude varies but is typically ¿50 uV (Kane et al., 2017). |
| Sharp | A sharp wave is similar to the spike, and its time limit is 70 200 ms (5 14 Hz). Amplitude is between 100 and 200 uV, and the phase is usually negative. |
| Spike and slow wave complex | An epileptiform pattern consisting of a spike and an associated slow wave following the spike, which can be clearly distinguished from the background activity; may be single or multiple (Kane et al., 2017). |
| Sharp and slow wave complex | An epileptiform pattern consisting of a sharp wave and an associated slow wave following the sharp wave, which can be clearly distinguished from the background activity; may be single or multiple (Kane et al., 2017). |
| Polyspike complex | A sequence of two or more spikes. |
| Polyspike and slow wave complex | An epileptiform pattern consisting of two or more spikes associated with one or more slow waves. |
| Spike rhythm | Refers to a widespread 10 25 Hz spike rhythm outbreak, with an amplitude of 100 200 uV and the highest voltage in the frontal area, lasting more than 1 s. |

### 5.2.4 Networks

The recurrent neural network is a type of neural network specialized for processing data sequences artificially [219]. This is suitable for tasks that involve sequential or time-related information. Unlike traditional feedforward neural networks, RNNs maintain an internal state and compute a new hidden state based on the input data of the current time step and hidden state of the previous time step. This mechanism allows the RNN to capture the context and patterns in a sequence, thereby enabling it to process sequence data of varying

Table 5.3. Types of Seizures and Abbreviations.

| **Seizure Type** | **Abbreviation** | **Description** |
|---|---|---|
| FNSZ | Focal nonspecific seizures | A large category of seizures occurring with specific focality. |
| GNSZ | Generalized seizures | A large category of seizures occurring in most, if not all, of the brain. |
| SPSZ | Simple partial seizures | Brief seizures that start in one location of the brain (and may spread) where the patient is fully aware and able to interact. |
| CPSZ | Complex partial seizures | Same as simple partial seizures but with impaired awareness. |
| ABSZ | Absence seizures | Brief, sudden seizure involving lapses in attention. It usually lasts for no more than 5 s and is commonly observed in children. |
| TNSZ | Tonic seizures | A seizure involving stiffening of the muscles. Usually associated with and annotated as tonic-clonic seizures, but not always (rarely, there is no clonic phase). |
| CNSZ | Clonic seizures | A seizure involving sustained rhythmic jerking. This is not seen in our datasets, as it is always associated with tonic-clonic seizures and is annotated as such. |
| TCSZ | Tonic-clonic seizures | A seizure involving loss of consciousness and violent muscle contractions. |
| ATSZ | Atonic seizures | A seizure involving loss of muscle tone in the body. It has never been observed, as it is always associated with an occasionally occurring phase before a tonic-clonic seizure. |
| MYSZ | Myoclonic seizures | A seizure associated with brief involuntary twitching or myoclonus. |
| NESZ | Nonepileptic seizures | Any non-epileptic seizure observed. No electrographic signs were observed. |

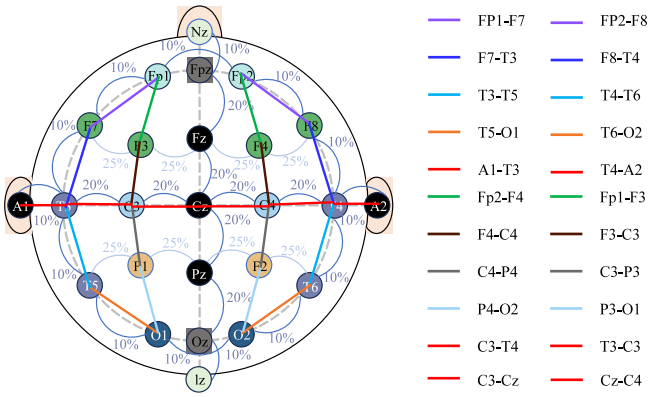



Fig. 5.1. Schematic of the 10-20 system numbered according to the ACNS TCP montage standard. A derivation is the voltage difference between electrodes: for example, Fp1-F3 is the voltage difference between electrodes Fp1 and F3.

Table 5.4. Basic descriptive statistics of the data.

| **Metric** | **Value** |
|---|---|
| Number of files (edf/csv/csv_bi) | 4,664 |
| Number of sessions | 1,175 |
| Number of patients | 579 |
| Number of files with seizures | 872 |
| Number of sessions with seizures | 352 |
| Number of patients with seizures | 208 |
| Total number of seizure events | 2,474 |
| Total duration | 3,277,229.00 secs |
| Total duration of files with seizures | 677,091.00 secs |
| Total background duration | 3,262,167.00 secs |
| Total seizure duration | 175,062.00 secs |

lengths. By continuously updating the hidden state at each time step, the information in the sequence is passed and accumulated to process the sequence data and capture temporal correlations in the sequence.

## 5.3 Experimentation 1 : Application of two-dimensionalization algorithms in epilepsy recognition

The raw EEG signal must be pre-processed to obtain relatively pure EEG data. Common EEG signal noise includes IF, ECG, EMG, and EEG noise [226]. The IF noise was mainly caused by the power supply of the device, and its frequency was 50 Hz. An ECG is generated by the rhythmic motion of the heart and has high amplitude. Because the heart is far from the head, the effect of the ECG signal on the EEG signal is typically ignored. EMG is generated by muscle contraction and its frequency is mainly concentrated in the high-frequency band at 100 Hz. EOG is an electrooculographic signal with a frequency between 0.1 and 100 Hz. The EOG noise is always measured when measuring prefrontal EEG signals. An EEG signal is generated by the heart, and its frequency is mainly in the high-frequency band of 100 Hz. The EEG spectrum has distinct waveforms in the four frequency bands in which seizures occur, as shown in Table 5.5 [227]. The clinical and physiological concerns range from 0.3 to 30 Hz. The frequency bands within this range are primarily categorized as $\delta$ (below 4 Hz), $\theta$ (4-8 Hz), $\alpha$ (8-13 Hz), and $\beta$ (13-30 Hz) [228]. Therefore, we pre-processed the EEG signals. The EEG signals were filtered using a bandpass filter of $0.3-30$ Hz. We then used independent component analysis to eliminate artifacts such as eye movements and EMG. With this well-established preprocessing framework, the accuracy of EEG signals is significantly improved, providing a solid foundation for subsequent in-depth analysis and research.

In this experiment, through data preprocessing, feature extraction, multi-channel convolution, and recurrent neural network operations, the original brainwave signals were transformed into images and classified, which realized accurate recognition and analysis for extracting useful information from complex brainwave data. The flow of the experiment as shown in the Fig. 5.2.

In this experiment, meticulous pre-processing was applied to the EEG signal data. Noise from EEG noise from head and neck muscle activity, ECG noise and power frequency mains noise was effectively eliminated by utilizing bandpass filtering techniques, leading to significant enhancement in data quality. To further minimize the impact of measurement noise, the differences between adjacent channels were calculated to improve data integrity. Subsequently, the data were segmented based on the start and stop times, as specified in *.csv* label files, ensuring alignment with the *.csv* file format, as shown in Table 5.6. The segmented data were then normalized to fit within the range (0–1). Linear transformation

Table 5.5. EEG Bands and Their Normal Manifestations

| **Band Name** | **Frequency Range** | **Normal Behavior** |
|---|---|---|
| Alpha Waves | 8-13 Hz | Normal adults exhibit alpha rhythms during relaxed and mentally inactive wakefulness. The amplitude is mostly below 50 $\mu$V and is most prominent in the occipital region. The alpha rhythm is blocked by eye opening (visual attention) and other mental activities, such as thinking. |
| Beta Waves | 13-30 Hz | Beta activity primarily observed in the anterior-central region with amplitudes smaller than the alpha rhythm. It increases during anticipation and tension states. |
| Theta Waves | 4-8 Hz | Theta frequency is present in normal infants and children, and during drowsiness and sleep in adults. Only a minimal theta rhythm is present in awake adults. A high theta activity in awake adults indicates abnormalities and pathological conditions. |
| Delta Waves | 0.5-4 Hz | Delta rhythm is slow brain activity that appears only during the deep sleep stage of normal adults. |

was applied to map the values to the interval (0, 255) to make the data suitable for an 8-bit image format.

The CyTex algorithm was employed to transform the EEG signals from each channel into images, with each image exclusively representing information from a single channel. Following the CyTex algorithm, data periodicity was determined using autocorrelation. One–dimensional time series data were transformed into two–dimensional data based on these periods. In this representation, the number of rows in the image corresponds to the number of periods, whereas the number of columns represents the length of each period. To create a dataset suitable for model training, the same filename data from the 22 channels were merged into one image, forming a multichannel image dataset. Subsequently, these image data were fed into a deep learning network to recognize different types of epileptic seizures. Image data processing was performed using multichannel parallel convolution.

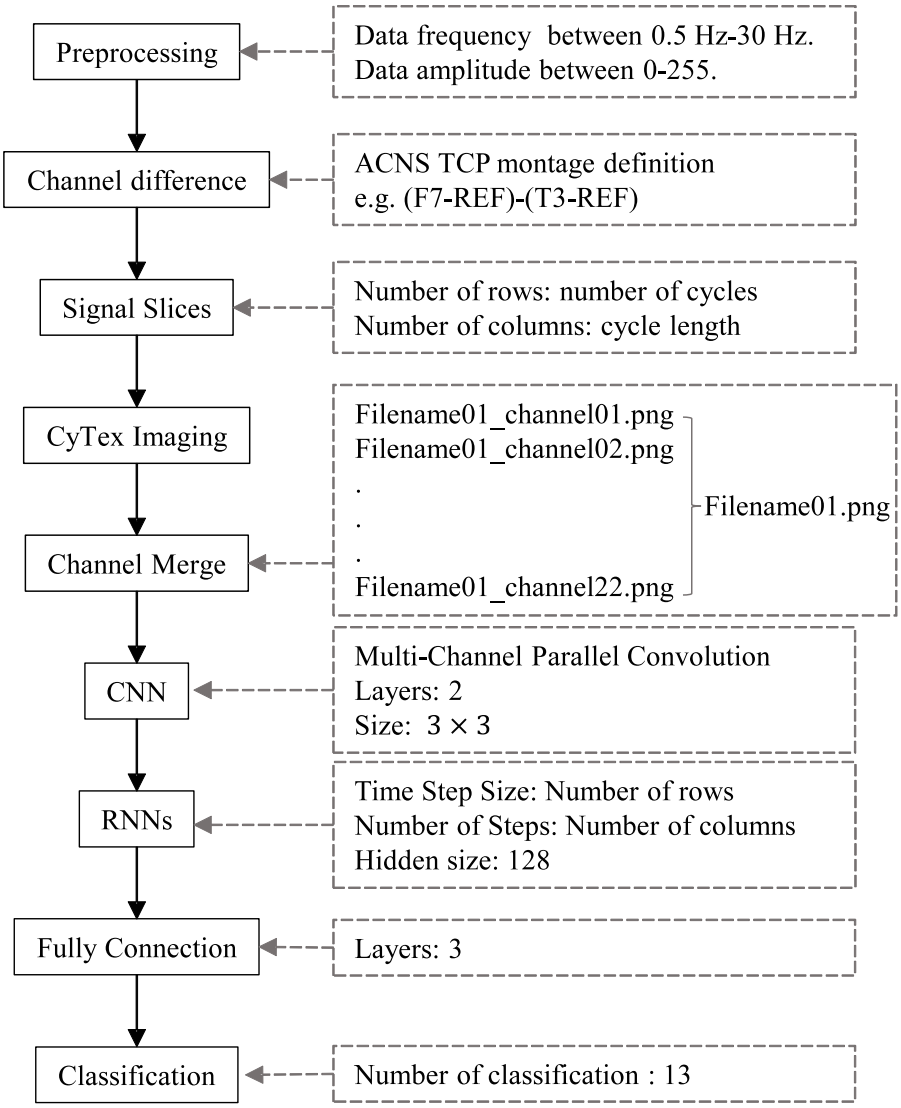



Fig. 5.2. Flow chart of the experiment.

The images were generated using the CyTex algorithm and named following the format *filename_channelnumber.png*. A dictionary was established to bundle 22 channel images with the same filename, resulting in channel–data fusion. Following this step, a data format was obtained with filename matching of 22 images as shown Fig. 5.3.

The number of rows in the images represents the number of cycles, and the number of columns represents the cycle length. To reduce the image dimensions, an image feature extraction process was performed using convolution. To ensure the integrity of each channel's data, a multichannel parallel convolution network was used, which was specifically designed for processing multi-channel data. The convolution operation reduces the size of the feature maps, extracts crucial features, and reduces the computational burden. After the convolution process, the results from each channel were fed into Recurrent Neural Networks (RNNs) to establish a temporal relationship with the features. The neural recursive

Table 5.6. Partial TUSZ database .CSV file

| Channel | Start time | Stop time | Label |
|---|---|---|---|
| FP1-F7 | 0.0000 | 36.8868 | bckg |
| FP1-F7 | 36.8868 | 183.3055 | cpsz |
| FP1-F7 | 183.3055 | 301.0000 | bckg |
| F7-T3 | 0.0000 | 36.8868 | bckg |
| F7-T3 | 36.8868 | 183.3055 | cpsz |
| F7-T3 | 183.3055 | 301.0000 | bckg |
| T3-T5 | 0.0000 | 36.8868 | bckg |

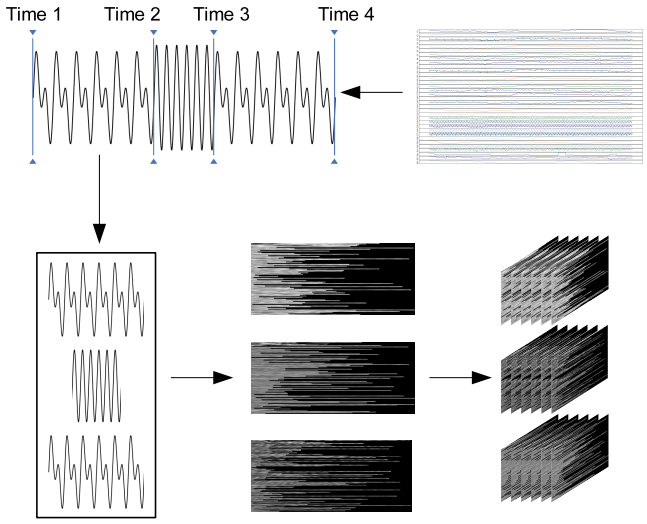



Fig. 5.3. EEG signal processing and conversion process.

network computes the feature values for each channel's data. Eventually, the data processed through the RNNs were transformed into one-dimensional vectors, with data from all 22 channels merged. After applying three fully connected layers and a Softmax activation function, the data were classified. The network architecture of this deep-learning model is depicted in the diagram. Through this sequence of processing steps, an accurate classification of epileptic seizure types was achieved. The network model for deep learning is shown in Fig. 5.4.

The design of this process allows us to extract useful information from raw brainwave signals and gradually transform them into data suitable for classification, taking full advantage of convolutional and recurrent neural networks. Using this approach, we can detect various patterns and characteristics in brainwave data, which supports research in

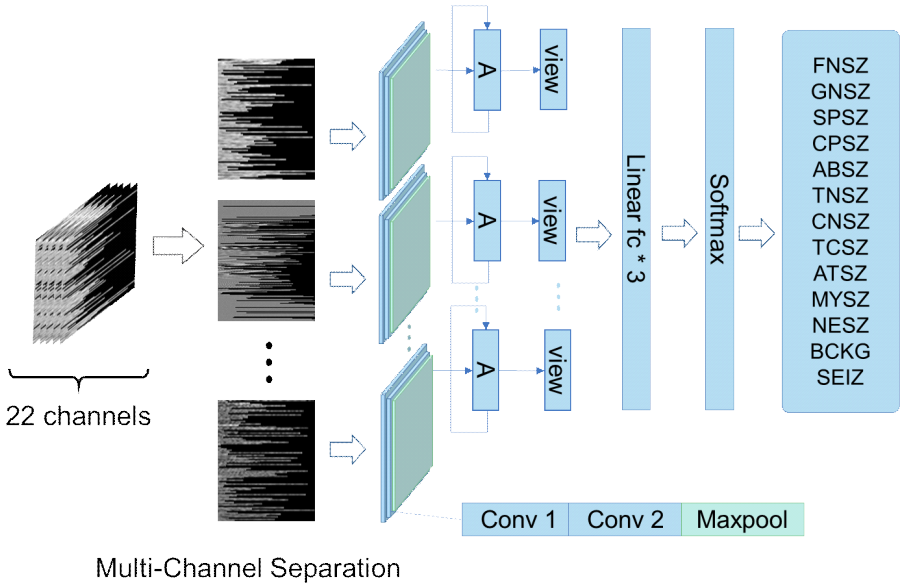



Fig. 5.4. Network model diagram for deep learning. The input data for the image had 22 channels, each with dimensions of height (H) and width (W). We maintained a single-channel output by performing a multichannel parallel convolution of the image for each channel, where each channel passed through two convolutional layers with a convolutional kernel size of 3 × 3. Next, we used a maximum pooling layer to reduce the size of the output image to half of its original size (H/2, W/2). We then utilized Recurrent Neural Networks (RNNs) to extract the feature values. The output is spread as a vector and merged with the results from the 22 channels. Finally, the classification task was accomplished using three fully connected layers and a softmax function. This processing flow can effectively extract and accurately classify key features from multichannel EEG images.

neuroscience and human-computer interaction. The software and parameter settings for the experiments are presented in Table 5.7.

### 5.3.1 Results and discussion

In this study, we employed a set of evaluation metrics to assess the performance of the proposed multiclassification model, which included four crucial metrics: TP, TN, FP, and FN. These metrics serve the following purposes.

1. **Precision**: precision measures the proportion of all samples classified into positive categories that are truly positive. For each category, precision was computed to determine the classification accuracy of the model. As shown in Eq. (5.1).

$$Precision = \frac{TP}{TP + FP} \tag{5.1}$$

Table 5.7. Experimental Setup.

| Experiment Description | Value/Parameter |
|---|---|
| Programming language and Vision | Python 3.11 and Pytorch |
| Operating System | Windows10 and Ubuntu 20.04 |
| Preprocessing process | Matlib R2013a and EEGlab |
| libraries | pandas, opencv, sklearn, itertools, torch |
| Batch size | 24 |
| Epochs | 100 |
| Hiddern size | 128 |
| Train ratio | 0.8 |
| Learning rate | 0.001 |
| Loss fuction | CrossEntropyLoss |
| Optimizer | Adam |

TP represents true positive cases (the number of samples correctly categorized as positive categories by the model) and FP represents false positive cases (the number of samples mistakenly categorized as positive categories by the model).

2. **Recall**: recall assesses the proportion of all truly positive category samples that were correctly categorized as positive. Recall was used for each category to gauge how effectively the model captured the positive categories. Recall is computed as shown in Eq. (5.2).

$$Recall = \frac{TP}{TP + FN} \tag{5.2}$$

Where TP represents true positive cases, and FN represents false negative cases (the number of samples misclassified as negative categories by the model).

3. **F1-score (F1 score)**: F1 score serves as a composite performance metric that balances precision and recall. It evaluates the performance of the model across different categories considering both classification accuracy and coverage. The F1-score was calculated as shown in Eq. (5.3).

$$F1\ score = \frac{2 \cdot Precision \cdot Recall}{Precision + Recall} \tag{5.3}$$

4. **F1-score (Accuracy)**: Accuracy is used to measure the overall performance of a classification model, indicating the proportion of correctly categorized samples, whether in positive or negative categories. Accuracy was calculated using Eq. (5.4).

$$F1\ score = \frac{TP + TN}{TP + TN + FP + FN} \tag{5.4}$$

In this equation, $TP$ stands for True Positive Examples, $TN$ is for True Negative Examples (the number of samples correctly categorized as negative), $FP$ represents False Positive Examples, and $FN$ denotes False Negative Examples.

These metrics collectively offer a comprehensive evaluation of the performance of the model for multiclass classification.

The classification results are presented in Table 5.9. The classification report provided precision, recall, F1 score, and number of supports for the samples. This provided a detailed analysis of the performance of each category. Precision is the proportion of samples that are actually positively categorized out of all samples predicted to be positively categorized by the model. Recall is the proportion of samples that the model successfully predicts as positive categories out of all samples that are actually positive. The F1 score is a weighted average of the precision and recall that balances the trade-off between precision and recall. It indicates the number of actual samples in each category. The *micro Avg* is the result of summarizing the performance of all categories. The support indicates the number of samples used in the experiment. In Table 5.8, the support value of 2533 indicates that 2533 data points were involved in this evaluation, and the same sample and sample size were used for the three different evaluation methods.The *Macro Avg* is the result obtained by averaging the performance metrics (Precision, Recall, F1-score) for each category. The *Weighted Avg* is the result obtained using a weighted average of the performance metrics for each category, where the weights are based on the number of supports in each category. As shown in Table 5.9, compared with existing epilepsy-type recognition methods [229], the accuracy rate was used as the evaluation metric. Accuracy is the number of samples correctly categorized by the model as a percentage of the total number of samples.

We conducted a comparative analysis of several classification studies based on the RNN and CNN algorithms using the same database. For a more comprehensive comparison, the histograms in Fig. 5.5. In the graph, distinct colors were used to represent different studies. The y-axis represents the various networks used in the different studies. Fig. 5.5(a) displays the results based on accuracy as an evaluation metric, the x-axis represents accuracy, indicated as a percentage. While Fig. 5.5(b) illustrates results based on the F1 score as an evaluation metric, the y-axis represents the F1 score with a maximum value of 1. Within the images, instances without color filling indicated that the respective authors did not provide data for that particular evaluation metric. Among the notable findings, Shankar et al. [230] employed the CNN algorithm for three-, four-, and five-

Table 5.8. Classification report of the experiment.

| | **Precision** | **Recall** | **F1 score** |
|---|---|---|---|
| Micro Avg | 0.77 | 0.77 | 0.77 |
| Macro Avg | 0.06 | 0.08 | 0.07 |
| Weighted Avg | 0.59 | 0.77 | 0.67 |

class classifications, achieving accuracy rates of 89.91%, 84.19%, and 84.20%, respectively. Similarly, Thundiyil et al. [231] utilized CC images, MI images, and Stacked images for epilepsy species recognition based on the Resnet18 network. Their results revealed accuracy and F1 scores of 93.45%, 0.936; 97.89%, 0.98; 95.50%, 0.956.

The experimental results showed that the model achieved 76.84% accuracy in an epilepsy category recognition task. Despite the fair performance in terms of accuracy, there is a need for further improvements in the model performance. This suggests that the new model can recognize different types of epileptic events. However, further optimization is required to provide higher reliability and stronger support for diagnosis and treatment in the medical field.

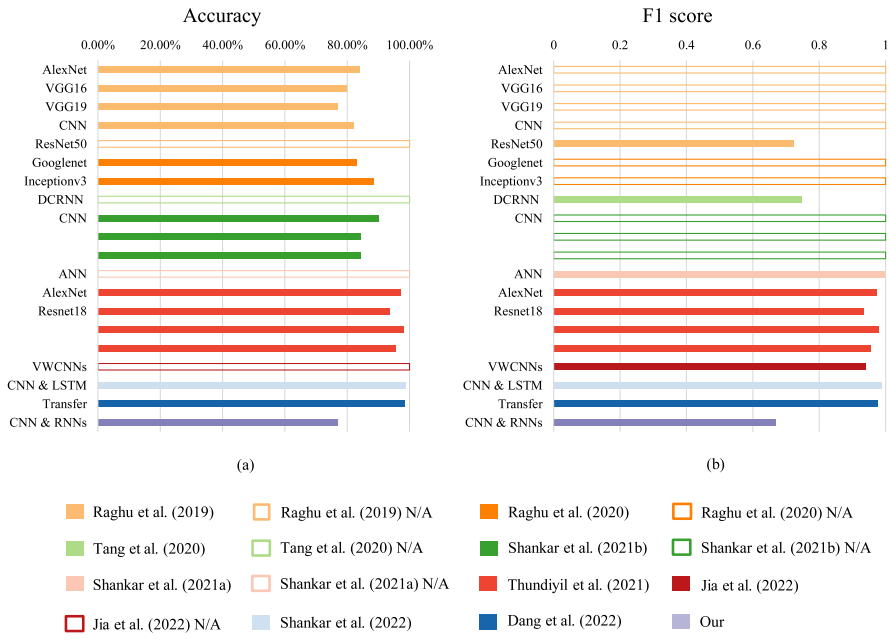



Fig. 5.5. Histogram of the results of epilepsy species identification based on the TUSZ database.

Table 5.9. Comparison of epilepsy species identification results based on TUSZ database.

| **Publications** | **Classifier** | **Accuracy** | **F1 score** |
|---|---|---|---|
| Raghu et al. (2019) [232] | AlexNet | 84.06% | N/A |
| | VGG16 | 79.71% | N/A |
| | VGG19 | 76.81% | N/A |
| | CNN | 82.1% | N/A |
| | ResNet50 | N/A | 0.722 |
| Raghu et al. (2020) [233] | Googlenet | 82.85% | N/A |
| | Inceptionv3 | 88.30% | N/A |
| Tang et al. (2021) [234] | DCRNN | N/A | 0.749 |
| Shankar et al. (2021b) [235] | CNN | 89.91% | N/A |
| | | 84.19% | N/A |
| | | 84.20% | N/A |
| Shankar et al. (2021a) [230] | ANN | N/A | 0.997 |
| Thundiyil et al. (2021) [236] | AlexNet | 97.15% | 0.975 |
| | Resnet18 | 93.45% | 0.936 |
| | | 97.89% | 0.98 |
| | | 95.50% | 0.956 |
| Jia et al. (2022) [231] | VWCNNs | N/A | 0.94 |
| Shankar et al. (2022) [237] | CNN & LSTM | 98.82% | 0.988 |
| Dang et al. (2022) [238] | Transfer | 98.48% | 0.976 |
| Our | CNN & RNNs | 76.84% | 0.67 |

## 5.4 Experimentation 2 : Improved Epilepsy Recognition Experiment

The method of obtaining the period depends on the type and characteristics of the data being processed. The following are some commonly used methods for obtaining the period of a signal.

- **Fourier Transform**: Used to convert a time domain signal into a frequency domain signal. Periodic signals usually show clear periodic characteristics in the spectrum [239].
- **Autocorrelation Analysis**: Autocorrelation analysis is a method used to determine the periodicity of a signal. It involves calculating the correlation between a signal and itself at different time lags. When a signal is periodic, the autocorrelation function will show clear periodic peaks. With autocorrelation analysis, you can estimate the major period of a signal. By calculating the correlation at different lag orders in a time series, you can find recurring patterns. In periodic data, ACF usually shows high autocorrelation for specific lags [240].
- **Moving Window Analysis (MWA)**: Using sliding windows to look at local features of the data and identify recurring patterns in the data [241].
- **Lomb-Scargle Periodogram Analysis**: Lomb-Scargle Periodogram is a method used to analyze the periodicity of non-uniform time series. It is a frequency domain analysis method that takes into account measurement intervals at different points in time. Lomb-Scargle periodograms can help you estimate the major periodic components of a signal [242].

The advantages and disadvantages are shown in the Table 5.10.

The experiments were conducted using the *Train* section of the TUSZ database. The specific description and data content of the database is shown in Table 5.11. It contains EEGs of 8 different epileptic events from 579 patients (men, women, and children), which is rich and sufficient data. The procedure of the experiment is shown in Figure 5.6. The experiment reconstructs the morphology of the data by segmenting the preprocessed EEG data according to the period feed line. The reconstructed data is fed into a neural network to learn, which enables the recognition of epilepsy species. The configuration of the experiment is shown in the Table 5.12.

### 5.4.1 Preprocessing

Preprocessing of EEG data based on the frequency characteristics of EEG waves is a common method. EEG waveforms have different symbols depending on the frequency band.

Table 5.10. Comparison of the advantages and disadvantages of signal period detection

| **Methods** | **Advantage** | **Disadvantage** |
|---|---|---|
| Fourier Transform | • Reveals frequency domain features<br>• Suitable for periodic signals | • Insensitive to non-periodic signals<br>• Requires signal stability |
| Autocorrelation Analysis | • Reveals signal periodicity<br>• Does not require signal model assumptions | • Sensitive to noise<br>• High computational complexity |
| Moving Window Analysis | • Suitable for detecting local features<br>• No need for prior assumptions | • Window selection affects results<br>• Not suitable for global periodicity |
| Lomb-Scargle Periodogram Analysis | • Suitable for non-uniform time series<br>• Considers time interval variations | • Sensitive to noise<br>• Requires parameter adjustments |

The main frequency bands of the brain waves obtained from the patient's measurements contain the states and pathologies that the patient himself is in [243]. The band symbols and states of the EEG are shown in the Table 5.13. Since the range of clinical and physiological interest is between 0.3 and 30 Hz [244], in this experiment, the EEG data were retained in the range of 0.5-30 Hz by means of a filter. In order to reduce the effects produced by outliers and noise in the experimental data [245], to increase the convergence speed of the model, and to prevent the ccurrence of overfitting phenomenon [246], the experiments performed regularization operations on the filtered EEG data. The EEG data amplitude was adjusted to the range of 0-255. The experiment used the longitudinal montage configuration of the International 10-20 system. The electrodes were arranged longitudinally to cover the prefrontal region to the parietal region, as shown in Fig. 5.7. The electrode pairs are shown in the Table 5.14. The distance between neighboring electrode pairs was kept consistent in the montage configuration and was selected according to the experimental purpose [247]. In this experiment, eight electrode pairs were selected ac-

Table 5.11. Number of documents per seizure category in the TUSZ training dataset

| Seizure Type | Decription | File Count |
|---|---|---|
| GNSZ | Generalized seizures. A large category of seizures occurring in most, if not all, of the brain. | 203 |
| FNSZ | Focal nonspecific seizures. A large category of seizures occurring with specific focality. | 516 |
| SPSZ | Simple partial seizures. Brief seizures that start in one location of the brain (and may spread) where the patient is fully aware and able to interact. | 1 |
| CPSZ | Complex partial seizures. Same as simple partial seizures but with impaired awareness. | 26 |
| MYSZ | Myoclonic seizures. A seizure associated with brief involuntary twitching or myoclonus. | 5 |
| ABSZ | Absence seizures. Brief, sudden seizure involving lapses in attention. It usually lasts for no more than 5 s and is commonly observed in children. | 56 |
| TNSZ | Tonic seizures. A seizure involving stiffening of the muscles. Usually associated with and annotated as tonic-clonic seizures, but not always (rarely, there is no clonic phase). | 11 |
| TCSZ | Tonic-clonic seizures. A seizure involving loss of consciousness and violent muscle contractions. | 37 |
| BCKG | Background. non-seizure annotation within the TUSZ background (BCKG) was used to identify the background. | 1784 |

cording to the region of seizure [246], namely, FP1-F7,FP1-F3,F7-T3,T3-T5,FP2-F8,FP2-F4,F8-T4,T4-T6, which covered prefrontal and temporal lobe regions [247], as shown in Fig. 5.8.

Table 5.12. Programming Environment Settings

| Setting | Value |
|---|---|
| Operating System | Ubuntu/Windows 10 |
| Programming Language | Python 3.11 |
| Deep Learning Framework | PyTorch |
| Libraries | csv, matplotlib, numpy, pandas, pyedflib, scipy, sklearn, torch |

Table 5.13. EEG Bands and Their Normal Manifestations

| **Band Name** | **Frequency Range** | **Normal Behavior** |
|---|---|---|
| Alpha Waves | 8-13 Hz | Normal adults exhibit alpha rhythms during relaxed and mentally inactive wakefulness. The amplitude is mostly below 50 $\mu$V and is most prominent in the occipital region. The alpha rhythm is blocked by eye opening (visual attention) and other mental activities, such as thinking. |
| Beta Waves | 13-30 Hz | Beta activity primarily observed in the anterior-central region with amplitudes smaller than the alpha rhythm. It increases during anticipation and tension states. |
| Theta Waves | 4-8 Hz | Theta frequency is present in normal infants and children, and during drowsiness and sleep in adults. Only a minimal theta rhythm is present in awake adults. A high theta activity in awake adults indicates abnormalities and pathological conditions. |
| Delta Waves | 0.5-4 Hz | Delta rhythm is slow brain activity that appears only during the deep sleep stage of normal adults. |

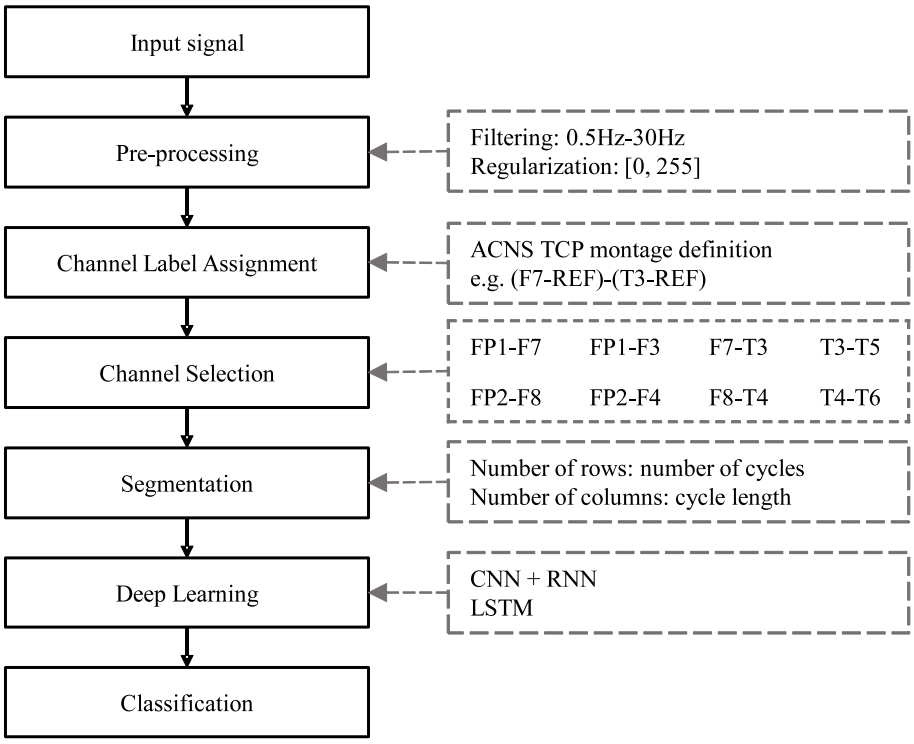



Fig. 5.6. Flow chart of the experimental process. The input EEG signal is first passed through a filter that retains the frequency range of 0.5-30 Hz band. Through regularization, it is compressed to between 0-255, which conforms to the 8-bit map standard. The pre-processed signals were used to calculate the signal difference at the corresponding location according to the ACNS TCP montage standard. Appropriate electrodes are selected according to the site of the seizure and the signal is reconstructed using the cycle recognition method. The reconstructed signal is transformed from a one-dimensional time series of length $m \times n$ into a two-dimensional array of (m,n). The number of rows of the array $m$ denotes the number of cycles; the number of columns of the array $n$ denotes the length of each cycle. The two-dimensional array is put into the deep learning neural network as an input signal to learn and finally complete the classification to realize the recognition of epilepsy species.

### 5.4.2 Segmentation by period

Research [248] pointed out that the periodicity of the data can be preserved by segmenting the time series signal according to the period. The preprocessed and filtered matched brain signal data is segmented by period, and the principle of segmentation is shown in Figure 5.9. The reconstructed data is learned using multi-channel parallel convolution and LSTM neural network respectively. There are two reasons for segmenting brainwave signals by period. One is because brainwaves are quasi-periodic signals, and the other is that in epileptic seizures, the waveforms of the brainwaves have typical characteristic

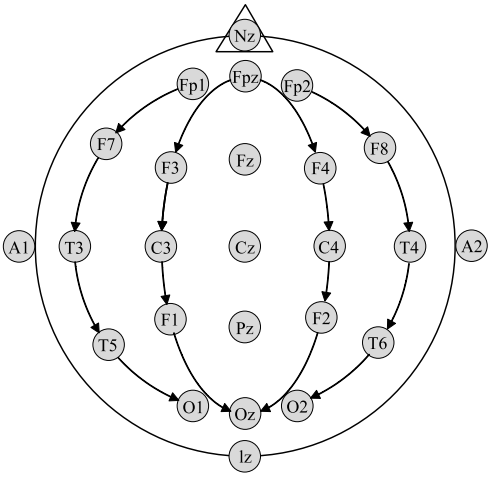



Fig. 5.7. Vertical standard montage schematic.

Table 5.14. EEG Electrode Layout

| | | | |
|---|---|---|---|
| Fp1-F7 | Fp1-F3 | Fp2-F4 | Fp2-F8 |
| F7-T3 | F3-C3 | F4-C4 | F8-T4 |
| T3-T5 | C3-F1 | C4-F2 | T4-T6 |
| T5-O1 | F1-O1 | F2-O2 | T6-O2 |
| A1-T3 | T3-C3 | C4-T4 | T4-A2 |
| | C3-Cz | Cz-C4 | |

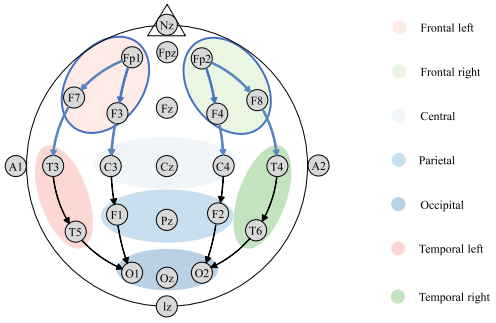



Fig. 5.8. Electrode pairs and coverage locations relevant to epilepsy.

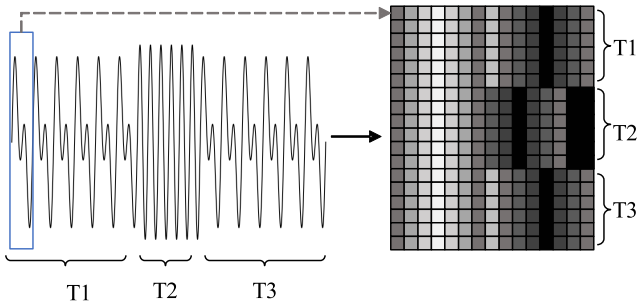



Fig. 5.9. Cycle methods

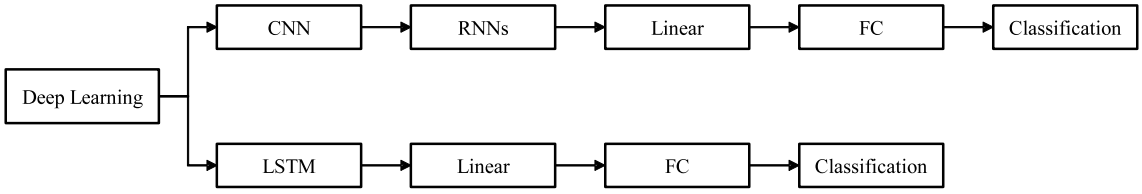



Fig. 5.10. DEEP CHART

patterns[249], as shown in the Table 5.15. Based on the quasi-periodic nature of brainwave signals, different methods (fft, autocorrelation, sliding window) are utilized to identify the period of the signals. The recognition results are shown in Fig. The signal is segmented according to the obtained period and the morphology of the data is reconstructed. In this study, the segmentation of brainwave data is based on 8Hz as the period corresponding to the frequency. 8Hz frequency is the minimum frequency of the alpha band [248], and its corresponding period is the maximum period of the band, as shown in Table. Thus it is a multiple of the period of the other bands, which preserves the periodicity of the period of the other bands. The data reconstructed according to the period is put into the neural network. Two different neural network methods are used for learning in this study. Multi-channel parallel convolution and RNN and LSTM neural network respectively, as shown in Fig. 5.10.

In order to preserve the temporal information in the brainwave signals and better understand and utilize the temporal dependencies in the signals, the experiments also used the neural network of LSTM to recognize the reconstructed brainwave signals. As shown in Fig. 5.11, the reconstructed brainwave signals were arranged in cycles. Each cycle was put into the LSTM network model as a time step.

## 5.5 Results and discussion

In previous studies, there are various methods for epilepsy species identification based on TUSZ database. We also worked on epilepsy species identification and made new contributions in this field. The brainwave data are segmented according to cycles, and each cycle

Table 5.15. Description of EEG Patterns

| **Pattern** | **Description** |
|---|---|
| Spike | The spikes are the most basic paroxysmal EEG activity, with a duration of 20–70 ms. Amplitude varies but is typically more than 50uV (Kane et al., 2017). |
| Sharp | A sharp wave is similar to the spike, and its time limit is 70–200 ms (5–14 Hz). Amplitude is between 100 and 200 uV, and the phase is usually negative. |
| Spike and slow wave complex | An epileptiform pattern consisting of a spike and an associated slow wave following the spike, which can be clearly distinguished from the background activity; may be single or multiple (Kane et al., 2017). |
| Sharp and slow wave complex | An epileptiform pattern consisting of a sharp wave and an associated slow wave following the sharp wave, which can be clearly distinguished from the background activity; may be single or multiple (Kane et al., 2017). |
| Polyspike complex | A sequence of two or more spikes. |
| Polyspike and slow wave complex | An epileptiform pattern consisting of two or more spikes associated with one or more slow waves. |
| Spike rhythm | Refers to a widespread 10–25 Hz spike rhythm outbreak, with an amplitude of 100–200 uV and the highest voltage in the frontal area, lasting more than 1s. |

preserves the temporal correlation of the brainwave signal. The commonly used methods for period recognition are Fourier transform, autocorrelation, moving windowing and lomb-scargle. As shown in the Fig. 5.12, experiments have been conducted to validate the ability of the above four algorithms to recognize the periods using multi-frequency filtering and multi-frequency superimposed waves. As can be seen from the figure, the four algorithms show different results in terms of period delineation. In Fig. 5.12(a) and Fig. 5.12(b), the identification of cycles using the Fourier variation analysis method is relatively poor, and the cycles cannot be accurately identified either for simple multifrequency waves or complex multifrequency superposition waves. Fig. 5.12(c) shows that for simple multifrequency waves, the autocorrelation function analysis method is more effective in identifying the period, on average, half of the period can be identified and segmented more

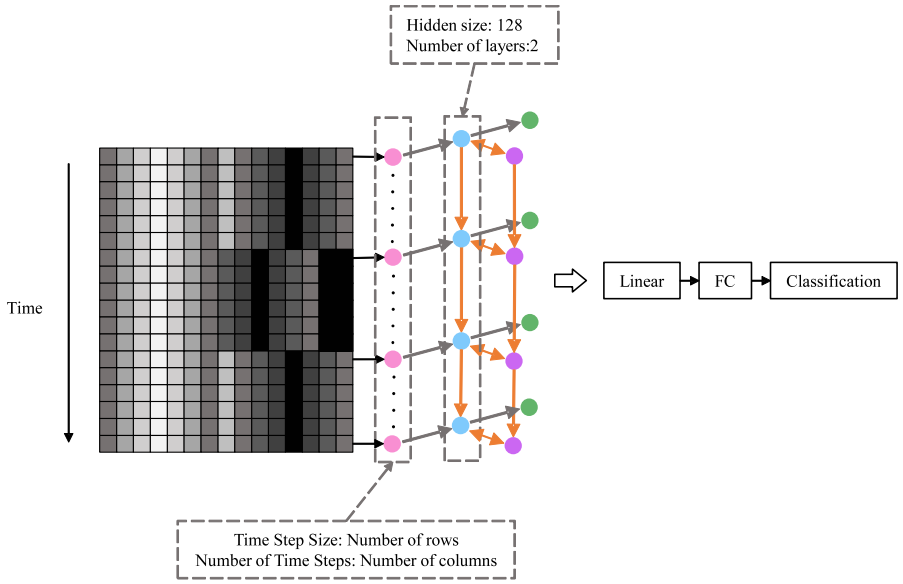
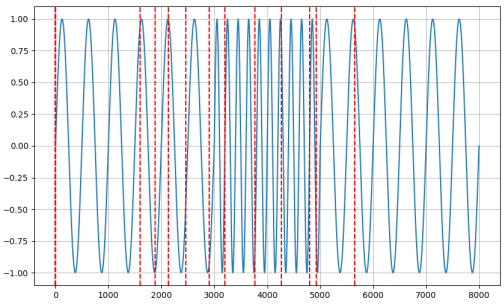


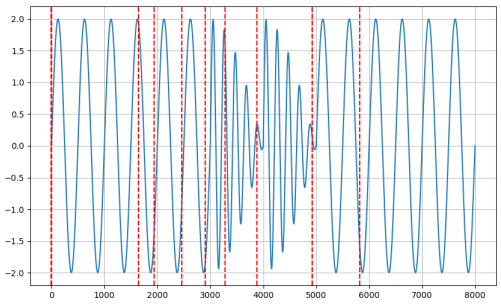

Fig. 5.11. Schematic diagram of the LSTM network. Each row of the reconstructed data represents one cycle. A cycle is put into the LSTM neural network as a time step. The number of steps is equal to the number of cycles.The LSTM contains 128 hidden layers, two LSTM operations are performed, the output feature values are expanded and tiled in chronological order, and the resulting one-dimensional array is operated by the fully connected layers to complete the recognition task. The final realization of epilepsy species recognition

accurately; however, for complex multifrequency superimposed waves Fig. 5.12(d), the autocorrelation function analysis method is generally effective in identifying the period, and the superimposed frequency cannot be identified completely. Compared with Fig. 5.12(c), Fig. 5.12(e) shows that the period identification of the sliding window analysis method is much finer, and on average, it can identify 1/4 period, but it also fails to accurately identify the complex multi-frequency superimposed waves, as in Fig. 5.12(f). While Fig. 5.12(g) and Fig. 5.12(h) indicate that the lgmb-scargle analysis method performs better in cycle identification and the results are more specific for the two waveforms. Periods of different frequencies are not recognized, but at the same time a complete cycle is not destroyed.

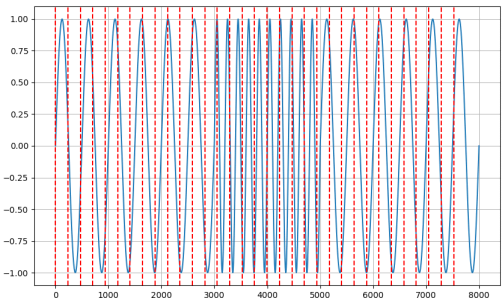
However, several methods are not ideal for segmentation of cycles. Therefore, the experiment was conducted to slice the brainwave signal based on the epileptic wave's performance on the alpha band (8Hz-13Hz) by calculating the brainwave period based on the frequency 8Hz. The experiments were compared with previous recognition results using accuracy rate and F1 score as the main evaluation indexes, as shown in the Table 5.16. Accuracy is the ratio of the number of samples correctly predicted by the model to the total number of samples in a classification problem. It is an important measure of the performance of a classification model and is usually expressed as a percentage. The F1 score



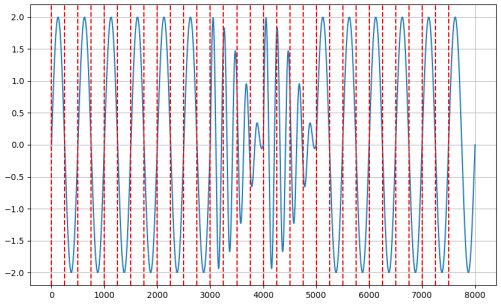
(a) Fourier Transform - Multi Frequency Periodic Waveform



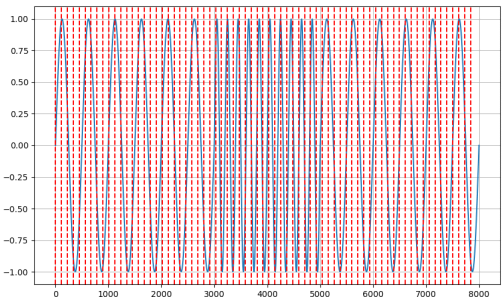
(b) Fourier Transform - Multi Frequency Superposition Periodic Waveform



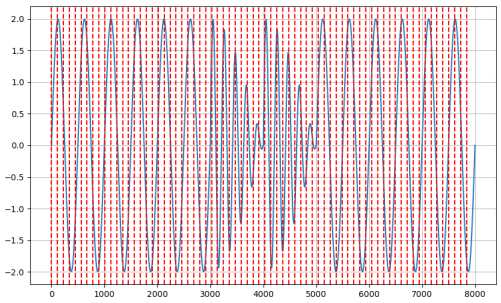
(c) Autocorrelation Function-Multiple Frequency Cycle Waveform



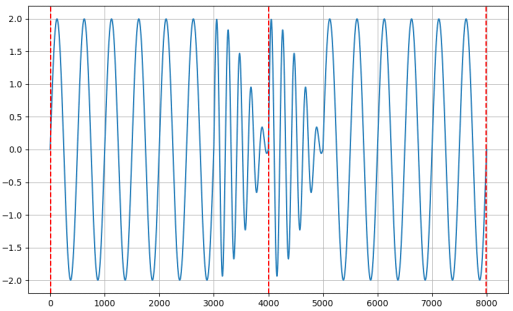
(d) Autocorrelation Function-Multiple Frequency Superposition Cycle Waveform



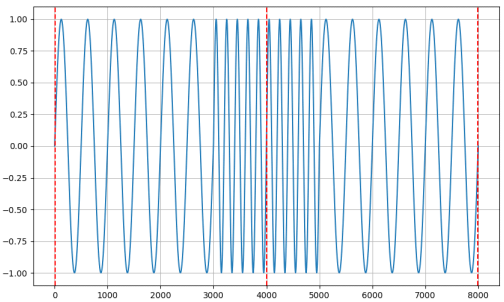
(e) Sliding Window Analysis-Multiple Frequency Periodogram



(f) Sliding Window Analysis-Multiple Frequency Superposition Periodogram



(g) Lomb Scargle Periodogram - Multi-Frequency Periodogram



(h) Lomb Scargle Periodogram - Multi-Frequency Superposition Periodogram

Fig. 5.12. The results of the four cycle recognition methods are shown. The first column shows the recognition results of a simple multi-frequency period waveform graph and the second column shows the recognition results of a complex multi-frequency superimposed period waveform graph. The recognized cycles are divided by dashed lines.

is an evaluation metric that combines model Precision and Recall. The F1 score provides a more comprehensive performance metric when dealing with unbalanced datasets or with category skew. In order to show more intuitively, the experiment plotted the accuracy rate notation over as a histogram, as shown in Fig. 5.13. From the figure, it can be seen that the experimentally proposed method gives excellent results for the classification of epilepsy. The experiment used both ROC curve and confusion matrix for further evaluation using LSTM method for evaluation. As shown in the Fig. 5.14, it can be seen in the ROC curve Fig. 5.14(a) and the confusion matrix Fig. 5.14(b). A problem in the results was mentioned and quantitative language was used to narrow down its significance. However, these results show that more temporal information can be retained while obtaining higher accuracy using period segmentation and LSTM network modeling compared to traditional epilepsy recognition.

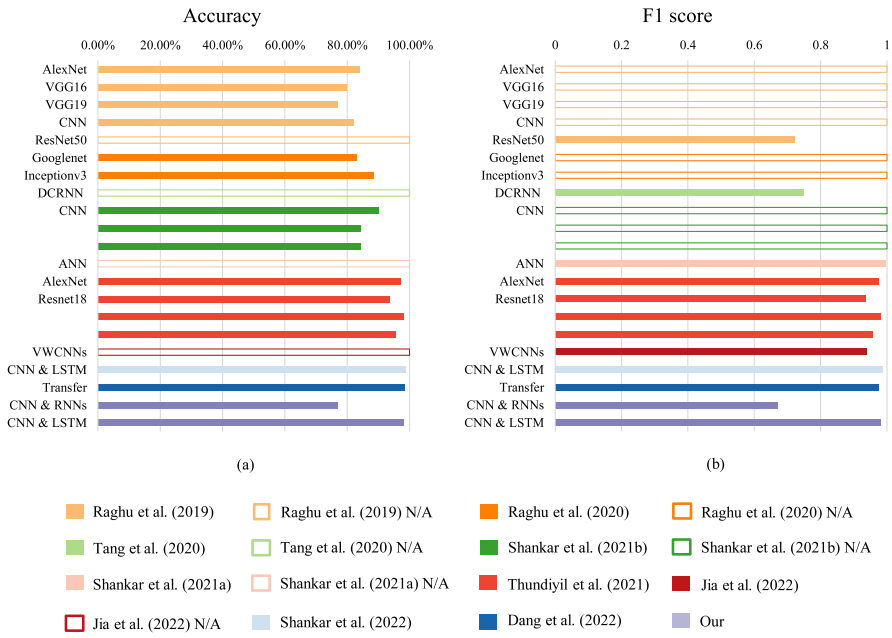
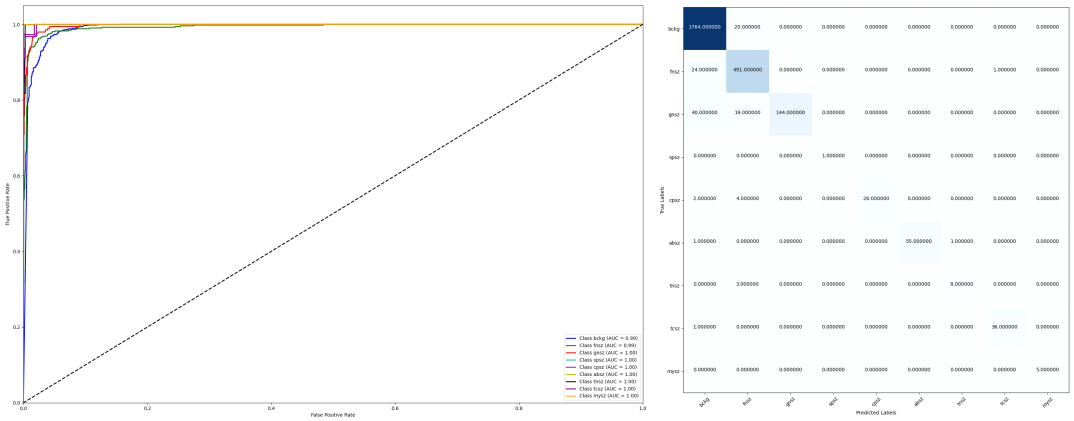



Fig. 5.13. Figure of the results of epilepsy species identification based on the TUSZ database.

In previous studies, some methods have focused primarily on analyzing brainwave signals in the frequency domain, for example, using the Fourier transform or wavelet transform. These methods are able to reveal the characteristics of the signal in the frequency domain, but may ignore the information in the time domain. Other methods focus more on time-domain analysis, using various statistical and time-domain feature extraction methods. These methods are capable of capturing signal variations over short periods of time, but may be relatively inadequate for modeling long-term temporal dependencies. In recent years, with the rise of deep learning, some researchers have begun to experiment with deep learning methods such as Convolutional Neural Networks (CNNs) and Recurrent Neural

Table 5.16. Comparison of epilepsy species identification results based on TUSZ database.

| **Publications** | **Classifier** | **Accuracy** | **F1 score** |
|---|---|---|---|
| Raghu et al. (2019) [232] | AlexNet | 84.06% | N/A |
| | VGG16 | 79.71% | N/A |
| | VGG19 | 76.81% | N/A |
| | CNN | 82.1% | N/A |
| | ResNet50 | N/A | 0.722 |
| Raghu et al. (2020) [233] | Googlenet | 82.85% | N/A |
| | Inceptionv3 | 88.30% | N/A |
| Tang et al. (2021) [234] | DCRNN | N/A | 0.749 |
| Shankar et al. (2021b) [235] | CNN | 89.91% | N/A |
| | | 84.19% | N/A |
| | | 84.20% | N/A |
| Shankar et al. (2021a) [230] | ANN | N/A | 0.997 |
| Thundiyil et al. (2021) [236] | AlexNet | 97.15% | 0.975 |
| | Resnet18 | 93.45% | 0.936 |
| | | 97.89% | 0.98 |
| | | 95.50% | 0.956 |
| Jia et al. (2022) [231] | VWCNNs | N/A | 0.94 |
| Shankar et al. (2022) [237] | CNN & LSTM | 98.82% | 0.988 |
| Dang et al. (2022) [238] | Transfer | 98.48% | 0.976 |
| Our | CNN & LSTM | 97.89% | 0.98 |



(a) ROC of epilepsy species identification based on the TUSZ database.

(b) Confusion matrix of epilepsy species identification based on the TUSZ database.

Fig. 5.14. ROC and confusion matrix of epilepsy species identification based on the TUSZ database.

Networks (RNNs, including LSTMs). These methods are usually better able to learn complex spatio-temporal features of signals, but they also require a large amount of labeled data and computational resources. Compared to the above methods, the innovation of this study is the cycle-based brainwave signal reconstruction method with a combination of multi-channel parallel convolutional neural networks and LSTM neural networks. Such a combination helps to fully utilize the time-frequency domain information and capture the temporal features in brainwave signals more comprehensively. In addition, we pay special attention to the diversity of epilepsy types and introduce the labeling process for multi-epilepsy signals, which enables the model to better adapt to multiple seizure situations in real clinical scenarios. These innovative approaches are expected to improve the accuracy and robustness of epilepsy species recognition.

## 5.6 Conclusions

In this study, an innovative method was used to convert EEG signals into image data. Convolutional neural networks (CNNs) and RNNs were utilized for processing, and accurate classification of different seizure types was successfully achieved. However, it is worth noting that the accuracy rate of 76.84% is not satisfactory. This result highlights the potential of this method in differentiating between various seizure types and also indicates room for improvement. The study built on that foundation and optimized it by using fewer channels for epilepsy analysis. This made the results more accurate, obtaining an accuracy of 97.89%.

It is important to note that the model cannot replace the expertise of a healthcare professional, especially when it comes to final diagnosis. Although it can assist in data

monitoring, it cannot replace the diagnostic ability of a physician. This limitation stems mainly from the possible interference caused by the large amount of irrelevant data. In the future, further data refinement is planned to minimize the impact of extraneous information. In addition, it is worth noting that EEG signals exhibit a time sequence during seizures, with different channels responding at different times. This phenomenon provides clues for early seizure detection and will be studied in depth in future studies to develop appropriate diagnostic methods.

In conclusion, this study introduces new perspectives and methods for recognizing epileptic seizure types, highlighting the great potential of deep learning and multichannel processing in the field of medical diagnosis. This study provides an opportunity to enhance the management and care of patients with epilepsy and has the potential to positively impact clinical practice.

# Chapter 6

# Summary and discussion

## 6.1 Conclusions and remarks

My journey through the realms of data analysis, computational techniques, and interdisciplinary collaboration has illuminated pathways towards innovation and progress across various domains. As I reflect on the methodologies explored and the insights gained, it becomes evident that each step taken has contributed to a broader understanding of complex challenges and potential solutions. In this concluding section, I delve deeper into the significance of my findings, discuss the implications for future research.

The exploration of GAF methodology marked the beginning of my quest for enhanced data presentation and analysis techniques. While navigating through challenges such as high memory requirements and computational complexity, I discovered the potential of downsampling methods and the SDM algorithm in speech emotion recognition. The superiority of the SDM model over traditional convolutional neural networks not only underscores its significance but also hints at the broader implications for data analysis methodologies. By embracing innovative algorithms and techniques, we can unlock new possibilities for understanding complex datasets and extracting meaningful insights.

Building upon the foundation laid by GAF methodology, my investigation into speech emotion recognition introduced a novel approach centered around short-time features and speech rate analysis. Through meticulous data preprocessing and model tuning, I achieved remarkable improvements in sentiment classification across multiple databases. The adoption of speech rate as a key parameter not only enhanced the accuracy and robustness of emotion recognition but also highlighted the importance of feature engineering in machine learning tasks. However, the limitations posed by dataset size underscore the need for larger and more diverse datasets to train and validate models effectively. Future research endeavors should focus on expanding datasets and exploring advanced feature engineering techniques to further improve model performance and generalization.

In the realm of data visualization, my exploration of transforming one-dimensional time-series data into visual representations opened new avenues for understanding and inter-

preting complex datasets. The efficacy of the Hilbert curve method in sentiment recognition showcases the potential of visual analytics in enhancing data interpretation and decision-making processes. By leveraging advanced visualization techniques, we can uncover hidden patterns and insights that may not be apparent through traditional analysis methods. Moreover, the interdisciplinary applications of data visualization extend beyond sentiment analysis to domains such as speech and brainwave signal analysis, offering new opportunities for cross-disciplinary collaboration and innovation.

Lastly, my endeavor to classify epileptic seizure types using deep learning techniques underscored the potential of multichannel processing in medical diagnosis. Despite initial challenges, I successfully optimized the classification process, achieving a remarkable accuracy rate. While recognizing the limitations of machine learning models in replacing healthcare professionals, my study emphasizes their potential to assist in early detection and improve patient care. Moving forward, further research efforts should focus on refining models and exploring novel diagnostic methods to enhance the management and care of patients with epilepsy.

## 6.2 Future works

In my future research, I will further delve into exploring the potential research directions of applying the time series binarization method to different areas. One of them is the use of this method on brainwave data to enable the identification of mental health problems such as depression. Research in this direction could provide new perspectives and tools for the mental health field and is expected to help improve the early diagnosis and treatment of depression.

Depression is a serious mental health disorder that is usually characterized by persistent low mood, loss of interest in daily activities, and multiple impairments in physical and cognitive functioning. Currently, the diagnosis of depression relies heavily on clinicians' experience and patients' self-reports. However, there are limitations in this subjective diagnostic approach; therefore, with the help of time-series binarization, we are expected to develop more objective and accurate diagnostic tools for depression.

In this line of research, I plan to process brainwave data into two-dimensional images using the time series two-dimensionalization method. This transformation is expected to capture important features in brainwave data, including information related to emotions and mental states. By designing neural network models applicable to brainwave images, I will attempt to extract and learn these features to enable accurate identification of depression.

In addition, I will explore the application of time series binarization methods to brainwave emotion recognition. Understanding an individual's emotional state is crucial for improving mental health, and brainwave data may contain information about emotional

experiences. By transforming time series into images, I will attempt to design feature extraction and classification models suitable for emotion recognition to achieve accurate judgments of an individual's emotional state.

The exploration of these research directions will introduce innovative methods and tools in the field of mental health and provide new possibilities for precision medicine and personalized treatment. By pushing time series dichotomization methods into new application areas, we expect to bring a positive impact on mental health research and clinical practice.

# Chapter 7

# Acknowledgement

As I stand on the cusp of completing my graduate studies, I am compelled to express my deepest gratitude to the many individuals who have supported and guided me throughout this academic journey. In this moment of reflection, I extend my heartfelt thanks to all those who have played a significant role in my academic and personal development.

First and foremost, I extend my sincere appreciation to my professor, Serikawa. Your guidance has been the beacon lighting my academic path. Your academic insights and personal support have been invaluable, shaping my research endeavors and providing me with profound lessons that will resonate throughout my life. Your patience and wisdom have resolved many of my academic dilemmas, allowing me to successfully complete my thesis. Thank you for your teachings and trust, which have been instrumental in my academic achievements. I would like to express gratitude to all members of the research lab. The collaborative spirit and mutual assistance within our team have created an environment conducive to academic growth. The experiences shared and knowledge gained during our collective efforts have enriched my academic journey. The friendships forged in this community will be cherished as enduring treasures. Special thanks are due to my family, particularly my parents and relatives, whose unwavering support has been my pillar of strength. In times of difficulty and setbacks, your encouragement and assistance have been my driving force. Your selfless dedication and hard work are the foundation of my achievements today. Thank you for your continual understanding, love, and support. Heartfelt appreciation goes out to all friends who have cared for and assisted me. Your companionship and encouragement have left indelible marks on my life, making our shared moments unforgettable. Together, we have navigated challenges and grown through mutual support, embodying the essence of true friendship.

I extend my thanks to the teachers, colleagues, and classmates who have contributed to my academic and personal growth. Your support and motivation have strengthened my resolve and confidence on the pursuit of knowledge. Lastly, I am grateful to the institution for providing me with the platform to pursue my studies. The opportunities for learning and growth have been immense, allowing me to meet outstanding individuals and gain a deeper understanding of the value of knowledge.

May this expression of gratitude reach each person who has contributed to my journey, and may the days of our collective efforts continue to radiate with brilliance in the future. Thank you all!

# Chapter 8

# Reference


[1] M. S. Kopp, A. Stauder, G. Purebl, I. Janszky, and A. Skrabski, "Work stress and mental health in a changing society," *European Journal of Public Health*, vol. 18, no. 3, pp. 238–244, 2008.

[2] M. S. Kopp, P. R. Falger, A. Appels, and S. Szedmak, "Depressive symptomatology and vital exhaustion are differentially related to behavioral risk factors for coronary artery disease," *Psychosomatic medicine*, vol. 60, no. 6, pp. 752–758, 1998.

[3] S. Dattani, L. Rod └ s-Guirao, H. Ritchie, and M. Roser, "Mental health," *Our World in Data*, 2023. https://ourworldindata.org/mental-health.

[4] C. Ruebeck, "Running head : Anxiety disorders anxiety disorders," 2009.

[5] K. S. Kumar, S. Srivastava, S. Paswan, A. S. Dutta, *et al.*, "Depression-symptoms, causes, medications and therapies," *The Pharma Innovation*, vol. 1, no. 3, Part A, p. 37, 2012.

[6] C. Habel, N. Feeley, B. Hayton, L. Bell, and P. Zelkowitz, "Causes of women's postpartum depression symptoms: Men's and women's perceptions," *Midwifery*, vol. 31, no. 7, pp. 728–734, 2015.

[7] J. G. Lee, Y. S. Woo, S. W. Park, D.-H. Seog, M. K. Seo, and W.-M. Bahk, "Neuromolecular etiology of bipolar disorder: possible therapeutic targets of mood stabilizers," *Clinical Psychopharmacology and Neuroscience*, vol. 20, no. 2, p. 228, 2022.

[8] J. Palmier-Claus, F. Lobban, W. Mansell, S. Jones, E. Tyler, C. Lodge, S. Bowe, A. Dodd, and K. Wright, "Mood monitoring in bipolar disorder: Is it always helpful?," *Bipolar disorders*, vol. 23, no. 4, pp. 429–431, 2021.

[9] R. O. Brady Jr, I. Gonsalvez, I. Lee, D. Öngür, L. J. Seidman, J. D. Schmahmann, S. M. Eack, M. S. Keshavan, A. Pascual-Leone, and M. A. Halko, "Cerebellar-prefrontal network connectivity and negative symptoms in schizophrenia," *American Journal of Psychiatry*, vol. 176, no. 7, pp. 512–520, 2019.

[10] A. Cohen, S. Chatterjee, and H. Minas, "Time for a global commission on mental health institutions," *World Psychiatry*, vol. 15, no. 2, p. 116, 2016.

[11] K. Zivin, M. Paczkowski, and S. Galea, "Economic downturns and population mental health: research findings, gaps, challenges and priorities," *Psychological medicine*,

vol. 41, no. 7, pp. 1343–1348, 2011.

[12] K. Wahlbeck and D. McDaid, "Actions to alleviate the mental health impact of the economic crisis," *World psychiatry*, vol. 11, no. 3, p. 139, 2012.

[13] H. Y. Chong, S. L. Teoh, D. B.-C. Wu, S. Kotirum, C.-F. Chiou, and N. Chaiyakunapruk, "Global economic burden of schizophrenia: a systematic review," *Neuropsychiatric disease and treatment*, pp. 357–373, 2016.

[14] G. . M. D. Collaborators *et al.*, "Global, regional, and national burden of 12 mental disorders in 204 countries and territories, 1990–2019: a systematic analysis for the global burden of disease study 2019," *The Lancet Psychiatry*, vol. 9, no. 2, pp. 137–150, 2022.

[15] M. Silva, D. Resurrección, A. Antunes, D. Frasquilho, and G. Cardoso, "Impact of economic crises on mental health care: a systematic review," *Epidemiology and psychiatric sciences*, vol. 29, p. e7, 2020.

[16] N. C. Coombs, W. E. Meriwether, J. Caringi, and S. R. Newcomer, "Barriers to healthcare access among us adults with mental health challenges: A population-based study," *SSM-population health*, vol. 15, p. 100847, 2021.

[17] F. Charlson, M. van Ommeren, A. Flaxman, J. Cornett, H. Whiteford, and S. Saxena, "New who prevalence estimates of mental disorders in conflict settings: a systematic review and meta-analysis," *The Lancet*, vol. 394, no. 10194, pp. 240–248, 2019.

[18] M. Gaiser, J. Buche, N. M. Baum, and K. L. Grazier, "Mental health needs due to disasters: Implications for behavioral health workforce planning during the covid-19 pandemic," *Public Health Reports*, vol. 138, no. 1_suppl, pp. 48S–55S, 2023.

[19] A. K. Tay, "The mental health needs of displaced people exposed to armed conflict," *The Lancet Public Health*, vol. 7, no. 5, pp. e398–e399, 2022.

[20] S. K. Padhy, S. Sarkar, M. Panigrahi, and S. Paul, "Mental health effects of climate change," *Indian journal of occupational and environmental medicine*, vol. 19, no. 1, p. 3, 2015.

[21] W. H. Organization *et al.*, "Management of physical health conditions in adults with severe mental disorders: Who guidelines," 2018.

[22] E. R. Walker, R. E. McGee, and B. G. Druss, "Mortality in mental disorders and global disease burden implications: a systematic review and meta-analysis," *JAMA psychiatry*, vol. 72, no. 4, pp. 334–341, 2015.

[23] Institute of Health Metrics and Evaluation, "Global health data exchange (ghdx)." `https://vizhub.healthdata.org/gbd-results/`, 2019. Accessed 14 May 2022.

[24] C. Hema and F. P. G. Marquez, "Emotional speech recognition using cnn and deep learning techniques," *Applied Acoustics*, vol. 211, p. 109492, 2023.

[25] W. Alsabhan, "Human–computer interaction with a real-time speech emotion recognition with ensembling techniques 1d convolution neural network and attention,"

*Sensors*, vol. 23, no. 3, p. 1386, 2023.

[26] R. W. Picard, "Affective computing: from laughter to ieee," *IEEE transactions on affective computing*, vol. 1, no. 1, pp. 11–17, 2010.

[27] C. Darwin and P. Prodger, *The expression of the emotions in man and animals*. Oxford University Press, USA, 1998.

[28] G. Pons and D. Masip, "Supervised committee of convolutional neural networks in automated facial expression analysis," *IEEE Transactions on Affective Computing*, vol. 9, pp. 343–350, 2018.

[29] B. Ilikci, L. Chen, H. Cho, and Q. Liu, "Heat-map based emotion and face recognition from thermal images," in *2019 Computing, Communications and IoT Applications (ComComAp)*, pp. 449–453, 2019.

[30] M. Kocon, "Rigid bones grouping scheme for facial expressions synthesis utilizing three-dimensional head model," 2010.

[31] D. R. Pant and R. Sthapit, "Analysis of micro facial expression by machine and deep learning methods: Haar, cnn, and rnn," 2021.

[32] G. Perna, A. Riva, A. Defillo, E. Sangiorgio, M. Nobile, and D. Caldirola, "Heart rate variability: Can it serve as a marker of mental health resilience: Special section on "translational and neuroscience studies in affective disorders" section editor, maria nobile md, phd," *Journal of Affective Disorders*, vol. 263, pp. 754–761, 2020.

[33] A. Barreto, J. Zhai, and M. Adjouadi, "Non-intrusive physiological monitoring for automated stress detection in human-computer interaction," in *Human–Computer Interaction: IEEE International Workshop, HCI 2007 Rio de Janeiro, Brazil, October 20, 2007 Proceedings 4*, pp. 29–38, Springer, 2007.

[34] A. B. Dollins, I. V. Zhdanova, R. J. Wurtman, H. J. Lynch, and M. H. Deng, "Effect of inducing nocturnal serum melatonin concentrations in daytime on sleep, mood, body temperature, and performance.," *Proceedings of the National Academy of Sciences*, vol. 91, no. 5, pp. 1824–1828, 1994.

[35] T. Malhotra, "Assessment of the validity of sentiment analysis as a tool to analyze the emotional content of text," 2019.

[36] S. N. S. M. Zain, N. A. Ramli, and R. A. Adnan, "Customer sentiment analysis through social media feedback: A case study on telecommunication company," *International Journal of Humanities Technology and Civilization*, 2022.

[37] M. R. Maarif, "Summarizing online customer review using topic modeling and sentiment analysis," *JISKA (Jurnal Informatika Sunan Kalijaga)*, 2022.

[38] Y. Prakash and D. K. Sharma, "A survey on aspect-based sentiment analysis for online purchased product feedback," *2023 10th International Conference on Computing for Sustainable Global Development (INDIACom)*, pp. 1342–1347, 2023.

[39] G. Garcia-Molina, T. Tsoneva, and A. Nijholt, "Emotional brain–computer interfaces," *International journal of autonomous and adaptive communications systems*,

vol. 6, no. 1, pp. 9–25, 2013.

[40] P. R. Bhise, S. B. Kulkarni, and T. A. Aldhaheri, "Brain computer interface based eeg for emotion recognition system: A systematic review," in *2020 2nd International Conference on Innovative Mechanisms for Industry Applications (ICIMIA)*, pp. 327–334, IEEE, 2020.

[41] V. Taschereau-Dumouchel, C. A. Cushing, and H. Lau, "Real-time functional mri in the treatment of mental health disorders," *Annual review of clinical psychology*, vol. 18, pp. 125–154, 2022.

[42] J. R. Wolpaw, "Brain–computer interfaces," in *Handbook of clinical neurology*, vol. 110, pp. 67–74, Elsevier, 2013.

[43] K. M. Naugle, C. J. Hass, D. Bowers, and C. M. Janelle, "Emotional state affects gait initiation in individuals with parkinson ¨ s disease," *Cognitive, Affective, & Behavioral Neuroscience*, vol. 12, pp. 207–219, 2012.

[44] G. Venture, H. Kadone, T. Zhang, J. Grèzes, A. Berthoz, and H. Hicheur, "Recognizing emotions conveyed by human gait," *International Journal of Social Robotics*, vol. 6, pp. 621–632, 2014.

[45] D. Janssen, W. I. Schöllhorn, J. Lubienetzki, K. Fölling, H. Kokenge, and K. Davids, "Recognition of emotions in gait patterns by means of artificial neural nets," *Journal of Nonverbal Behavior*, vol. 32, pp. 79–92, 2008.

[46] C. L. Roether, L. Omlor, A. Christensen, and M. A. Giese, "Critical features for the perception of emotion from gait," *Journal of vision*, vol. 9, no. 6, pp. 15–15, 2009.

[47] B. J. Abbaschian, D. Sierra-Sosa, and A. Elmaghraby, "Deep learning techniques for speech emotion recognition, from databases to models," *Sensors*, vol. 21, no. 4, p. 1249, 2021.

[48] D. Mamieva, A. B. Abdusalomov, A. Kutlimuratov, B. Muminov, and T. K. Whangbo, "Multimodal emotion detection via attention-based fusion of extracted facial and speech features," *Sensors*, vol. 23, no. 12, p. 5475, 2023.

[49] P. Srinivas and P. Mishra, "Human emotion recognition by integrating facial and speech features: An implementation of multimodal framework using cnn," *International Journal of Advanced Computer Science and Applications*, vol. 13, no. 1, 2022.

[50] M. G. Huddar, S. S. Sannakki, and V. S. Rajpurohit, "Multimodal emotion recognition using facial expressions, body gestures, speech, and text modalities," *International Journal of Engineering and Advanced Technology (IJEAT)*, vol. 8, no. 5, pp. 2453–2459, 2019.

[51] G. Du, Y. Zeng, K. Su, C. Li, X. Wang, S. Teng, D. Li, and P. X. Liu, "A novel emotion-aware method based on the fusion of textual description of speech, body movements, and facial expressions," *IEEE Transactions on Instrumentation and Measurement*, vol. 71, pp. 1–16, 2022.

[52] M. El Ayadi, M. S. Kamel, and F. Karray, "Survey on speech emotion recognition: Features, classification schemes, and databases," *Pattern recognition*, vol. 44, no. 3, pp. 572–587, 2011.

[53] R. P. Gadhe, D. Babasaheb, R. Deshmukh, and D. Babasaheb, "Emotion recognition from isolated marathi speech using energy and formants," *International Journal of Computer Applications*, vol. 975, p. 8887, 2015.

[54] R. W. Picard, *Affective computing*. MIT press, 2000.

[55] M. B. Akçay and K. Oğuz, "Speech emotion recognition: Emotional models, databases, features, preprocessing methods, supporting modalities, and classifiers," *Speech Communication*, vol. 116, pp. 56–76, 2020.

[56] M. Lech, M. Stolar, C. Best, and R. Bolia, "Real-time speech emotion recognition using a pre-trained image classification network: Effects of bandwidth reduction and companding," *Frontiers in Computer Science*, vol. 2, p. 14, 2020.

[57] S. Madanian, T. Chen, O. Adeleye, J. M. Templeton, C. Poellabauer, D. Parry, and S. L. Schneider, "Speech emotion recognition using machine learning – a systematic review," *Intelligent Systems with Applications*, p. 200266, 2023.

[58] S. Ramakrishnan and I. M. El Emary, "Speech emotion recognition approaches in human computer interaction," *Telecommunication Systems*, vol. 52, pp. 1467–1478, 2013.

[59] N. Kwon, S. Hossain, N. Blaylock, H. O ¨ Connell, N. Hachen, and J. Gwin, "Detecting anxiety and depression from phone conversations using x-vectors," in *Proc. Workshop on Speech, Music and Mind*, pp. 1–5.

[60] D. Dukes, A. Samson, and E. Walle, *The Oxford handbook of emotional development*. Oxford University Press, 2022.

[61] D. Indira, B. Prasanna, C. Pavani, and G. Vandana, "Assessment of patient health condition based on speech emotion recognition (ser) using deep learning algorithms," *European Journal of Molecular & Clinical Medicine*, vol. 7, no. 4, p. 2020, 2020.

[62] L. Cen, F. Wu, Z. L. Yu, and F. Hu, "A real-time speech emotion recognition system and its application in online learning," in *Emotions, technology, design, and learning*, pp. 27–46, Elsevier, 2016.

[63] J. Lin, R. Khade, and Y. Li, "Rotation-invariant similarity in time series using bag-of-patterns representation," *Journal of Intelligent Information Systems*, vol. 39, pp. 287–315, 2012.

[64] T. Rakthanmanon and E. Keogh, "Fast shapelets: A scalable algorithm for discovering time series shapelets," in *proceedings of the 2013 SIAM International Conference on Data Mining*, pp. 668–676, SIAM, 2013.

[65] P. Senin and S. Malinchik, "Sax-vsm: Interpretable time series classification using sax and vector space model," in *2013 IEEE 13th international conference on data mining*, pp. 1175–1180, IEEE, 2013.

[66] D. F. Silva, V. M. De Souza, and G. E. Batista, "Time series classification using compression distance of recurrence plots," in *2013 IEEE 13th International Conference on Data Mining*, pp. 687–696, IEEE, 2013.

[67] M. G. Baydogan, G. Runger, and E. Tuv, "A bag-of-features framework to classify time series," *IEEE transactions on pattern analysis and machine intelligence*, vol. 35, no. 11, pp. 2796–2802, 2013.

[68] T. Rakthanmanon and E. Keogh, "Fast shapelets: A scalable algorithm for discovering time series shapelets," in *proceedings of the 2013 SIAM International Conference on Data Mining*, pp. 668–676, SIAM, 2013.

[69] Z. Wang, T. Oates, *et al.*, "Encoding time series as images for visual inspection and classification using tiled convolutional neural networks," in *Workshops at the twenty-ninth AAAI conference on artificial intelligence*, vol. 1, AAAI Menlo Park, CA, USA, 2015.

[70] A. Bakhshi, A. Harimi, and S. Chalup, "Cytex: Transforming speech to textured images for speech emotion recognition," *Speech Communication*, vol. 139, pp. 62–75, 2022.

[71] F. Lederbogen, L. Haddad, and A. Meyer-Lindenberg, "Urban social stress–risk factor for mental disorders. the case of schizophrenia," *Environmental pollution*, vol. 183, pp. 2–6, 2013.

[72] M. S. Kopp, A. Stauder, G. Purebl, I. Janszky, and A. Skrabski, "Work stress and mental health in a changing society," *European Journal of Public Health*, vol. 18, no. 3, pp. 238–244, 2008.

[73] M. S. Kopp, P. R. Falger, A. Appels, and S. Szedmak, "Depressive symptomatology and vital exhaustion are differentially related to behavioral risk factors for coronary artery disease," *Psychosomatic medicine*, vol. 60, no. 6, pp. 752–758, 1998.

[74] M. S. Kopp and J. Réthelyi, "Where psychology meets physiology: chronic stress and premature mortality ! the central-eastern european health paradox," *Brain research bulletin*, vol. 62, no. 5, pp. 351–367, 2004.

[75] J. Swanson, M. Swartz, S. Estroff, R. Borum, R. Wagner, and V. Hiday, "Psychiatric impairment, social contact, and violent behavior: evidence from a study of outpatient-committed persons with severe mental disorder," *Social Psychiatry and Psychiatric Epidemiology*, vol. 33, pp. S86–S94, 1998.

[76] V. A. Hiday, "The social context of mental illness and violence," *Journal of Health and Social Behavior*, pp. 122–137, 1995.

[77] S. M. Rabbitt, A. E. Kazdin, and B. Scassellati, "Integrating socially assistive robotics into mental healthcare interventions: Applications and recommendations for expanded use," *Clinical psychology review*, vol. 35, pp. 35–46, 2015.

[78] S. Lucas and T. Wade, "An examination of the power of the voices in predicting the mental state of people experiencing psychosis," *Behaviour Change*, vol. 18, no. 1,

pp. 51–57, 2001.

[79] N. Cummins, S. Scherer, J. Krajewski, S. Schnieder, J. Epps, and T. F. Quatieri, "A review of depression and suicide risk assessment using speech analysis," *Speech communication*, vol. 71, pp. 10–49, 2015.

[80] M. Cannizzaro, B. Harel, N. Reilly, P. Chappell, and P. J. Snyder, "Voice acoustical measurement of the severity of major depression," *Brain and cognition*, vol. 56, no. 1, pp. 30–35, 2004.

[81] J. Nicholson, K. Takahashi, and R. Nakatsu, "Emotion recognition in speech using neural networks," *Neural computing & applications*, vol. 9, pp. 290–296, 2000.

[82] M. Song, M. You, N. Li, and C. Chen, "A robust multimodal approach for emotion recognition," *Neurocomputing*, vol. 71, no. 10-12, pp. 1913–1920, 2008.

[83] R. Farzanfar, S. Frishkopf, R. Friedman, and K. Ludena, "Evaluating an automated mental health care system: making meaning of human–computer interaction," *Computers in Human Behavior*, vol. 23, no. 3, pp. 1167–1182, 2007.

[84] A. A. Scoglio, E. D. Reilly, J. A. Gorman, and C. E. Drebing, "Use of social robots in mental health and well-being research: systematic review," *Journal of medical Internet research*, vol. 21, no. 7, p. e13322, 2019.

[85] B. W. Schuller, "Speech emotion recognition: Two decades in a nutshell, benchmarks, and ongoing trends," *Communications of the ACM*, vol. 61, no. 5, pp. 90–99, 2018.

[86] P. Mermelstein, "Distance measures for speech recognition, psychological and instrumental," *Pattern recognition and artificial intelligence*, vol. 116, pp. 374–388, 1976.

[87] J. Shen, R. Pang, R. J. Weiss, M. Schuster, N. Jaitly, Z. Yang, Z. Chen, Y. Zhang, Y. Wang, R. Skerrv-Ryan, *et al.*, "Natural tts synthesis by conditioning wavenet on mel spectrogram predictions," in *2018 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, pp. 4779–4783, IEEE, 2018.

[88] T. L. Nwe, S. W. Foo, and L. C. De Silva, "Speech emotion recognition using hidden markov models," *Speech communication*, vol. 41, no. 4, pp. 603–623, 2003.

[89] F. Eyben, K. R. Scherer, B. W. Schuller, J. Sundberg, E. André, C. Busso, L. Y. Devillers, J. Epps, P. Laukka, S. S. Narayanan, *et al.*, "The geneva minimalistic acoustic parameter set (gemaps) for voice research and affective computing," *IEEE transactions on affective computing*, vol. 7, no. 2, pp. 190–202, 2015.

[90] C. S. Ooi, K. P. Seng, L.-M. Ang, and L. W. Chew, "A new approach of audio emotion recognition," *Expert systems with applications*, vol. 41, no. 13, pp. 5858–5869, 2014.

[91] C. S. Ooi, K. P. Seng, L.-M. Ang, and L. W. Chew, "A new approach of audio emotion recognition," *Expert systems with applications*, vol. 41, no. 13, pp. 5858–5869, 2014.

[92] Z. Wang and T. Oates, "Imaging time-series to improve classification and imputation," *arXiv preprint arXiv:1506.00327*, 2015.

[93] A. S. Campanharo, M. I. Sirer, R. D. Malmgren, F. M. Ramos, and L. A. N. Amaral, "Duality between time series and networks," *PloS one*, vol. 6, no. 8, p. e23378, 2011.

[94] E. J. Keogh and M. J. Pazzani, "Scaling up dynamic time warping for datamining applications," in *Proceedings of the sixth ACM SIGKDD international conference on Knowledge discovery and data mining*, pp. 285–289, 2000.

[95] D. H. Douglas and T. K. Peucker, "Algorithms for the reduction of the number of points required to represent a digitized line or its caricature," *Cartographica: the international journal for geographic information and geovisualization*, vol. 10, no. 2, pp. 112–122, 1973.

[96] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *arXiv preprint arXiv:1409.1556*, 2014.

[97] Pypi, "visvalingamwyatt," 2022.

[98] S. Steinarsson, *Downsampling time series for visual representation.* PhD thesis, 2013.

[99] J. R. Bunch, L. Kaufman, and B. N. Parlett, "Decomposition of a symmetric matrix," *Numerische Mathematik*, vol. 27, no. 1, pp. 95–109, 1976.

[100] E. P. Wigner, "On the distribution of the roots of certain symmetric matrices," *Annals of Mathematics*, vol. 67, no. 2, pp. 325–327, 1958.

[101] M. El Ayadi, M. S. Kamel, and F. Karray, "Survey on speech emotion recognition: Features, classification schemes, and databases," *Pattern recognition*, vol. 44, no. 3, pp. 572–587, 2011.

[102] J. Nicholson, K. Takahashi, and R. Nakstsu, "Emotion recognition in speech using neural networks, neural information processing, 1999," in *Proceedings. ICONIP*, vol. 99, pp. 495–501.

[103] M. Song, M. You, N. Li, and C. Chen, "A robust multimodal approach for emotion recognition," *Neurocomputing*, vol. 71, no. 10-12, pp. 1913–1920, 2008.

[104] T. L. Nwe, S. W. Foo, and L. C. De Silva, "Speech emotion recognition using hidden markov models," *Speech communication*, vol. 41, no. 4, pp. 603–623, 2003.

[105] K. Vicsi and D. Sztahó, "Emotional state recognition in customer service dialogues through telephone line," in *2011 2nd International Conference on Cognitive Infocommunications (CogInfoCom)*, pp. 1–4, IEEE, 2011.

[106] C. M. Lee and S. S. Narayanan, "Toward detecting emotions in spoken dialogs," *IEEE transactions on speech and audio processing*, vol. 13, no. 2, pp. 293–303, 2005.

[107] J. F. Sánchez-Rada and C. A. Iglesias, "Social context in sentiment analysis: Formal definition, overview of current trends and framework for comparison," *Information Fusion*, vol. 52, pp. 344–356, 2019.

[108] M. Wankhade, A. C. S. Rao, and C. Kulkarni, "A survey on sentiment analysis

methods, applications, and challenges," *Artificial Intelligence Review*, vol. 55, no. 7, pp. 5731–5780, 2022.

[109] M. Cannizzaro, B. Harel, N. Reilly, P. Chappell, and P. J. Snyder, "Voice acoustical measurement of the severity of major depression," *Brain and cognition*, vol. 56, no. 1, pp. 30–35, 2004.

[110] N. Cummins, S. Scherer, J. Krajewski, S. Schnieder, J. Epps, and T. F. Quatieri, "A review of depression and suicide risk assessment using speech analysis," *Speech communication*, vol. 71, pp. 10–49, 2015.

[111] J. C. Mundt, P. J. Snyder, M. S. Cannizzaro, K. Chappie, and D. S. Geralts, "Voice acoustic measures of depression severity and treatment response collected via interactive voice response (ivr) technology," *Journal of neurolinguistics*, vol. 20, no. 1, pp. 50–64, 2007.

[112] B. Schuller, F. Weninger, Y. Zhang, F. Ringeval, A. Batliner, S. Steidl, F. Eyben, E. Marchi, A. Vinciarelli, K. Scherer, *et al.*, "Affective and behavioural computing: Lessons learnt from the first computational paralinguistics challenge," *Computer Speech & Language*, vol. 53, pp. 156–180, 2019.

[113] U. Petti, S. Baker, and A. Korhonen, "A systematic literature review of automatic alzheimer ¨ s disease detection from speech and language," *Journal of the American Medical Informatics Association*, vol. 27, no. 11, pp. 1784–1797, 2020.

[114] K. C. Fraser, K. Lundholm Fors, M. Eckerström, F. Öhman, and D. Kokkinakis, "Predicting mci status from multimodal language data using cascaded classifiers," *Frontiers in aging neuroscience*, vol. 11, p. 205, 2019.

[115] C. Nass, I.-M. Jonsson, H. Harris, B. Reaves, J. Endo, S. Brave, and L. Takayama, "Improving automotive safety by pairing driver emotion and car voice emotion," in *CHI'05 extended abstracts on Human factors in computing systems*, pp. 1973–1976, 2005.

[116] L. Zhu, L. Chen, D. Zhao, J. Zhou, and W. Zhang, "Emotion recognition from chinese speech for smart affective services using a combination of svm and dbn," *Sensors*, vol. 17, no. 7, p. 1694, 2017.

[117] B. Schuller, S. Steidl, A. Batliner, F. Burkhardt, L. Devillers, C. Müller, and S. Narayanan, "The interspeech 2010 paralinguistic challenge," in *Proc. INTERSPEECH 2010, Makuhari, Japan*, pp. 2794–2797, 2010.

[118] R. A. Khalil, E. Jones, M. I. Babar, T. Jan, M. H. Zafar, and T. Alhussain, "Speech emotion recognition using deep learning techniques: A review," *IEEE Access*, vol. 7, pp. 117327–117345, 2019.

[119] R. Banse and K. R. Scherer, "Acoustic profiles in vocal emotion expression.," *Journal of personality and social psychology*, vol. 70, no. 3, p. 614, 1996.

[120] F. Eyben, K. R. Scherer, B. W. Schuller, J. Sundberg, E. André, C. Busso, L. Y. Devillers, J. Epps, P. Laukka, S. S. Narayanan, *et al.*, "The geneva minimalistic

acoustic parameter set (gemaps) for voice research and affective computing," *IEEE transactions on affective computing*, vol. 7, no. 2, pp. 190–202, 2015.

[121] M. Tahon and L. Devillers, "Towards a small set of robust acoustic features for emotion recognition: challenges," *IEEE/ACM transactions on audio, speech, and language processing*, vol. 24, no. 1, pp. 16–28, 2015.

[122] Z.-T. Liu, M. Wu, W.-H. Cao, J.-W. Mao, J.-P. Xu, and G.-Z. Tan, "Speech emotion recognition based on feature selection and extreme learning machine decision tree," *Neurocomputing*, vol. 273, pp. 271–280, 2018.

[123] A. K. Samantaray, K. Mahapatra, B. Kabi, and A. Routray, "A novel approach of speech emotion recognition with prosody, quality and derived features using svm classifier for a class of north-eastern languages," in *2015 IEEE 2nd International Conference on Recent Trends in Information Systems (ReTIS)*, pp. 372–377, IEEE, 2015.

[124] J. G. Rázuri, D. Sundgren, R. Rahmani, A. Moran, I. Bonet, and A. Larsson, "Speech emotion recognition in emotional feedbackfor human-robot interaction," *International Journal of Advanced Research in Artificial Intelligence (IJARAI)*, vol. 4, no. 2, pp. 20–27, 2015.

[125] Z. Huang, J. Epps, and E. Ambikairajah, "An investigation of emotion change detection from speech," in *Sixteenth Annual Conference of the International Speech Communication Association*, 2015.

[126] A. Rawat and P. K. Mishra, "Emotion recognition through speech using neural network," *Int. J*, vol. 5, pp. 422–428, 2015.

[127] J. Lee and I. Tashev, "High-level feature representation using recurrent neural network for speech emotion recognition," in *Interspeech 2015*, 2015.

[128] ChineseLDC, "Chinese academy of sciences emotional speech database." `https://www.ChineseLDC.Org`, Accessed April 4, 2021.

[129] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural computation*, vol. 9, no. 8, pp. 1735–1780, 1997.

[130] J. D. Markel and A. J. Gray, *Linear prediction of speech*, vol. 12. Springer Science & Business Media, 2013.

[131] L. Zhu, L. Chen, D. Zhao, J. Zhou, and W. Zhang, "Emotion recognition from chinese speech for smart affective services using a combination of svm and dbn," *Sensors*, vol. 17, no. 7, p. 1694, 2017.

[132] L. R. Rabiner, *Digital processing of speech signals*. Pearson Education India, 1978.

[133] S. G. Koolagudi and R. S. Krothapalli, "Two stage emotion recognition based on speaking rate," *International Journal of Speech Technology*, vol. 14, pp. 35–48, 2011.

[134] S. Ramakrishnan, "Recognition of emotion from speech: A review," *Speech Enhancement, Modeling and recognition–algorithms and Applications*, vol. 7, pp. 121–137, 2012.

[135] R. W. Wescott, "Linguistic iconism," *Language*, pp. 416–428, 1971.

[136] P. Dan-ling and Y. Hui, "The phonological processing of chinese phonograms," *Asia Pacific Journal of Speech, Language and Hearing*, vol. 2, no. 3, pp. 177–194, 1997.

[137] M. Sugishita, K. Otomo, S. Kabe, and K. Yunoki, "A critical appraisal of neuropsychological correlates of japanese ideogram (kanji) and phonogram (kana) reading," *Brain*, vol. 115, no. 5, pp. 1563–1585, 1992.

[138] R. G. Bachu, S. Kopparthi, B. Adapa, and B. D. Barkana, "Voiced/unvoiced decision for speech signals based on zero-crossing rate and energy," in *Advanced techniques in computing sciences and software engineering*, pp. 279–282, Springer, 2010.

[139] M. Guzman, S. Correa, D. Munoz, and R. Mayerhoff, "Influence on spectral energy distribution of emotional expression," *Journal of voice*, vol. 27, no. 1, pp. 129–e1, 2013.

[140] M. Jalil, F. A. Butt, and A. Malik, "Short-time energy, magnitude, zero crossing rate and autocorrelation measurement for discriminating voiced and unvoiced segments of speech signals," in *2013 The international conference on technological advances in electrical, electronics and computer engineering (TAEECE)*, pp. 208–212, IEEE, 2013.

[141] R. W. Schafer and L. R. Rabiner, "Digital representations of speech signals," *Proceedings of the IEEE*, vol. 63, no. 4, pp. 662–677, 1975.

[142] J. Makhoul, "Linear prediction: A tutorial review," *Proceedings of the IEEE*, vol. 63, no. 4, pp. 561–580, 1975.

[143] I. R. Titze, "Where has all the power gone? energy production and loss in vocalization," *Speech communication*, vol. 101, pp. 26–33, 2018.

[144] R. Bachu, S. Kopparthi, B. Adapa, and B. Barkana, "Separation of voiced and unvoiced using zero crossing rate and energy of the speech signal," in *American Society for Engineering Education (ASEE) zone conference proceedings*, pp. 1–7, American Society for Engineering Education, 2008.

[145] X. Yang, B. Tan, J. Ding, J. Zhang, and J. Gong, "Comparative study on voice activity detection algorithm," in *2010 International Conference on Electrical and Control Engineering*, pp. 599–602, IEEE, 2010.

[146] M. Jalil, F. A. Butt, and A. Malik, "Short-time energy, magnitude, zero crossing rate and autocorrelation measurement for discriminating voiced and unvoiced segments of speech signals," in *2013 The international conference on technological advances in electrical, electronics and computer engineering (TAEECE)*, pp. 208–212, IEEE, 2013.

[147] M. Jalil, F. A. Butt, and A. Malik, "Short-time energy, magnitude, zero crossing rate and autocorrelation measurement for discriminating voiced and unvoiced segments of speech signals," in *2013 The international conference on technological advances in electrical, electronics and computer engineering (TAEECE)*, pp. 208–212, IEEE,

2013.

[148] Y. Qi and R. E. Hillman, "Temporal and spectral estimations of harmonics-to-noise ratio in human voice signals," *The Journal of the Acoustical Society of America*, vol. 102, no. 1, pp. 537–543, 1997.

[149] P. J. Murphy and O. O. Akande, "Noise estimation in voice signals using short-term cepstral analysis," *The Journal of the Acoustical Society of America*, vol. 121, no. 3, pp. 1679–1690, 2007.

[150] M. Sampaio, M. L. Vaz Masson, M. F. de Paula Soares, J. E. Bohlender, and M. Brockmann-Bauser, "Effects of fundamental frequency, vocal intensity, sample duration, and vowel context in cepstral and spectral measures of dysphonic voices," *Journal of Speech, Language, and Hearing Research*, vol. 63, no. 5, pp. 1326–1339, 2020.

[151] P. de Boves Harrington, "Support vector machine classification trees based on fuzzy entropy of classification," *Analytica chimica acta*, vol. 954, pp. 14–21, 2017.

[152] W. Zaremba, I. Sutskever, and O. Vinyals, "Recurrent neural network regularization," *arXiv preprint arXiv:1409.2329*, 2014.

[153] L. S.-T. Memory, "Long short-term memory," *Neural computation*, vol. 9, no. 8, pp. 1735–1780, 2010.

[154] Y. Zhou, C. Xiong, and R. Socher, "Improved regularization techniques for end-to-end speech recognition," *arXiv preprint arXiv:1712.07108*, 2017.

[155] L. R. Rabiner and M. R. Sambur, "An algorithm for determining the endpoints of isolated utterances," *Bell System Technical Journal*, vol. 54, no. 2, pp. 297–315, 1975.

[156] S. Kshirsagar, A. Pendyala, and T. H. Falk, "Task-specific speech enhancement and data augmentation for improved multimodal emotion recognition under noisy conditions," *Frontiers in Computer Science*, vol. 5, p. 1039261, 2023.

[157] J. Zhao, X. Mao, and L. Chen, "Speech emotion recognition using deep 1d & 2d cnn lstm networks," *Biomedical signal processing and control*, vol. 47, pp. 312–323, 2019.

[158] S. K. Pandey, H. S. Shekhawat, and S. M. Prasanna, "Deep learning techniques for speech emotion recognition: A review," in *2019 29th International Conference Radioelektronika (RADIOELEKTRONIKA)*, pp. 1–6, IEEE, 2019.

[159] N. Senthilkumar, S. Karpakam, M. G. Devi, R. Balakumaresan, and P. Dhilipkumar, "Speech emotion recognition based on bi-directional lstm architecture and deep belief networks," *Materials Today: Proceedings*, vol. 57, pp. 2180–2184, 2022.

[160] J. Ancilin and A. Milton, "Improved speech emotion recognition with mel frequency magnitude coefficient," *Applied Acoustics*, vol. 179, p. 108046, 2021.

[161] M. S. Saleem, A. S. N. Isha, M. I. Awan, Y. B. Yusop, and G. M. A. Naji, "Fostering academic engagement in post-graduate students: Assessing the role of positive emo-

tions, positive psychology, and stress," *Frontiers in Psychology*, vol. 13, p. 920395, 2022.

[162] P. Starosotskaya, "Emotions as a driving motive of human behavior. a review of robert frank's book on the key role of emotions in decision making," *Philosophy. Journal of the Higher School of Economics*, 2018.

[163] R. S. Lazarus, "Emotions and interpersonal relationships: Toward a person-centered conceptualization of emotions and coping," *Journal of personality*, vol. 74, no. 1, pp. 9–46, 2006.

[164] G. Stephanou and K. Athanasiadou, "Interpersonal relationships: Cognitive appraisals, emotions and hope.," *European Journal of Psychology and Educational Research*, vol. 3, no. 1, pp. 13–38, 2020.

[165] G. Sarvani, "A comparison between bagavad gita and modern mechanism of emotions and attitudes in human behavior.," *Indian Journal of Positive Psychology*, vol. 6, no. 4, 2015.

[166] L. O'Connor, "How social workers understand and use their emotions in practice: A thematic synthesis literature review," *Qualitative Social Work*, vol. 19, no. 4, pp. 645–662, 2020.

[167] C. Marinetti, P. Moore, P. Lucas, and B. Parkinson, "Emotions in social interactions: Unfolding emotional experience," *Emotion-oriented systems: The humaine handbook*, pp. 31–46, 2011.

[168] P. Starosotskaya, "Emotions as a driving motive of human behavior. a review of robert frank's book on the key role of emotions in decision making," *Philosophy. Journal of the Higher School of Economics*, 2018.

[169] T. Krettenauer, T. Colasante, M. Buchmann, and T. Malti, "The development of moral emotions and decision-making from adolescence to early adulthood: A 6-year longitudinal study," *Journal of youth and adolescence*, vol. 43, pp. 583–596, 2014.

[170] E. A. Kemp, A. L. Borders, N. A. Anaza, and W. J. Johnston, "The heart in organizational buying: marketers ¨ understanding of emotions and decision-making of buyers," *Journal of Business & Industrial Marketing*, vol. 33, no. 1, pp. 19–28, 2018.

[171] M. Escadas, M. S. Jalali, and M. Farhangmehr, "Why bad feelings predict good behaviours: The role of positive and negative anticipated emotions on consumer ethical decision making," *Business Ethics: A European Review*, vol. 28, no. 4, pp. 529–545, 2019.

[172] K. hülya Akdemir, "How main negative emotions effect people ¨ s political decision making process in a positive way? : In the sense of ^ fear, anxiety and anger ¯," 2021.

[173] M. El Ayadi, M. S. Kamel, and F. Karray, "Survey on speech emotion recognition: Features, classification schemes, and databases," *Pattern recognition*, vol. 44, no. 3,

pp. 572–587, 2011.

[174] M. Lech, M. Stolar, C. Best, and R. Bolia, "Real-time speech emotion recognition using a pre-trained image classification network: Effects of bandwidth reduction and companding," *Frontiers in Computer Science*, vol. 2, p. 14, 2020.

[175] S. Madanian, T. Chen, O. Adeleye, J. M. Templeton, C. Poellabauer, D. Parry, and S. L. Schneider, "Speech emotion recognition using machine learning – a systematic review," *Intelligent Systems with Applications*, p. 200266, 2023.

[176] A. Davletcharova, S. Sugathan, B. Abraham, and A. P. James, "Detection and analysis of emotion from speech signals," *Procedia Computer Science*, vol. 58, pp. 91–96, 2015.

[177] S. Chamishka, I. Madhavi, R. Nawaratne, D. Alahakoon, D. De Silva, N. Chilamkurti, and V. Nanayakkara, "A voice-based real-time emotion detection technique using recurrent neural network empowered feature modelling," *Multimedia Tools and Applications*, vol. 81, no. 24, pp. 35173–35194, 2022.

[178] A. Hashem, M. Arif, and M. Alghamdi, "Speech emotion recognition approaches: A systematic review," *Speech Communication*, p. 102974, 2023.

[179] A. Jain and H. R. Sah, "Student's feedback by emotion and speech recognition through deep learning," in *2021 International Conference on Computing, Communication, and Intelligent Systems (ICCCIS)*, pp. 442–447, IEEE, 2021.

[180] M. Higuchi, M. Nakamura, S. Shinohara, Y. Omiya, T. Takano, S. Mitsuyoshi, S. Tokuno, *et al.*, "Effectiveness of a voice-based mental health evaluation system for mobile devices: prospective study," *JMIR formative research*, vol. 4, no. 7, p. e16455, 2020.

[181] D. M. Low, K. H. Bentley, and S. S. Ghosh, "Automated assessment of psychiatric disorders using speech: A systematic review," *Laryngoscope investigative otolaryngology*, vol. 5, no. 1, pp. 96–116, 2020.

[182] Y. Wang, L. Liang, Z. Zhang, X. Xu, R. Liu, H. Fang, R. Zhang, Y. Wei, Z. Liu, R. Zhu, *et al.*, "Fast and accurate assessment of depression based on voice acoustic features: a cross-sectional and longitudinal study," *Frontiers in Psychiatry*, vol. 14, p. 1195276, 2023.

[183] J. F. Sánchez-Rada and C. A. Iglesias, "Social context in sentiment analysis: Formal definition, overview of current trends and framework for comparison," *Information Fusion*, vol. 52, pp. 344–356, 2019.

[184] S. Abeysinghe, I. Manchanayake, C. Samarajeewa, P. Rathnayaka, M. J. Walpola, R. Nawaratne, T. Bandaragoda, and D. Alahakoon, "Enhancing decision making capacity in tourism domain using social media analytics," in *2018 18th International Conference on Advances in ICT for Emerging Regions (ICTer)*, pp. 369–375, IEEE, 2018.

[185] M. Wankhade, A. C. S. Rao, and C. Kulkarni, "A survey on sentiment analysis

methods, applications, and challenges," *Artificial Intelligence Review*, vol. 55, no. 7, pp. 5731–5780, 2022.

[186] F. Noroozi, M. Marjanovic, A. Njegus, S. Escalera, and G. Anbarjafari, "Audio-visual emotion recognition in video clips," *IEEE Transactions on Affective Computing*, vol. 10, no. 1, pp. 60–75, 2017.

[187] C. Zheng, C. Wang, and N. Jia, "An ensemble model for multi-level speech emotion recognition," *Applied Sciences*, vol. 10, no. 1, p. 205, 2019.

[188] T. Anvarjon, Mustaqeem, and S. Kwon, "Deep-net: A lightweight cnn-based speech emotion recognition system using deep frequency features," *Sensors*, vol. 20, no. 18, p. 5212, 2020.

[189] S. Zhang, R. Liu, X. Tao, and X. Zhao, "Deep cross-corpus speech emotion recognition: Recent advances and perspectives," *Frontiers in neurorobotics*, vol. 15, p. 784514, 2021.

[190] M. Swain, B. Maji, P. Kabisatpathy, and A. Routray, "A dcrnn-based ensemble classifier for speech emotion recognition in odia language," *Complex & Intelligent Systems*, vol. 8, no. 5, pp. 4237–4249, 2022.

[191] A. Batliner, B. Schuller, D. Seppi, S. Steidl, L. Devillers, L. Vidrascu, T. Vogt, V. Aharonson, and N. Amir, *The automatic recognition of emotions in speech*. Springer, 2011.

[192] C.-N. Anagnostopoulos, T. Iliou, and I. Giannoukos, "Features and classifiers for emotion recognition from speech: a survey from 2000 to 2011," *Artificial Intelligence Review*, vol. 43, pp. 155–177, 2015.

[193] M. Ayadi, M. Kamel, and F. Karray, "Survey on speech recognition: Resources, features and methods," *Pattern Recognition*, vol. 44, no. 3, pp. 572–587, 2011.

[194] Y. Yang, C. Fairbairn, and J. F. Cohn, "Detecting depression severity from vocal prosody," *IEEE transactions on affective computing*, vol. 4, no. 2, pp. 142–150, 2012.

[195] J. C. Mundt, A. P. Vogel, D. E. Feltner, and W. R. Lenderking, "Vocal acoustic biomarkers of depression severity and treatment response," *Biological psychiatry*, vol. 72, no. 7, pp. 580–587, 2012.

[196] J. C. Mundt, P. J. Snyder, M. S. Cannizzaro, K. Chappie, and D. S. Geralts, "Voice acoustic measures of depression severity and treatment response collected via interactive voice response (ivr) technology," *Journal of neurolinguistics*, vol. 20, no. 1, pp. 50–64, 2007.

[197] C. Hema and F. P. G. Marquez, "Emotional speech recognition using cnn and deep learning techniques," *Applied Acoustics*, vol. 211, p. 109492, 2023.

[198] A. Bakhshi, A. Harimi, and S. Chalup, "Cytex: Transforming speech to textured images for speech emotion recognition," *Speech Communication*, vol. 139, pp. 62–75, 2022.

[199] D. Hilbert and D. Hilbert, "Über die stetige abbildung einer linie auf ein

flächenstück," *Dritter Band: Analysis· Grundlagen der Mathematik· Physik Verschiedenes: Nebst Einer Lebensgeschichte*, pp. 1–2, 1935.

[200] ChineseLDC, "Chinese academy of sciences emotional speech database." `https://www.ChineseLDC.Org`, Accessed April 4, 2021.

[201] O. S. Kayhan and J. C. v. Gemert, "On translation invariance in cnns: Convolutional layers can exploit absolute spatial location," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 14274–14285, 2020.

[202] Z.-T. Liu, M. Wu, W.-H. Cao, J.-W. Mao, J.-P. Xu, and G.-Z. Tan, "Speech emotion recognition based on feature selection and extreme learning machine decision tree," *Neurocomputing*, vol. 273, pp. 271–280, 2018.

[203] W. H. Organization, "Epilepsy," 2017.

[204] M. J. Moroney, *Facts from figures*, vol. 236. Penguin books Harmondsworth, Middlesex, 1956.

[205] W. H. Organization *et al.*, "Epilepsy: a public health imperative," 2019.

[206] Mayoclinic, "Epilepsy-symptoms& causes," 2023.

[207] H. Anwar, Q. U. Khan, N. Nadeem, I. Pervaiz, M. Ali, and F. F. Cheema, "Epileptic seizures," *Discoveries*, vol. 8, no. 2, 2020.

[208] U. R. Acharya, S. L. Oh, Y. Hagiwara, J. H. Tan, and H. Adeli, "Deep convolutional neural network for the automated detection and diagnosis of seizure using eeg signals," *Computers in biology and medicine*, vol. 100, pp. 270–278, 2018.

[209] R. C. Petersen, J. C. Stevens, M. Ganguli, E. G. Tangalos, J. L. Cummings, and S. T. DeKosky, "Practice parameter: Early detection of dementia: Mild cognitive impairment (an evidence-based review)[retired]: Report of the quality standards subcommittee of the american academy of neurology," *Neurology*, vol. 56, no. 9, pp. 1133–1142, 2001.

[210] W. Pravdich-Neminsky, "Ein versuch der registrierung der elektrischen gehirnerscheinungen," *Zentralbl Physiol*, vol. 27, pp. 951–960, 1912.

[211] H. Viles, "'unswept stone, besmeer'd by sluttish time': Air pollution and building stone decay in oxford, 1790-1960," *Environment and History*, vol. 2, no. 3, pp. 359–372, 1996.

[212] F. A. Gibbs, H. Davis, and W. G. Lennox, "The electro encephalogram in epilepsy and in conditions of impaired consciousness," *American Journal of EEG Technology*, vol. 8, no. 2, pp. 59–73, 1968.

[213] H. Adeli, Z. Zhou, and N. Dadmehr, "Analysis of eeg records in an epileptic patient using wavelet transform," *Journal of neuroscience methods*, vol. 123, no. 1, pp. 69–87, 2003.

[214] U. R. Acharya, S. V. Sree, G. Swapna, R. J. Martis, and J. S. Suri, "Automated eeg analysis of epilepsy: a review," *Knowledge-Based Systems*, vol. 45, pp. 147–165, 2013.

[215] O. Faust, U. R. Acharya, H. Adeli, and A. Adeli, "Wavelet-based eeg processing for computer-aided seizure detection and epilepsy diagnosis," *Seizure*, vol. 26, pp. 56–64, 2015.

[216] U. R. Acharya, H. Fujita, V. K. Sudarshan, S. Bhat, and J. E. Koh, "Application of entropies for automated diagnosis of epilepsy using eeg signals: A review," *Knowledge-based systems*, vol. 88, pp. 85–96, 2015.

[217] S. B. Xu Minpeng, "Seizure prediction features based on eeg signals," 2023.

[218] A. Bakhshi, A. Harimi, and S. Chalup, "Cytex: Transforming speech to textured images for speech emotion recognition," *Speech Communication*, vol. 139, pp. 62–75, 2022.

[219] W. Zaremba, I. Sutskever, and O. Vinyals, "Recurrent neural network regularization," *arXiv preprint arXiv:1409.2329*, 2014.

[220] G. Schalk, D. J. McFarland, T. Hinterberger, N. Birbaumer, and J. R. Wolpaw, "Bci2000: a general-purpose brain-computer interface (bci) system," *IEEE Transactions on biomedical engineering*, vol. 51, no. 6, pp. 1034–1043, 2004.

[221] C. Huang, X. Shan, Y. Lan, L. Liu, H. Cai, W. Che, Y. Hao, Y. Cheng, and Y. Peng, "A hybrid active contour segmentation method for myocardial d-spect images," *IEEE Access*, vol. 6, pp. 39334–39343, 2018.

[222] C. Huang, Y. Xie, Y. Lan, Y. Hao, F. Chen, Y. Cheng, and Y. Peng, "A new framework for the integrative analytics of intravascular ultrasound and optical coherence tomography images," *IEEE Access*, vol. 6, pp. 36408–36419, 2018.

[223] G. Xu, X. Shen, S. Chen, Y. Zong, C. Zhang, H. Yue, M. Liu, F. Chen, and W. Che, "A deep transfer convolutional neural network framework for eeg signal classification," *IEEE Access*, vol. 7, pp. 112767–112776, 2019.

[224] U. Herwig, P. Satrapi, and C. Schönfeldt-Lecuona, "Using the international 10-20 eeg system for positioning of transcranial magnetic stimulation," *Brain topography*, vol. 16, pp. 95–99, 2003.

[225] V. Shah, E. Von Weltin, S. Lopez, J. R. McHugh, L. Veloso, M. Golmohammadi, I. Obeid, and J. Picone, "The temple university hospital seizure detection corpus," *Frontiers in neuroinformatics*, vol. 12, p. 83, 2018.

[226] H. Cai, Z. Qu, Z. Li, Y. Zhang, X. Hu, and B. Hu, "Feature-level fusion approaches based on multimodal eeg data for depression recognition," *Information Fusion*, vol. 59, pp. 127–138, 2020.

[227] H. Cai, Z. Qu, Z. Li, Y. Zhang, X. Hu, and B. Hu, "Feature-level fusion approaches based on multimodal eeg data for depression recognition," *Information Fusion*, vol. 59, pp. 127–138, 2020.

[228] H. Adeli, Z. Zhou, and N. Dadmehr, "Analysis of eeg records in an epileptic patient using wavelet transform," *Journal of neuroscience methods*, vol. 123, no. 1, pp. 69–87, 2003.

[229] N. McCallan, S. Davidson, K. Y. Ng, P. Biglarbeigi, D. Finlay, B. L. Lan, and J. McLaughlin, "Epileptic multi-seizure type classification using electroencephalogram signals from the temple university hospital seizure corpus: A review," *Expert Systems with Applications*, p. 121040, 2023.

[230] A. Shankar, S. Dandapat, and S. Barma, "Classification of seizure types based on statistical variants and machine learning," in *2021 IEEE 18th India Council International Conference (INDICON)*, pp. 1–6, IEEE, 2021.

[231] G. Jia, H.-K. Lam, and K. Althoefer, "Variable weight algorithm for convolutional neural networks and its applications to classification of seizure phases and types," *Pattern Recognition*, vol. 121, p. 108226, 2022.

[232] N. Sriraam, Y. Temel, S. V. Rao, P. L. Kubben, *et al.*, "A convolutional neural network based framework for classification of seizure types," in *2019 41st annual international conference of the IEEE engineering in medicine and biology society (EMBC)*, pp. 2547–2550, IEEE, 2019.

[233] S. Raghu, N. Sriraam, Y. Temel, S. V. Rao, and P. L. Kubben, "Eeg based multi-class seizure type classification using convolutional neural network and transfer learning," *Neural Networks*, vol. 124, pp. 202–212, 2020.

[234] S. Tang, J. A. Dunnmon, K. Saab, X. Zhang, Q. Huang, F. Dubost, D. L. Rubin, and C. Lee-Messer, "Self-supervised graph neural networks for improved electroencephalographic seizure analysis," *arXiv preprint arXiv:2104.08336*, 2021.

[235] A. Shankar, S. Dandapat, and S. Barma, "Seizure type classification using eeg based on gramian angular field transformation and deep learning," in *2021 43rd Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC)*, pp. 3340–3343, IEEE, 2021.

[236] S. Thundiyil, M. Thungamani, and S. Hariprasad, "Big eeg data images for convolutional neural networks," in *2021 IEEE Signal Processing in Medicine and Biology Symposium (SPMB)*, pp. 1–6, IEEE, 2021.

[237] A. Shankar, S. Dandapat, and S. Barma, "Seizure types classification by generating input images with in-depth features from decomposed eeg signals for deep learning pipeline," *IEEE Journal of Biomedical and Health Informatics*, vol. 26, no. 10, pp. 4903–4912, 2022.

[238] N. Dang, K. Shao, L. Chen, and M. Yang, "Multi-model decision-making seizure types classification based on transfer learning," in *2022 International Symposium on Control Engineering and Robotics (ISCER)*, pp. 192–201, IEEE, 2022.

[239] M. Tohyama, "Chapter 1 - introduction," in *Acoustic Signals and Hearing* (M. Tohyama, ed.), pp. 1–24, Academic Press, 2020.

[240] T. Puech, M. Boussard, A. D'Amato, and G. Millerand, "A fully automated periodicity detection in time series," in *Advanced Analytics and Learning on Temporal Data: 4th ECML PKDD Workshop, AALTD 2019, Würzburg, Germany, September*

*20, 2019, Revised Selected Papers 4*, pp. 43–54, Springer, 2020.

[241] A. Hagen-Zanker, "A computational framework for generalized moving windows and its application to landscape pattern analysis," *International Journal of Applied Earth Observation and Geoinformation*, vol. 44, pp. 205–216, 2016.

[242] M. Zechmeister and M. Kürster, "The generalised lomb-scargle periodogram-a new formalism for the floating-mean and keplerian periodograms," *Astronomy & Astrophysics*, vol. 496, no. 2, pp. 577–584, 2009.

[243] P. Kellaway, "An orderly approach to visual analysis: characteristics of the normal eeg of adults and children," *Clinical practice of clinical electroencephalography*, 1997.

[244] H. Adeli, Z. Zhou, and N. Dadmehr, "Analysis of eeg records in an epileptic patient using wavelet transform," *Journal of neuroscience methods*, vol. 123, no. 1, pp. 69–87, 2003.

[245] U. M. Khaire and R. Dhanalakshmi, "Stability of feature selection algorithm: A review," *Journal of King Saud University-Computer and Information Sciences*, vol. 34, no. 4, pp. 1060–1073, 2022.

[246] I. H. Sarker, "Machine learning: Algorithms, real-world applications and research directions," *SN computer science*, vol. 2, no. 3, p. 160, 2021.

[247] J. N. Acharya, A. J. Hani, P. Thirumala, and T. N. Tsuchida, "American clinical neurophysiology society guideline 3: a proposal for standard montages to be used in clinical eeg," *The Neurodiagnostic Journal*, vol. 56, no. 4, pp. 253–260, 2016.

[248] A. Bakhshi, A. Harimi, and S. Chalup, "Cytex: Transforming speech to textured images for speech emotion recognition," *Speech Communication*, vol. 139, pp. 62–75, 2022.

[249] Q. Li, J. Gao, Z. Zhang, Q. Huang, Y. Wu, and B. Xu, "Distinguishing epileptiform discharges from normal electroencephalograms using adaptive fractal and network analysis: A clinical perspective," *Frontiers in Physiology*, vol. 11, p. 828, 2020.