



Sentiment Analysis of Dialectal Speech: Unveiling Emotions through Deep Learning Models

تحليل مشاعر اللهجة المنطوقة: اكتشاف العواطف من خلال نماذج التعلم العميق

by

KHALED MOHAMED KHALED EZZELDIN

**A thesis submitted in partial fulfilment
of the requirements for the degree of
DOCTOR OF PHILOSOPHY IN COMPUTER SCIENCE**

at

The British University in Dubai

April 2024



The
British University
in Dubai

Sentiment Analysis of Dialectal Speech: Unveiling Emotions through Deep Learning Models

تحليل مشاعر اللهجة المنطوقة: اكتشاف العواطف من خلال نماذج التعلم العميق

by

KHALED MOHAMED KHALED EZZELDIN

A thesis submitted to the Faculty of Engineering and IT
in partial fulfilment of the requirements for the degree of
DOCTOR OF PHILOSOPHY IN COMPUTER SCIENCE

at

The British University in Dubai

April 2024

Thesis Supervisor
Prof. Khaled Shaalan

Approved for award:

Name
Designation

Name
Designation

Name
Designation

Name
Designation

Date: _____

Declaration

I warrant that the content of this research is the direct result of my own work and that any use made in it of published or unpublished copyright material falls within the limits permitted by international copyright conventions.

I understand that a copy of my research will be deposited in the University Library for permanent retention.

I hereby agree that the contents of this dissertation, for which I am the author and copyright holder, may be copied and distributed by The British University in Dubai for the purposes of research, private study, or education, and that The British University in Dubai may recover from purchasers the costs incurred in such copying and distribution, where appropriate.

I understand that the British University in Dubai may make a digital copy available in the institutional repository.

I understand that I may apply to the university to retain the right to withhold or restrict access to my thesis for a period that shall not normally exceed four calendar years from the congregation at which the degree is conferred. The length of the period is specified in the application, along with the precise reasons for making that application.



Signature of the Student

Copyright and Information to Users

The author, whose copyright is declared on the title page of the work, has granted the British University in Dubai the right to lend his/her research work to users of its library as an open source; and to make partial or single copies for educational and research use.

The author has also granted permission to the University to keep or make a digital copy for similar use and for the purposes of preserving the work digitally.

Multiple copies of this work for scholarly purposes may be granted by either the Author, the Registrar, or the Dean.

Copying for financial gain shall only be allowed with the Author's express permission.

Any use of this work in whole or in part shall respect the moral rights of the Author to be acknowledged and to reflect, in good faith and without detriment, the meaning of the content and the original authorship.

Abstract

Dialect Speech Sentiment Analysis is an evolutionary field where machine learning algorithms are utilized to detect emotions in spoken language. However, Arabic, particularly Egyptian Arabic, remains underrepresented, lacking a dedicated speech sentiment database. This thesis introduces a novel dataset specifically created for sentiment and emotion detection in the Egyptian Arabic dialect, generated from publicly available YouTube videos and annotated across seven emotional categories: anger, happiness, sadness, disgust, fear, romantic, and neutrality. The proposed solution involves leveraging a multi-stage machine learning pipeline that first extracts spectral features such as MFCC and mel spectrograms from acoustic speech waves using Fourier transformation. These features are then classified using a range of Deep Learning Models, including convolutional neural networks (CNN), bidirectional long-short-term memory (BI-LSTM), gated recurrent units (GRU), and Artificial Neural Networks (ANNs). A key contribution of this work is the development and evaluation of hybrid Deep Learning Models that combine CNN-BI-LSTM, CNN-GRU, GRU-CNN, GRU-BI-LSTM, and GRU-ANN architectures. The results demonstrate the superiority of the hybrid CNN-BI-LSTM model, achieving an accuracy of 93%, significantly outperforming individual deep-learning models such as CNN (87%) and BI-LSTM (83%). Additionally, the GRU-CNN hybrid model attained a notable accuracy of 90%. These findings establish the robustness and effectiveness of hybrid architectures in enhancing emotion recognition accuracy in Arabic speech data, presenting a novel approach for Arabic dialect sentiment analysis.

Keywords: sentiment analysis, speech recognition, Arabic speech recognition, deep learning methods, hybrid deep learning approaches.

خلاصة

تحليل تعابير الكلمات هو مجال متنام الاطراف يستخدم خوارزميا التعلم الآلي لتحديد المشاعر من الكلام. ومع ذلك، فإن اللغة العربية هي اللغة الأكثر ندرة من حيث نطاق الابحاث، ولا توجد قاعدة بيانات خاصة بالخطاب العربي لاسيما باللغة المصرية. تهدف هذه الأطروحة إلى التعرف على العواطف الإنسانية باستخدام البيانات المنطوقة والكلام باللغة العربية المصرية. وقد تم تجميع هذه المجموعة من البيانات من أشرطة فيديو يوتيوب التي يمكن للجمهور الاطلاع عليها، وجرى شرحها بسبع فئات عاطفية متميزة: الغضب، والسعادة، والحزن، والاشمئزاز، والخوف، والرومانسية، والحياد. وقُسمت إشارة الكلام الصوتي إلى إطارات صغيرة، وأجري تحويل سريع من فورير. واستُعيدت الخصائص ذات الصلة باستخدام معامل التصفير الميل - ترددي (MFCC) وأطياف الميل كسمات طيفية. واستُخدمت في تصنيف السمات المستخرجة نماذج مختلفة للتعلم الآلي، بما في ذلك الشبكة العصبية التعاقبية، والذاكرة القصيرة الأجل الثنائية الاتجاه، والوحدة المتكررة البوّابة، والشبكات الخلقية الاصطناعية. ويقدم البحث تحليلا شاملا لأداء الاعتراف بالعاطفة باستخدام خوارزمية هجينة للتعلم العميق تتألف من CNN-BI-LSTM، و CNN-GRU، و GRU-CNN، و GRU-BI-LSTM، و GRU-ANN. وتم الحصول على أفضل النتائج عن طريق نموذج CNN ونموذج BI-LSTM، حيث بلغت معدلات الدقة 87 في المائة و 83 في المائة على التوالي، باستخدام نهج تعليمي عميق وحيد. غير أن تنفيذ نموذج هجين يجمع بين CNN-BI-LSTM أسفر عن معدل دقة قدره 93 في المائة، وحقق النموذج الهجين في GRU-CNN معدل دقة بنسبة 90 في المائة. وتبين هذه النتائج مدى قوة وفعالية نهج التصنيف المقترحة تعريف عاطفة الكلام باللغة العربية.

الكلمات الأساسية: تحليل المشاعر، تحليل العواطف، التعرف على الكلام، التعرف على الكلام العربي، أساليب التعلم العميق، أساليب التعلم العميق الهجين.

Dedication

In this journey of academic pursuit, the writing of this thesis stands as a bridge between my dreams and their realization. The support and love I have received along this path have been immeasurable, and it is with deep gratitude that I dedicate this work to those who have been pivotal in my journey.

To my late parents, whose memory and teachings continue to guide me, I owe an immeasurable debt of gratitude. Though they are no longer with us, their spirit and lessons remain a beacon in my life, illuminating my path and inspiring me to strive for excellence. Their absence is felt deeply, but their influence endures on every page of this work.

To my children, Mohamed, Marim, and Ahmed, whose love and understanding have been my sanctuary. Your patience and support, especially during times when my academic commitments demanded so much of our shared moments, have been a source of strength and encouragement. This thesis is as much your achievement as it is mine, for your love has been the wind beneath my wings, propelling me forward even in the face of adversity.

Acknowledgements

I want to begin by extending my heartfelt thanks to the Almighty Allah for granting me the capacity and fortitude to complete this dissertation. I am deeply grateful to my thesis advisors, Prof. Khaled Shaalan and Dr. Manar Alkhatib, for their continuous guidance, unwavering patience, and steadfast support throughout my PhD endeavor. Their constructive feedback and insightful remarks played a pivotal role in shaping my research and assisting me in surmounting various challenges.

In addition, I am appreciative of all the doctors and professors who played a role in this journey. I want to express my sincere gratitude to all the individuals who provided their valuable time, shared their experiences, and offered their invaluable advice.

I extend my thanks to each and every one of you for your significant contributions and unwavering support.

Table of Contents

Chapter 1: Introduction	1
1.1 Background of the Study	1
1.2 Limitations of the Previous Research	9
1.3 Challenges and Constraints in Speech Emotions Recognition	10
1.4 The Relationship Between Egyptian Arabic and MSA	12
1.5 Differences Between Egyptian Arabic and MSA	13
1.6 Challenges in the Navigation Between Egyptian Arabic and MSA	15
1.7 Problem Statement	17
1.8 Motivation.....	18
1.9 Research Aim.....	19
1.10 Research Objectives.....	19
1.11 Research Questions	21
1.12 Thesis Structure	21
Chapter 2: Background.....	24
2.1 Sentiment Analysis Overview.....	25
2.1.1 <i>Functions of Sentiment Analysis</i>	29
2.1.2 <i>Sentiment Analysis Applications</i>	30
2.1.3 <i>Sentiment Analysis Challenges</i>	33
2.1.4 <i>Sentiments Analysis on Social Media Platforms</i>	34

2.2 Arabic Sentiment Analyses.....	36
2.2.1 Arabic Language Processing Characteristic.....	38
2.2.2 Challenges in Arabic Language Processing.....	39
2.3.3 Analyzing Related Works in Sentiment Analysis of Arabic.....	45
2.3 Speech Emotion Recognition.....	48
2.3.1 Speech Emotion Recognition Features	48
2.3.2 Algorithms & Techniques of Speech Emotion Recognition	49
2.3.3 Representation Learning Methods for Recognizing Emotions in Speech.....	53
2.3.4 Arabic Speech Emotion Recognition	56
2.4 Speech Recognition Deep Learning and Hybrid Models	58
2.4.1 Tuning the Parameters.....	58
2.5 Deep Learning Approaches.....	60
2.5.1 Convolutional Neural Networks	61
2.5.2 A Bidirectional LSTM	67
2.5.3 Gated Recurrent Unit Networks.....	71
2.5.4 Artificial Neural Network	76
2.6 Hybrid Deep Learning Approaches	79
2.6.1 CNN-Bi-LSTM Hybrid Model	81
2.6.2 CNN-GRU Hybrid Model	84
2.6.3 GRU-BI-LSTM Hybrid Model.....	86
2.6.4 GRU-CNN Hybrid Model	87
2.6.5 GRU-ANN Hybrid Model.....	88

2.7 Conclusion	90
Chapter 3: Literature Review	91
3.1 Emotions in Human Being.....	91
3.1.1 Emotions Depiction.....	91
3.2 Arabic Speech Emotion Recognition Overview	92
3.3 Arabic Emotion Speech Recognition Architecture	93
3.4 Feature Extraction	95
3.5 Neural Network Acoustic Model	99
3.6 Modeling the Language	103
3.7 Decoding	104
3.8 Evaluation Metrics	105
3.9 Learning Curve in Deep Learning Models	108
3.10 Conclusion	109
Chapter 4: Dataset Corpus	110
4.1 Data Collection	110
4.1.1 YouTube as a Source of Emotion-Rich Data	112
4.2 Data Annotation	114
4.2.1 Annotation Process for Egyptian Dialect Speech Recognition.....	116
4.2.2 Annotation.....	119
4.3 Arabic Dataset Corpus	122

4.4 Dataset Splitting.....	126
4.5 Conclusion	127
Chapter 5: Methodology.....	129
5.1 Data Collection	133
5.2 Data Preprocessing Methodology	134
5.2.1 Pre-Processing.....	134
5.2.2 Data Splitting	138
5.2.3 Feature Extraction	141
5.3 Deep Learning Approaches.....	144
5.3.1 Convolutional Neural Network (CNN)	144
5.3.2 Bidirectional Long Short-Term Memory (BI-LSTM).....	147
5.3.3 Artificial Neural Network (ANN)	149
5.3.4 Gated Recurrent Unit (GRU) Networks:	152
5.4 Hybrid Deep Learning Models	154
5.4.1 CNN-Bi-LSTM Hybrid Model	154
5.4.2 CNN-GRU Hybrid Model	156
5.4.3 GRU-Bi-LSTM Hybrid Model.....	159
5.4.4 GRU-CNN Hybrid Model	161
5.4.5 GRU-ANN Hybrid approach.....	164
5.5 Evaluation Metrics	167
5.6 Conclusion	168
Chapter 6: Experiment and Results.....	169

6.1 Operating Environment.....	169
6.1.1 Jupyter Notebook	169
6.1.2 TensorFlow	169
6.1.3 Google Colaboratory (Colab).....	170
6.2 Experiment Setup.....	170
6.2.1 Hyperparameters of the Preprocessing	172
6.2.2 Deep Learning Models.....	172
6.2.3 Hybrid Deep Learning Models	175
6.3 Results and Discussion	180
6.4 Experiments results	181
6.5 Results Analysis.....	248
6.6 Conclusion	250
Chapter 7: Conclusion.....	251
7.1 How Objectives Were Achieved.....	251
7.2 Summary of Contributions.....	252
7.3 Arabic Dialects Corpus	252
7.4 Arabic Speech Recognition.....	253
7.5 Future Work	255
References	257

List of Tables

Table 1 Arabic Speech Emotion in Literature	57
Table 2 Binary Classification.....	107
Table 3 The Emotions Categories.....	118
Table 4 Emotions with their Labels	120
Table 5 Labeling the Filenames	121
Table 6 Sentiment Analysis with Related Emotions.....	124
Table 7 Distribution of Emotions Among Male and Female Speakers	125
Table 8 The Preprocessing Hyperparameters	172
Table 9 Hyperparameters Values For CNN.....	173
Table 10 Hyperparameters Values For BI-LSTM	173
Table 11 Hyperparameters Values For ANN.....	174
Table 12 Hyperparameters Values For GRU.....	175
Table 13 The Hyperparameters For CNN-BI-LSTM	175
Table 14 Hyperparameters Values For CNN-GRU	176
Table 15 GRU-BI-LSTM Hyperparameters Values	178
Table 16 GRU-CNN Hyperparameters Values.....	179
Table 17 GRU-CNN Hyperparameters Values.....	180
Table 18 CNN Performance.....	181
Table 19 The Performance of BI-LSTM Model	187
Table 20 The Performance of ANN on Speech Recognition.....	196
Table 21 The Performance of GRU Model.....	201
Table 22 The Performance of CNN-BI-LSTM Hybrid Model.....	208

Table 23 The CNN-GRU Hybrid Model	214
Table 24 The Performance of GRU-BI-LSTM Model	223
Table 25 The Performance of GRU-CNN Hybrid Model	232
Table 26 The Performance of GRU-ANN Hybrid Model	239

List of Figures

Figure 1 Number of Internet Users in Egypt	36
Figure 2 Arab World Map.....	37
Figure 3 The Increasing Trend in Sentiment Analysis Research for the Arabic Language (El-Masri et al., 2017)	38
Figure 4 Type of Research Works by the Time	46
Figure 5 Neural Network Block Diagram.....	60
Figure 6 CNN Framework	62
Figure 7 Simple Architecture of CNN	63
Figure 8 CNN Architecture.....	64
Figure 9 Spectrogram of Speech Signal.....	65
Figure 10 Inserting CNN Architecture for Speech Recognition.....	66
Figure 11 BI-LSTM Layer Architecture	70
Figure 12 GRU Architecture.....	74
Figure 13 Reset Gate.....	75
Figure 14 Updated Gate	75
Figure 15 CNN-BI-LSTM Architecture	83
Figure 16 CNN-GRU Architecture	85
Figure 17 GRU-BI-LSTM Architecture	86
Figure 18 GRU-CNN Architecture	87
Figure 19 ASR Architecture	94
Figure 20 Feature Extraction Diagram	97
Figure 21 Transferring Audio to Mel-Spectrum.....	98

Figure 22 Transferring Audio to Mel-Spectrum	101
Figure 23 Sentiment Analysis with Distributed Emotions	123
Figure 24 Emotion Distribution for Men and Women Speakers	126
Figure 25 Framework of the Speech Sentiment Recognition	130
Figure 26 The Flowchart of the Methodology	132
Figure 27 Data Pre-Processing for Arabic Speech Recognition	137
Figure 28 Distribution of Emotions in Training Data.....	139
Figure 29 Distribution of Emotions in Validation Data.....	140
Figure 30 Distribution of Emotions in Validation Data.....	140
Figure 31 The Labeling Table.....	141
Figure 32 The Waveform of Happiness Emotion	142
Figure 33 The Waveforms for the SAD Emotion	142
Figure 34 The Waveforms for the Disgust Emotion.....	143
Figure 35 The Waveforms for the Romantic Emotion	143
Figure 36 The Waveforms for the Neutral Emotion	143
Figure 37 The Waveforms for the Fear Emotion.....	144
Figure 38 The Waveforms for the Anger Emotion	144
Figure 39 CNN Architecture.....	146
Figure 40 CNN Performance	182
Figure 41 The Confusion Matrix for The CNN Model.....	184
Figure 42 The Training and Validation Losses of CNN Model	186
Figure 43 The Training and Validation Accuracy of CNN Model	186
Figure 45 The Performance of BI-LSTM Model.....	188

Figure 45 The Confusion Matrix for BI-LSTM.....	191
Figure 46 The Training and Validation Losses In BI-LSTM	195
Figure 47 The Training and Validation Accuracy In BI-LSTM.....	195
Figure 48 ANN Performance	197
Figure 49 ANN Confusion Matrix.....	199
Figure 50 The Losses in The Train and Validation Using ANN Model.....	200
Figure 51 The Accuracy in the Train and Validation Using ANN Model.....	200
Figure 52 GRU Performance	201
Figure 53 GRU Confusion Matrix	205
Figure 54 Training and Validation Losses in GRU	207
Figure 55 Training and Validation Accuracy in GRU.....	208
Figure 56 The Performance of CNN-BI-LSTM Model.....	209
Figure 57 The Performance of CNN-BI-LSTM Model.....	211
Figure 58 The Training and Validation Losses in CNN-BI-LSTM Hybrid Model.....	213
Figure 59 The Training and Validation Accuracy In CNN-BI-LSTM Hybrid Model	214
Figure 60 The CNN-GRU Performance	215
Figure 61 The Confusion Matrix for the CNN-GRU Model	218
Figure 62 The Training and Validation Losses In CNN-BI-LSTM Hybrid Model.....	221
Figure 63 The Training and Validation Accuracy In CNN-BI-LSTM Hybrid Model	222
Figure 64 The Performance Of GRU-BI-LSTM Hybrid Model.....	223
Figure 65 The Confusion Matrix for the GRU-BI-LSTM Hybrid Model	227
Figure 66 The Training and Validation Losses Curve in the GRU-BI-LSTM Hybrid Model ...	231
Figure 67 The Training and Validation Accuracy Curve for GRU-BI-LSTM Hybrid Model ...	232

Figure 68 The Performance of the GRU-CNN Hybrid Model	233
Figure 69 The Confusion Matrix for the GRU-CNN.....	236
Figure 70 Training and Validation Accuracy Curve in GRU-CNN Hybrid Model	238
Figure 71 Training and Validation Losses Curve in GRU-CNN Hybrid Model.....	239
Figure 72 The Performance Report Of GRU-ANN Hybrid Model	240
Figure 73 The Confusion Matrix of the GRU-ANN Hybrid Model.....	242
Figure 74 The Training and Validation Losses Curve for the GRU-ANN Hybrid Model	244
Figure 75 The Training and Validation Accuracy Curve for the GRU-ANN Hybrid Model.....	245
Figure 76 The Accuracy in the Deep Learning Models Model	246

Chapter 1: Introduction

1.1 Background of the Study

Over time, surveys have served as the primary means of gathering and investigating perspectives on a specific subject. The conventional approach involves selecting a sample of participants and using a structured questionnaire to assess opinions (Rubin and ER, 2009). However, with the rise of social media and widespread Internet access, such as through platforms like YouTube and Facebook, the survey method has undergone changes. Social media has become a significant factor in user interactions, providing a new avenue for understanding and analyzing people's opinions, especially regarding popular topics or products. This novel approach is known as Sentiment Analysis (SA) and will be the focus of this thesis.

Individuals can convey their viewpoints on diverse subjects through emotions. Emotions hold a significant role in both personal and social spheres, providing a means for individuals to communicate through diverse channels and modalities. In the 20th century, advancements in experimental psychology and neuroscience led to the development of various emotion models, grounded in empirical testing and fundamental concepts. Scientific insights reveal that emotion influences multiple domains of animal and human cognition and behavior, such as attention, memory, decision-making, and action. Consequently, the exploration of different facets of emotion across various disciplines has given rise to an interdisciplinary field known as "affective science" (Sander and Scherer, 2009). Affective Science draws on insights from psychology, philosophy, neuroscience, sociology, computer science, economics, literature, and history (Coppin and Sander, 2021) to comprehend, model, measure, and predict affective reactions.

In Darwin's seminal work (Darwin, 1998; Ekman and Friesen, 1971), it is asserted that specific emotional states are universally expressed by people worldwide, even in isolated regions

with minimal external influence. This suggests that these emotional expressions are unlikely to be learned behaviors or influenced by cultural factors (LeDoux, 2012). However, cultural differences can impact emotions in diverse ways, affecting aspects such as the form of expression and intensity (Lim, 2016). Philosophers, dating back to the era of Socrates and spanning subsequent centuries, have engaged in debates about emotions (Deonna and Teroni, 2012). Recent philosophical discussions on emotions have primarily focused on understanding others without direct access to their thoughts or feelings, a quandary known as the "problem of other minds." This issue is explored from two perspectives: the conceptual, addressing reliable conceptualizations, and the epistemological, examining the accessibility of another person's mental states.

In the contemporary era, marked by advancements across diverse technological domains and the evolution of modern living, human existence and the surrounding environment have become increasingly sophisticated. Machines, serving as integral components of this new lifestyle, play a pivotal role in various applications that involve intricate interactions with humans. Consequently, the recognition of emotions emerges as a fundamental requirement for fostering seamless interaction and communication between humans and machines. This parallels the essential role of affective information in human-to-human interactions, underscoring the significance of emotion recognition in the broader landscape of technological and interpersonal engagement (Tzirakis et al., 2017).

Humans perceive and interpret emotions through diverse channels, with the human brain integrating information from various modalities to comprehend the underlying emotional states. Mehrabian's communication model ((Mehrabian, 1969) emphasizes the significance of non-verbal elements in personal communication, attributing 55% of the message to facial expression

and body language, 38% to voice intonation, and 7% to spoken words (Mehrabian, 1981)(Mehrabian, 2017). This model underscores the crucial role played by non-verbal cues in conveying and understanding feelings and attitudes. In recent decades, advancements in affect recognition have arisen from the integration of multiple modalities, including physiological signals (Maaoui et al., 2010), speech signals (Petridis et al., 2018); (Zhang et al., 2016), video (Kaya et al., 2017), hand gestures (Noroozi et al., 2019), and body movement (Daoudi et al., 2017). These developments have found applications in various domains such as safe driving, security, and mobile health (Marechal et al., 2019). The fusion of modalities has significantly contributed to the progress of affect recognition, expanding its utility and impact.

The Arabic language holds a prominent position globally, being one of the five key languages and the official language in approximately 25 countries (Zerari et al., 2019). With over 400 million native speakers and an additional more than 270 million individuals who speak it as a second language, Arabic ranks as the fourth most spoken language, following Mandarin, Spanish, and English (Awajan and Mohamad, 2018). Its significance is further underscored as the language of the Quran, the holy book of Islam, with 1.8 billion Muslims globally in 2015, a number projected to reach 3 billion by 2060 (“Why Muslims are the world’s fastest-growing religious group,” 2017).

Despite its widespread use, Arabic has received comparatively limited attention in speech recognition research compared to other languages (Alkhatib and Shaalan, 2018). Arabic is characterized by its morphological richness, where affixes, both prefixes and suffixes, play a crucial role in word formation (Alkhatib and Shaalan, 2018). Being one of the oldest Semitic languages, Arabic exhibits high variation and is categorized into three types: 1. Classical Arabic, the language of the Quran, Hadith, and ancient Arabic poetry. 2. modern standard Arabic (MSA),

derived from classical Arabic but with modifications in syntax and diacritics. MSA is employed in formal communication, news, and modern literature. 3. Dialectal Arabic, with multiple regional forms used in everyday spoken communication in informal settings. Dialectal Arabic is also written on social media platforms.

Each type of Arabic possesses distinct grammatical, lexical, and morphological standards, posing challenges for the development of Arabic Natural Language Processing (NLP) applications that must handle diverse linguistic variations (Qasim and Abdulbaqi, 2022).

Dialectal Arabic, in particular, presents additional complexities due to informal introductions of prefixes and suffixes during everyday language use, limited availability of text data compared to MSA, and challenges in increasing vocabulary size without encountering out-of-vocabulary issues (Lippmann, 1989).

In the realm of communication, speech stands out as the easiest, fastest, and most natural form of interaction among humans. However, this seamless interaction does not necessarily extend to communication between humans and machines. Despite advancements in voice interaction, the communication remains distant from being truly interactive and natural. One critical factor contributing to this gap is the machines' inability to comprehend human emotions.

Recognizing the significance of emotional understanding in human-machine interaction, researchers, exemplified by (Al-Dujaili and Ebrahimi-Moghadam, 2023) and (Mishra et al., 2023), have delved into methods aimed at efficiently discerning human emotions. This pursuit involves extracting the emotional state of a speaker from their speech, giving rise to the field known as Speech Emotion Recognition (SER). The expressive nature of emotions through speech provides a valuable avenue for extracting meaningful semantics from spoken words, thereby enhancing the efficacy of speech recognition (Kaur and Singh, 2023).

The integration of Speech Emotion Recognition (SER) systems holds significant utility across various applications that necessitate human interaction. For instance, SER can prove beneficial in in-car systems, where it can track the driver's physical condition (Kaur and Singh, 2023). Additionally, call centers can leverage SER to detect and comprehend the emotions of callers (Madanian et al., 2023). Furthermore, SER systems can serve as effective supportive tools in psychological medical diagnosis, particularly for therapists (Landau, 2008).

Speech datasets employed for training and testing SER systems can be categorized into four types based on the methods used to induce emotions: natural, semi-natural, acted, and elicited (Abdel-Hamid, 2020; Abdel-Hamid, 2020). Natural datasets involve emotions extracted from spontaneous speech in TV or radio programs or call centers. Semi-natural datasets are derived from spontaneous reactions to emotions, such as emotional scenes in movies. Acted datasets consist of emotions recorded by professional actors for data collection purposes. Lastly, elicited or induced emotion datasets are created by fabricating artificial emotional situations, with actors involved in these emotional conversations not necessarily being professional actors.

Automatic Speech Recognition (ASR) is the automated process of transcribing speech into text, facilitating machines in comprehending human voice. Over the past few decades, ASR technologies have played a pivotal role in various domains, including education, personal computers, robotics, mobile phones, dictation, military applications, health, and security systems. The significance of ASR lies in its ability to enable seamless communication among people through speech, the simplest form of interaction. Active development and widespread adoption of ASR systems across diverse applications underscore their functionality and simplicity.

One notable application of ASR is in customer care services, where users can engage with voice-enabled systems, reducing reliance on human interactions and addressing service

queues efficiently. Another example is the integration of ASR in smart homes, allowing users to control heating, lighting, and other appliances through voice commands. Within the realm of ASR, Automatic Digit/Command Recognition emerges as a particularly challenging domain, gaining importance with the rise of applications facilitating human-machine interaction through natural language, especially command systems relying on pronounced digits (Nelci et al., 2022).

Speech datasets utilized for training and evaluating Speech Emotion Recognition (SER) systems can be categorized into four types based on the methods used to induce emotions: natural, semi-natural, acted, and elicited (Abdel-Hamid, 2020). Natural datasets involve emotions extracted from spontaneous speech in settings like TV or radio programs and call centers. Semi-natural datasets capture emotions from reactions in unscripted situations, such as emotional scenes in movies. Acted datasets involve emotions recorded by professional actors for the explicit purpose of data collection. Elicited/induced emotion datasets are created by simulating artificial emotional scenarios, with non-professional actors engaging in these emotional conversations.

Given the undeniable success of deep neural networks in addressing complex machine learning challenges, researchers across various fields, including medical sciences, psychology, big data analysis, and paralinguistics, are increasingly turning to these algorithms to find optimal solutions for their specific problems. This active area of research has garnered significant interest from both academia and industry, leading to groundbreaking improvements in practical issues such as computer vision, human-machine interaction, speech, and natural language processing (Ramachandram and Taylor, 2017). In the realm of emotion recognition, the majority of previous work has relied on handcrafted features like scale-invariant feature transform (SIFT) (Zhang et al., 2016) and histogram of oriented gradients (HOG) features (Singh et al., 2017) for facial

expression, as well as Mel-frequency Cepstral coefficients (MFCC), perceptual linear prediction (PLP) coefficients, and supra-segmental features for speech (Ringeval et al., 2015). However, recent research trends lean towards end-to-end optimization with minimal prior human knowledge. This shift is motivated by the understanding that networks can learn an intermediate representation directly from unprocessed input or raw data, a crucial aspect of real-world applications (Graves and Jaitly, 2014).

In contrast to the prevailing trend in the machine learning community, which increasingly adopts deep learning approaches in various domains, including speech recognition for Arabic language (Abdel-Hamid, (2020); Badshah et al., (2017); Qasim and Abdulbaqi, 2022 Chen et al., 2012), and visual and audiovisual speech recognition (Petridis et al., 2017). While a few instances exist in the context of speech emotion recognition (Sivanagaraja et al., 2017; Trigeorgis et al., 2016) and multimodal emotion recognition (Tzinis and Potamianos, 2017), these approaches often rely on audio-visual information that may not be entirely reliable for accurately discerning genuine emotions.

For Speech Emotion Recognition (SER) systems, various learning methods are proposed by researchers, including Artificial Neural Network (ANN), Long Short-Term Memory (LSTM), Convolutional Neural Network (CNN), and Support Vector Machine (SVM). Convolutional Neural Networks (CNNs), for instance, are deep learning architectures designed to extract a hierarchical representation of input data, facilitating categorization (Hema and Garcia Marquez, 2023). CNNs, originally developed for recognizing two-dimensional shapes, prove effective in utilizing the dimensional information encoded in waveform points (Zhang et al., 2016). The adaptive feature extraction properties of the CNN model make it suitable for applications such as voice emotion recognition and image recognition.

Studying Arabic speech emotion recognition is particularly vital as research in this field is still limited. Despite the increasing relevance of emotion recognition technology, there is a noticeable gap in understanding and addressing the specific nuances of emotions expressed in Arabic speech. The scarcity of comprehensive datasets and research studies focused on Arabic emotion recognition impedes the development of accurate and culturally sensitive systems for this linguistic context. By delving into Arabic speech emotion recognition, researchers can contribute to filling this gap and advancing the state of the art in emotion recognition technology. This endeavor not only addresses the current limitations but also opens up new opportunities for applications tailored to the Arabic-speaking population. It is an essential step towards fostering inclusivity and effectiveness in emotion-aware technologies within the Arabic linguistic and cultural landscape.

In this thesis, we detail our endeavors in constructing a dataset focused on the Egyptian dialect and Modern Standard Arabic. The dataset encompasses seven fundamental emotions frequently assessed in Speech Emotion Recognition (SER) literature, namely anger, happiness, sadness, disgust, and neutral, adhering to established standards (Mustafa et al., 2018). Employing spectral features for extraction, as endorsed for optimal feature extraction in SER systems (Mustafa et al., 2018), the dataset preparation is systematically outlined. Subsequent to feature extraction, three robust deep learning classification methods, namely Convolution Neural Network (CNN) (Etienne et al., 2018), Bi directional Long Short Term Memory (BI-LSTM) (Efanov et al., 2022), and Gated Recurrent Unit (GRU) (Shewalkar et al., 2019) were employed due to their proven efficacy in emotion recognition from speech. Additionally, three hybrid models which are CNN-BI-LSTM (Mishra et al., 2023), CNN-GRU and GRU-BI-LSTM were implemented known for its efficiency in classification and pattern recognition.

1.2 Limitations of the Previous Research

The following are limitations of the previous research:

1. Limited focus on Arabic dialects: Many studies focus on modern standard Arabic (MSA) or specific dialects (e.g., Egyptian, Saudi) but do not address the diversity of Arabic dialects comprehensively. This leads to a limited generalization of emotion recognition models across various Arabic-speaking regions. For example, the study by (Aljuhani et al., 2021) focuses only on the Saudi dialect, while (Abdel-Hamid, 2020) is limited to the Egyptian dialect.

2. Dataset Constraints: Most studies rely on a few well-known datasets like the KSUEmotions dataset (Hifny and Ali, 2019; Meftah et al., 2021), while comprehensive, may not represent the full range of Arabic dialects or emotions. Another example, is the dataset used by (Abdel-Hamid, 2020) is partly natural, which might not capture a diverse range of emotions or reflect real-world scenarios as effectively as larger, more varied datasets.

3. Feature extraction and classifier limitations: Some of the earlier works like (Meftah et al., 2020) and (Aljuhani et al., 2021) rely on traditional machine learning classifiers such as SVM and KNN, which have limited scalability and adaptability when compared to deep learning approaches. While Deep Learning Models like CNN-BLSTM-DNN employed by (Hifny and Ali, 2019) achieve higher accuracies, they are not always accessible due to computational demands, and the studies often do not provide sufficient performance evaluation across different dialects or use cases.

4. Limited use of advanced neural architectures: Even though Hifny and Ali (2019) employ a hybrid CNN-BLSTM-DNN model, more recent advancements in transformer-based models, such as BERT or Wav2Vec2.0, which have demonstrated superior performance in various speech and text recognition tasks, have not been fully explored in Arabic SER.

This study addresses these gaps as illustrated below.

1. **Broader Dialect Coverage:** Our research encompasses a broader spectrum of Arabic dialects beyond MSA, Egyptian, or Saudi dialects, striving for a more thorough comprehension of Arabic emotions across various places.
2. **Improved Dataset Representation:** By assembling and refining a more diversified and balanced dataset encompassing a broad spectrum of real emotions and dialects, our research can rectify the deficiencies of current datasets, which frequently suffer from constraints in size and diversity.
3. **Advanced Deep Learning Models:** Rather than depending on conventional classifiers such as SVM and KNN, our study can utilize more sophisticated neural networks, including transformer-based models (e.g., BERT, Wav2Vec), which have demonstrated substantial advancements in other languages for SER tasks.
4. **Feature Engineering Enhancements:** Our research can investigate advanced feature extraction techniques, including the integration of spectral, prosodic, and wavelet characteristics with deep learning representations to enhance classification performance across various moods and dialects.

By addressing these gaps, our research will offer a more thorough, scalable, and precise solution for Arabic Speech Emotion Recognition, thereby enhancing the state of the art in this field.

1.3 Challenges and Constraints in Speech Emotions Recognition

(Schuller, 2018) identifies SER "in the wild" as a major difficulty in the field in a recent assessment. It should function properly regardless of the speaker's language or cultural background in order to be a suitable SER in the wild solution. The ideal "in the wild" answer, however, is by no means found in the existing literature.

Finding the speaker's true "inner emotion" would be a significant difficulty in SER, according to (Schuller, 2018). The majority of the statistics and assessments available today focus on the emotional state that is visible to the outside world. That being said, the speaker's genuine emotional state may differ from the emotion that is visible to the outside world. He emphasizes once more how important it is to define the emotion tags connected to feelings that can be represented by biological data.

One of the challenges in SER, according to (Anagnostopoulos et al., 2012) and (El Ayadi et al., 2011), is the lack of a universal emotional representation and a precise understanding of emotions. Either the dimensional representation, valence, and arousal (Cowie et al., 2001) or the six fundamental emotions (anger, disgust, fear, happiness, sadness, and surprise) specified by (Ekman, 1992) are used in current SER primary research. Both formats can be used to annotate available corpora, however, some corpora have annotations in both.

Data sparseness has plagued SER research since its beginning and continues to be a major problem for the discipline (Parry et al., 2019) (Zhang et al., 2011). Speech signal characteristics can change depending on a number of variables, such as the language spoken by the speakers, their cultural background, and the recording's surroundings. Moreover, it is expensive to create an annotated set of emotional voice data. Because of this, the existing corpora of emotional speech are very variable among themselves and do not fully capture the range of speech features.

The fact that different SER corpora are available further contributes to the difficulty of SER in the wild. A major restriction in the SER area, as noted (Cummins and Schuller, 2019) in a recent review, is the absence of actual labelled data. Since most datasets are professionally acted, the emotions are mimicked. Despite the fact that acted facts may be viewed as usual, they lack the nuanced information connected to the natural emotion produced by the neurological system

(Li et al., 2015) (Anagnostopoulos et al., 2012). Additionally, it has been discovered that the robustness of natural speech is adversely affected when acted databases are used in systems (Madanian et al., 2023) and (Li et al., 2015)).

In addition, a collection of audio files undergoes a tagging process to produce the annotated corpora, irrespective of the type of gathering, acted, evoked, or natural emotions. A group of assessors hears the speech and assigns a label or a continuous value spanning all emotional qualities. On the other hand, evaluators don't always give the same continuous value or label. Averaging continuous values or using a majority vote are two methods for assigning emotion labels. Due to their large variance in the ratings given by the evaluators, these inconsistencies add noise to the ground truth. The impact of this labelling inconsistency on SER tasks has been studied by (Dang et al., 2017). According to the study, when there is less uncertainty offered, emotion prediction is more reliable, and vice versa. One known constraint in the existing SER corpora is the inconsistent evaluators.

A speech signal is a very rich source of information. Finding the ideal combination of data that is associated with emotion is generally another difficult task. Low resource demanding models are necessary for consumer applications, and this can be related to the difficulty of determining the ideal set of emotional information (Cummins and Schuller, 2019).

1.4 The Relationship Between Egyptian Arabic and MSA

MSA represents a linguistic progression from Classical Arabic, the parlance applied in the Quran and classical literary works, and serves as the official norm across numerous Arab countries. MSA is predominantly employed in written discourse, official documentation, news telecasts, and formal orations. Conversely, Egyptian Arabic has developed from Classical Arabic but has been markedly shaped by historical interactions with a multitude of other languages,

including Coptic, Turkish, Italian, French, and English, among others. Egyptian Arabic is chiefly utilized in quotidian spoken exchanges, as well as in television productions, cinematic works, and musical compositions.

The interplay between Egyptian Arabic and modern standard Arabic (MSA) is characterized by both continuity and divergence. MSA functions as a shared linguistic framework that integrates Arabic speakers across diverse geographical regions, whereas Egyptian Arabic embodies the distinct socio-cultural identity of Egypt. Despite the fact that both dialects originate from a common linguistic source and possess numerous foundational linguistic attributes, the discrepancies are sufficiently pronounced to pose difficulties for mutual intelligibility, particularly for individuals who are non-native speakers or learners of the Arabic language.

1.5 Differences Between Egyptian Arabic and MSA

Below are the differences between Egyptian and Arabic MSA:

1. Phonological Differences. A significant differentiation between Egyptian Arabic and MSA is evident in their respective phonetic articulations. For instance:

- In MSA, the letter "ج" is pronounced as /dʒ/ (like the English "j" in "judge"), while in Egyptian Arabic, it is often pronounced as /g/ (like the "g" in "go"). For example, the word "جمل" (camel) is pronounced as "jamal" in MSA but "gamal" in Egyptian Arabic.
- The letter "ق" is pronounced as /q/ (a voiceless uvular plosive) in MSA but is often replaced with a glottal stop /ʔ/ in Egyptian Arabic. For example, the word "قلب" (heart) is pronounced as "qalb" in MSA but "ʔalb" in Egyptian Arabic.
- Vowel length and stress patterns also differ, which can affect the intelligibility of words. Egyptian Arabic tends to have a more simplified vowel system compared to the richer vowel inventory in MSA.

2. *Morphological and Syntactic Differences.* Morphology and syntax in Egyptian Arabic have diverged significantly from MSA:

- Verb conjugation in Egyptian Arabic is simpler than in MSA. For example, in MSA, the past tense of the verb "to write" (كتب) for "I wrote" is "katabtu," whereas in Egyptian Arabic, it is simply "katabt."
- The definite article in MSA is "ال" (al-), which is also used in Egyptian Arabic, but the pronunciation and assimilation rules may differ. In Egyptian Arabic, certain consonants after "al-" cause the "l" sound to assimilate, which is similar to MSA but often applied more liberally.
- Word order in sentences can differ. While both MSA and Egyptian Arabic typically follow a VSO (Verb-Subject-Object) order in formal contexts, Egyptian Arabic often prefers an SVO (Subject-Verb-Object) structure in informal speech. For example, In MSA the sentence "ذهب الولد الي المدرسة" (went-the-boy-to-the-school), whereas in Egyptian Arabic, it is "الولد راح المدرسة" (the-boy-went-to-the-school).

3. *Lexical Differences.* The vocabulary between Egyptian Arabic and MSA differs significantly:

- Egyptian Arabic incorporates a variety of loanwords from languages like Coptic, Turkish, French, and English. For example, "ترابيزة" (tarabeiza) for "table" comes from Italian "tavola," whereas in MSA, "table" is "طولة" (ṭāwīlah).
- Certain everyday words are entirely different. For example, "now" is "الآن" (al-ān) in MSA but "دلوقتي" (dilwaʔti) in Egyptian Arabic.

4. *Semantic Differences.* Some words may exist in both MSA and Egyptian Arabic but carry different connotations or meanings:

- The word "ساعة" (sa'ah) means "hour" in both MSA and Egyptian Arabic but can also mean "watch" or "clock" in Egyptian Arabic.
- "عالم" (‘ālam) means "world" in MSA, while in Egyptian Arabic, it could also imply "people" or "crowd."

5. *Pragmatic and Cultural Differences*. Pragmatics and contextual usage frequently dictate the specific variety of Arabic that is employed. MSA is regarded as the "high" register utilized in formal contexts, whereas Egyptian Arabic is considered the "low" register appropriate for informal, quotidian exchanges. The application of MSA may imply a sense of formality, educational attainment, or authoritative presence, while Egyptian Arabic denotes familiarity and is typically employed in narrative discourse, humour, and routine interpersonal communications.

1.6 Challenges in the Navigation Between Egyptian Arabic and MSA

Below is the list of challenges in navigating between Egyptian Arabic and MSA:

1. Educational Challenges

Teaching Arabic as a foreign parlance involves deciding whether to concentrate on MSA, that has wider official applicability, or on a vernacular like Egyptian Arabic, which is more frequently oral and might be increased immediately effective in certain contexts. The divergence within MSA and Egyptian Arabic can be confusing for learners who have been exposed to only one form.

2. Translation and Interpretation Challenges

The process of translation through Egyptian Arabic and MSA is often fraught with difficulties owing to substantial lexical, syntactic, and semantic divergences. The informal register and vernacular idioms prevalent in Egyptian Arabic frequently lack precise counterparts within the more formalized structure of MSA.

3. Media and Communication Challenges

Media institutions in Egypt frequently amalgamate modern standard Arabic (MSA) with Egyptian Arabic, particularly within the contexts of television programming, news broadcasts, and commercial advertisements. This convergence may engender a diglossic environment wherein viewers are anticipated to transition fluidly between the two linguistic registers, thereby presenting difficulties for individuals who lack familiarity with either variant.

4. Technological Challenges in Natural Language Processing (NLP)

The distinctions between modern standard Arabic (MSA) and Egyptian Arabic pose particular obstacles for Natural Language Processing (NLP) applications, including but not limited to speech recognition, sentiment analysis, and machine translation. Algorithms developed utilizing MSA datasets may encounter significant difficulties in effectively processing Egyptian Arabic as a result of variances in lexicon, syntactic framework, and phonetic articulation.

The distinctions between modern standard Arabic (MSA) and Egyptian Arabic pose particular obstacles for Natural Language Processing (NLP) applications, including but not limited to speech recognition, sentiment analysis, and machine translation. Algorithms developed utilizing MSA datasets may encounter significant difficulties in effectively processing Egyptian Arabic as a result of variances in lexicon, syntactic structure, and phonetic articulation.

5. Social and Political Challenges

There is a debate among Arabic speakers and scholars regarding the status of dialects versus MSA. Some argue that promoting dialects could lead to fragmentation of the Arabic language, while others see dialects as living languages that reflect the cultural and social identities of their speakers.

The relationship between Egyptian Arabic and Modern Standard Arabic is a testament to the linguistic diversity within the Arabic language. While MSA provides a unifying standard for formal communication, Egyptian Arabic offers a rich, vibrant medium for everyday interaction and cultural expression. Understanding the differences and challenges between the two is crucial for effective communication, education, and technology development in the Arabic-speaking world. Balancing the use of MSA and Egyptian Arabic remains an ongoing challenge that requires careful consideration of linguistic, cultural, and social factors.

1.7 Problem Statement

Within the domain of human-machine interaction, the ability to comprehend and correctly interpret human emotions is crucial for facilitating successful communication and involvement. Speech, as a primary form of human communication, contains important emotional signals that, when understood, may greatly improve the quality of interactions between people and computers. Although significant advancements have been achieved in Speech Emotion Recognition (SER) for other languages, there remains a noticeable deficiency in research and development specifically for the Arabic language.

Arabic, being a prominent global language, is spoken by millions in many places and has significant cultural importance. Although Arabic Speech Emotion Recognition is widespread and significant, there is a lack of thorough studies and resources dedicated to this field. The current Speech Emotion Recognition (SER) systems mostly prioritize languages such as English, German, and Spanish, resulting in a lack of representation for the Arabic language in this crucial field.

The issue arises from the distinct features of the Arabic language, which is not only linguistically complex but also displays significant diversity in various dialects. modern standard

Arabic (MSA), the official language of communication, exists alongside other spoken dialects, each exhibiting unique linguistic characteristics. The wide range of languages spoken is a significant challenge to the creation of reliable and precise Arabic Speech Emotion Recognition (SER) systems.

Furthermore, the limited availability of annotated speech emotion datasets for Arabic exacerbates the issue. The current datasets lack sufficient representation of the subtleties in Arabic emotional expressions, and their restricted availability hampers the training and assessment of models for Speech Emotion Recognition (SER). This inadequacy becomes more evident when considering the cultural milieu, where emotions may be articulated in distinct ways, affected by cultural norms and subtle subtleties specific to Arabic-speaking groups.

Alleviating the deficiency in Arabic oral communication Emotion recognition is not just a language obstacle, but also a societal necessity. Acquiring a thorough comprehension of Arabic emotional expressions is crucial for a wide range of applications, including human-machine interaction in smart homes and psychological medical diagnostics. Hence, there is a pressing requirement for research initiatives that focus on creating strong, culturally attuned, and linguistically varied Arabic Speech Emotion Recognition (SER) systems. This will help bridge the current gap and contribute to the wider field of emotion-aware artificial intelligence.

1.8 Motivation

The rapid progress of technology has made it possible to accurately detect human emotions using technologies. With the expectation that machines would eventually outperform humans in this field, this advancement is positioned to enable a significant level of seamless communication between people and machines. Automatic emotion recognition systems have the potential to be applied in several research fields, such as diagnosing psychological states,

enhancing smart gaming experiences, detecting deception, improving the efficiency of contact centres, and enhancing educational programs.

1.9 Research Aim

In the realm of NLP, the capacity to discern emotions conveyed through speech is essential for improving the capabilities of automated systems. This work seeks to comprehensively examine Arabic Speech Emotion Recognition (SER), considering the diverse cultural and linguistic variations among the Arabic-speaking community. The main goal is to develop robust models that both understand the intricacies of emotional expressions and account for the various linguistic differences present in modern standard Arabic (MSA) and its spoken dialects.

1.10 Research Objectives

The main target of this study is to achieve the following objectives:

Development of a corpus for speech analysis: The primary goal is to create a novel Arabic Speech Emotion Corpus that surpasses the constraints of current datasets. This corpus seeks to provide a culturally appropriate and varied base for training and assessing models that recognize and analyze a wide range of emotions, including happiness, sorrow, anger, contempt, fear, romance, and neutrality. The primary focus is to capture genuine emotional emotions in various circumstances and language variants.

Feature Extraction Strategies: In order to fully utilize the expressive capabilities of Arabic speech, the study focuses on developing feature extraction techniques specifically designed for the distinct qualities of the language. The study will investigate linguistic, prosodic, and spectral characteristics, focusing on identifying unique aspects that are distinctive to

emotional patterns in Arabic. The purpose of this aim is to provide the foundation for the accurate portrayal of emotional signals in the speech signal.

Deep Learning Models: The study intends to utilize and assess advanced deep learning structures for Arabic Speech Emotion Recognition, going beyond traditional methods. The ability of Convolutional Neural Networks (CNNs), Gated Recurrent Unit (GRU), and Long Short-Term Memory networks (LSTMs) to identify temporal relationships and hierarchical characteristics in Arabic audio data will be examined.

Hybrid Models for Improved Performance: This objective is to investigate the possible synergies between conventional machine learning and deep learning by utilizing hybrid models. The study aims to combine feature-based algorithms and deep neural networks in order to leverage the advantages of both approaches, hence improving the accuracy and resilience of Arabic SER systems. The study will analyze CNN-BI_LSTM, CNN-GRU, and GRU-BI-LSTM hybrid models to assess their capacity to detect temporal relationships and hierarchical features in Arabic speech data. The research examines the language variety within Arabic and investigates how dialectal variables affect SER performance, emphasizing the need of cross-dialect analysis and cultural sensitivity. Developing strategies to manage these variances and guarantee cultural sensitivity in the detection of emotions are crucial components of this goal.

The use of benchmarking criteria, adherence to standards, and the dedication to open-source contributions are crucial for promoting cooperation and furthering research on Arabic speech recognition. This study not only paves the way for significant advancements in emotion identification but also provides essential resources to the wider academic and scientific community.

1.11 Research Questions

The study aims to address the following questions:

- 1) What methods may be used to effectively classify the sentiment of Arabic Speech emotion recognition?
- 2) What is the effect of applying sentiment analysis algorithms specifically designed for the Arabic language on the accuracy of classification?
- 3) Can the effectiveness of sentiment analysis in Arabic speech emotion recognition be enhanced by utilizing hybrid Deep Learning Models?
- 4) What is the distribution of sentiment categorization percentages in the speech?

1.12 Thesis Structure

Chapter 2 offers a thorough overview of the background information about the main topics covered in this study. This text provides a comprehensive analysis of the theoretical aspects of Arabic scripts, including a detailed examination of Arabic letters, diacritics, and the Arabic keyboard. In addition, a succinct summary of Modern Standard Arabic is presented, emphasizing its distinctive traits and essential elements. The chapter also explores Dialectic Arabic, which includes variations in Arabic phonology. Moreover, this chapter presents the issues related to Arabic dialects. This chapter presents the important features, which are then summarized in the conclusion. It offers a comprehensive introduction to Arabic Speech Emotion Recognition, encompassing topics such as the nature of human emotions and their many forms of expression. The chapter delves into the topics of feature extraction, acoustic modelling, and decoding approaches. Additionally, it presents the assessment measures used to analyze the performance of Arabic Speech Emotion Recognition systems. The chapter concludes with a recapitulation of the key elements that were explored. This chapter presents Speech Recognition

Deep Learning Models, explaining deep learning techniques such as Convolutional Neural Networks (CNN), Bidirectional Long Short-Term Memory (BI-LSTM), and Gated Recurrent Units (GRU). The chapter further presents an exposition of hybrid Deep Learning Models, including CNN-BI-LSTM, CNN-GRU, and GRU-BI-LSTM. It concludes by summarizing the key themes discussed.

Chapter 3 presents a comprehensive literature overview on sentiment analysis, including topics such as sentiment classification, different levels of sentiment analysis, the purposes of sentiment analysis, and its applications across diverse domains, including social media platforms. The chapter also explores research that specifically examines Arabic Sentiment Analysis. Moreover, it delves into the investigation of speech emotion detection. The chapter finishes by providing a concise overview of the main discoveries derived from the examined literature. A conclusion succinctly outlines the fundamental aspects of this chapter.

Chapter 4 provides a comprehensive description of a dataset corpus, including the methodology for collecting data, the use of YouTube as a source of data with high emotional content, and the process of annotating the dataset. The chapter further presents the Arabic dataset corpus and delineates the process for dataset partitioning. The chapter finishes by providing a concise overview of the main topics discussed.

Chapter 5 provides a detailed explanation of the methods used to accomplish the goals of this thesis. The chapter introduces the data preparation portion, providing an overview of the processes required to prepare the data, and ends with a summary. This includes the process of gathering data, preparing the data for analysis, utilizing deep learning techniques, and employing a combination of deep learning approaches. The chapter presents evaluation measures and finishes by providing a concise summary of the main themes discussed.

Chapter 6 provides a thorough explanation of the implementation and experimentation methods. The operating environment is defined to include Jupyter Notebook, TensorFlow, and Google Colaboratory. The comprehensive description of the experimental setup encompasses the processes of data collecting and preprocessing, the implementation of Deep Learning Models, and the utilization of hybrid deep learning techniques. The chapter finishes by providing a concise overview. The chapter also presents the results and arguments arising from the use of algorithms to assess the effectiveness of the Speech Emotion Recognition (SER) model. The findings are thoroughly examined and extensively deliberated about. Moreover, this chapter culminates by providing a concise overview of the thesis and delineates prospective directions for future research. Finally the chapter represents the last stage of this thesis, providing a thorough and inclusive conclusion that covers several important components. The chapter starts with a concise overview of the contributions made throughout the investigation. Subsequently, the ramifications of the discoveries for Arabic dialects emerge. Therefore, the final outcome of the Arabic speech recognition is expressed. The chapter concludes by presenting a prospective viewpoint, delineating the potential future endeavours that might expand upon the existing research.

Chapter 2: Background

This section provides an overview of the existing research and fundamental studies in sentiment analysis, specifically focusing on the analysis of sentiment in the Arabic language. The text provides a clear explanation of the classification of sentiment analysis, including its many phases, tasks, and related challenges. The chapter further addresses the process of feature selection and provides insight into sentiment analysis within the realm of social media. Furthermore, it highlights the methodologies used to analyze sentiment in social network data and the many aspects of sentiment analysis investigated in academic literature. Finally, the part explores Arabic sentiment analysis, including linguistic characteristics, challenges, and academic contributions.

The chapter explores the subject of human emotions. The text covers several topics, including the representation of emotions, the identification of emotions in speech, the factors involved in speech emotion detection, and the methodologies and algorithms employed in this procedure. Furthermore, the chapter provides a comprehensive examination of the difficulties and limitations linked to this field of study, along with a detailed elucidation of representation learning methods employed for the identification of emotions in speech.

Arabic, an integral constituent of the Semitic parlances with a rich historical lineage, is characterized by a complex diglossic system, in which modern standard Arabic (MSA) and varied territorial dialects coexist simultaneously. Egyptian Arabic, colloquially known as "Masri," ranks among the majority prevalently utilized dialects within the Arab region and exhibits substantial divergence from MSA, which performs as the formal variant of the terminology employed in media, literature, and official discourse. The interplay among Egyptian Arabic and MSA embodies both linguistic diversity and complexities, particularly concerning

grammatical structure, lexicon, phonetic realization, and contextual application. Grasping these distinctions is imperative for linguists, educators, and language learners as they navigate the intricate landscape of Arabic linguistics.

2.1 Sentiment Analysis Overview

Before exploring sentiment analysis in any given language, it is essential to establish a mutual comprehension of the term "sentiment." Within the literature, several names are employed interchangeably to denote this notion, including opinion, subjectivity in text, feeling, emotion, appraisal, belief, and conjecture. Each of these phrases refers to an individual's subjective condition that cannot be objectively witnessed or confirmed (Birjali et al., 2021). The varied nomenclature used in the field of computational analysis includes terms such as opinion mining, sentiment analysis, or subjectivity analysis (Alkhatib and Shaalan, 2017) and (Pang and Lee, n.d.), which can sometimes cause confusion for readers or beginners in this domain.

Textual information may generally be classified into two distinct categories. The initial category comprises factual information, which just conveys objective facts regarding entities or occurrences. The second type pertains to subjective information, which discloses the writer's genuine emotions or viewpoints on things and occurrences. In this dissertation, sentiment is defined as the subjective information that represents an individual's emotions or opinions on a certain topic or issue (Turney, 2002).

Sentiment analysis is a technique used to discern the feelings or opinions of individuals towards a specific subject. This field is encompassed within the domains of machine learning, natural language processing, and computational linguistics. Essentially, it aims to assess and extract people's sentiments from their written expressions.

Sentiment analysis is a method employed to ascertain the emotions or viewpoints of persons regarding a certain issue. This field is within the purview of machine learning, natural language processing, and computational linguistics. Its primary objective is to evaluate and extract individuals' feelings from their written utterances.

In addition, the mood expressed in a writing might be either clearly stated or indirectly implied. In overt instances, the language explicitly expresses a feeling, as exemplified by the phrase "This is a splendid panorama," "يا له من منظر جميل" whereas in covert instances, the text suggests a sentiment, as demonstrated by the statement "The vacation lasts just for two days." "العطلة لمدة يومين فقط." Sentiment analysis may be technically defined as the process of evaluating and categorizing the emotions and opinions expressed in a given text. The goal is to use computational methods to assign polarity labels p from a set of polarities P to a given text t , in a way that appropriately represents the overall prevalent polarity within the collection of texts (Pang and Lee, n.d.). The first stage in sentiment analysis is to ascertain if the text's content is subjective or objective, which is known as subjectivity categorization (Pang and Lee, n.d.). Subsequently, the subjective language is examined to determine the precise sentiment polarity it conveys (Pang and Lee, n.d.). The magnitude of this polarity might fluctuate between different viewpoints. For example, when it comes to customer evaluations of a product, it may be necessary to categorize them as either positive or negative, demonstrating a binary polarity. The work becomes further intricate when the polarity is broadened to encompass more than two categories, such as the inclusion of a neutral class. Another kind of sentiment analysis pertains to emotions such as Sorry, Hugs, You Rock, Wow, and others (Socher et al., 2011), which poses a problem in classifying many classes.

The classification process can be applied at different levels within the text: individual terms, phrases, sentences, or entire documents. The results obtained at each level typically serve as input for the subsequent level. For instance, the output from evaluating sentences is utilized and extended for document classification. In our work, we will focus on identifying sentiment at both the sentence and document levels.

The categorization process can be implemented at several levels within the text, including individual terms, phrases, sentences, or complete texts. The outcomes achieved at each stage often function as input for the following stage. For example, the results obtained from analyzing sentences are used and expanded upon for the purpose of categorizing documents. Our efforts will primarily concentrate on discerning sentiment at both the sentence and document levels.

Another variant of sentiment analysis entails discerning the sentiment target or ascertaining the specific object or subject towards which the sentiment is directed. The majority of previous research in sentiment analysis concentrates on evaluating sentiments associated with a broad subject or objective, such as user critiques of a film or product. Identifying the topic in such circumstances is uncomplicated, as it is presumed that the evaluation is related to a certain product. However, in the context of feature-based sentiment analysis, accurately identifying the specific parts of the product that the user is referring to and determining their opinion becomes increasingly difficult when the target is unfamiliar. As a result, the first step is to determine the specific characteristics that a user is discussing, which is often done using feature extraction techniques (Etzioni et al., 2005). Afterwards, the mood or attitude about these traits is assessed. This dissertation does not discuss the specific process in sentiment analysis that is being referred to.

Sentiment analysis, or opinion mining, seeks to analyze individuals' perspectives on a wide range of themes, including individuals, topics, issues, services, organizations, goods, events, and their associated characteristics. This domain comprises the examination and analysis of assessments, attitudes, sentiments, views, emotions, and appraisals, making it a multifaceted area of inquiry.

Sentiment analysis, which is sometimes referred to as opinion extraction, subjectivity analysis, emotion analysis, opinion mining, sentiment mining, review mining, or affect analysis, is a subfield of either sentiment analysis or opinion mining. This subject is a primary emphasis in applied linguistics, and its importance extends to different fields such as political science, education, and marketing. Opinion mining, also known as sentiment analysis, is the process of extracting insights from given materials by recognizing negative, positive, or neutral opinions (Mohammed, 2016). The extraction procedure depends on fields such as machine learning, natural language processing, and statistics to ascertain the polarity of a certain record (Awajan and Mohamad, 2018). The feedback, critique, or remarks acquired through this procedure might be a helpful indication for several purposes.

In addition, in order to automate the classification of a document, a collection of categories X is required. Every category corresponds to a unique field of study, denoted as

$$X = \{x_1, x_2, x_3, x_n\} \quad (1)$$

Here, 'n' represents the total number of categories in C . In addition, A is defined as a collection of records within a set as A (Prabowo and Thelwall, 2009). This collection is expressed as in Eq. 2:

$$A = \{a_1, a_2, a_3, a_m\} \quad (2)$$

Let 'a' represent the total number of records in the collection. Automated categorization is the process of using a classifier algorithm to allocate a record to a specified category. The findings provide pairs, where each pair consists of a record 'xi' and a category 'aj', where $\{x_i, a_j\} \in R \times C$. The symbols signify that the element 'ri' belonging to the set of real numbers (R) is categorized or allocated to the element 'xj' belonging to the set A, as stated by (Prabowo and Thelwall, 2009).

2.1.1 Functions of Sentiment Analysis

The tasks of sentiment analysis may be classified into five categories: polarity classification, identifying degrees of arousal or valence on a certain scale, going beyond polarity, separating objectivity/subjectivity, and aspect-based or feature-specific sentiment analysis (Kron et al., 2015). Below is a concise overview of the sentiment analysis function:

1- Classification of the polarity

Polarity classification is the primary focus of sentiment analysis research. It entails evaluating the sentiment of a given viewpoint in a sentence or document with respect to a certain characteristic or attribute of a target. The categorization involves categorizing the sentiment as either neutral, positive, or negative (Kron et al., 2015).

2- A quantification of the level of valence or arousal on a certain scale.

Multiple scales can be utilized to convey the emotion of a particular document. For example, the conventional binary classification job of categorizing something as either bad or good can be extended to include a rating scale, such as four or five stars, as a result of the prediction process (Kron et al., 2015).

3-Precise examination

In order to do a more comprehensive examination of comments, researchers identify subjective words inside phrases, as well as their corresponding objectives and sources. This detailed analysis focuses on the identification of several sorts of entities in comments, such as targets, expressions, sources, holders, elements of the target, as well as geographical and temporal patterns in information (Kron et al., 2015).

4- Identification of objectivity or subjectivity

Determining whether a statement is objective or subjective is a component of sentiment analysis. Although it may appear to be a work of recognition at first, in some settings, it might be seen as an effort of classification. This process of sentiment analysis entails categorizing a given text as either objective or subjective, which might be more complex than the traditional work of categorizing polarity. Subjectivity of a word or term depends on the context. Even a seemingly neutral magazine article might include a person's subjective opinions. Furthermore, the results of this research heavily depend on annotated data that have been identified as subjective. However, it may be beneficial to remove factual statements from a text before determining its emotional tone (Kron et al., 2015). Subjective words or expressions are used to express emotional and cognitive states such as views, feelings, estimates, and opinions (Akkaya et al., 2009).

5- Implement sentiment analysis functionality

Feature sentiment analysis, often referred to as aspect-based sentiment analysis, is a specialized kind of sentiment analysis that concentrates on assessing the sentiment conveyed towards particular elements or characteristics of a product, service, or entity. Feature sentiment analysis differs from general sentiment analysis by specifically identifying sentiments linked to certain qualities or traits, rather than providing an overall sentiment score for a piece of text.

2.1.2 Sentiment Analysis Applications

Sentiment analysis, or opinion mining, is an advanced method that automates the process of identifying and comprehending sentiments, views, and emotions conveyed in written or spoken language. The applications of this technology are extensive and have been quite beneficial in a range of different sectors (Medhat et al., 2014). Sentiment analysis is widely applied in many major areas:

1. Customer Feedback and Market Research: Sentiment analysis aids organizations in comprehending the way their customers view their products, services, and overall brand. Examining client evaluations, polls, and social media comments yields significant knowledge for the enhancement and advancement of products.

2. Social Media Monitoring: Companies employ sentiment analysis to monitor references, remarks, and discussions pertaining to their brand on sites such as Twitter, Facebook, and Instagram. This facilitates the assessment of public sentiment, identification of possible concerns, and the management of reputations.

3. In the field of e-commerce and retail, sentiment analysis is employed to assess customer comments and reviews regarding their items. This data facilitates the management of inventories, the development of pricing strategies, and the identification of prevailing patterns.

4. Brand Reputation Management: Sentiment analysis plays a vital role in the management and preservation of a favourable brand image. It aids in promptly recognizing and resolving unfavourable feelings, enabling firms to properly respond and minimize possible harm.

5. Market intelligence and competitive analysis benefit from sentiment analysis as it allows businesses to obtain valuable insights into their rivals. This entails comprehending customer views of competing items or services, and pinpointing areas where they may possess a competitive advantage.

6. Political Analysis: attitude analysis is used in politics to assess popular attitudes towards political parties, politicians, and policies. This information is vital for election campaigns and policy formulation.

7. Client Support and Service Enhancement: Through the analysis of client feedback, specifically obtained from customer service contacts, firms may pinpoint areas in need of development within their support procedures. It assists in resolving areas of discomfort and enhancing client satisfaction.

8. Finance and Stock Market Analysis: Sentiment analysis is employed in the finance industry to assess public sentiment on stocks, currencies, and financial instruments. It facilitates the process of making well-informed investing decisions.

9. Healthcare and Medical Research: Sentiment analysis is utilized in the healthcare industry to comprehend patient input, evaluations of medical facilities, and attitudes towards certain treatments or drugs. This information is really helpful for enhancing healthcare services.

10. Sentiment analysis aids in comprehending consumer responses at all stages of product launches and marketing efforts, including before, during, and after. This data informs modifications to marketing tactics and product positioning.

11. Legal and Compliance: Sentiment analysis may be utilized in legal settings to examine extensive quantities of legal papers, aiding attorneys and legal experts in identifying crucial attitudes and viewpoints that are pertinent to their cases.

12. Academic Research: Sentiment analysis is employed in many academic disciplines such as psychology, sociology, and linguistics to examine extensive sentiment patterns in texts and comprehend human emotions.

Sentiment analysis is a flexible instrument that has a crucial impact on the development of strategies, decision-making processes, and customer-focused methods in many businesses and fields. The capacity to get significant insights from unorganized data renders it an indispensable asset in the contemporary information-centric society.

2.1.3 Sentiment Analysis Challenges

The utilization of sentiment analysis has inherent difficulties that necessitate attention from scholars and practitioners (Birjali et al., 2021). These are some of the primary obstacles in sentiment analysis:

1. Ambiguity and Polysemy: Words can possess various interpretations based on the surrounding circumstances. For instance, the term "cool" can denote a low temperature or be employed as informal language to convey approval. Accurately determining the intended meaning is essential for precise sentiment analysis.

2. Identifying sarcasm and irony in literature can provide a challenge since they frequently employ words or expressions in a manner that defies their literal interpretation.

3. Negation and Modifiers: Expressions such as "not bad" or "pretty good" might be problematic since they contain negation or modifiers that can alter or amplify the feeling.

4. Context Dependency: The sentiment conveyed in a statement can be significantly influenced by the surrounding context. For example, the attitude towards the "battery life" of a device may vary in a review of a smartphone as opposed to a laptop.

5. Cultural and regional variations can lead to differences in sentiment expression. Words or phrases can carry varying implications or moods depending on the language or cultural environment.

6. **Domain-Specific Language:** Sentiment analysis models that are trained on a certain domain may exhibit worse performance when applied to a distinct domain. For instance, a model that has been trained on product evaluations may not be as efficient in evaluating the sentiment of healthcare-related content.

7. **Data imbalance** refers to the uneven distribution of sentiment labels in numerous datasets. For instance, there can be a substantial imbalance between the number of favourable evaluations and bad reviews, which can affect the model's capacity to learn from both categories.

8. **Dealing with Multimodal Data:** Sentiment analysis may need the processing of many types of data, including text, photos, videos, or audio. Integrating data from several modalities and comprehending them in a unified manner might pose difficulties.

9. **Emotion and Intensity Detection:** Sentiment analysis commonly concentrates on categorizing sentiments into binary or ternary categories, such as positive, negative, or neutral. Expressing subtle feelings or varying degrees of sentiment can be a more intricate task.

10. **Insufficiency of Labelled Data:** Obtaining labelled data for sentiment analysis, particularly for specialized topics or languages, maybe a laborious and costly process. The scarcity of labelled data might impede the training of models.

2.1.4 Sentiments Analysis on Social Media Platforms

Within the domain of politics, applications that provide voting recommendations and gather feedback play a vital role in conducting sentiment research. This study allows campaign advisers to closely observe and manipulate public opinion on many matters, while also evaluating how politicians' remarks and actions might affect voter inclinations. A comprehensive analysis of tweets pertaining to the U.S. presidential elections is carried out. This sentiment study evaluates the popularity of the two primary presidential contenders, Donald Trump and Hillary

Clinton, on the social media platform Twitter. The average sentiment score of tweets referencing each contender is calculated on a daily basis. The qualities linked to the tweets, along with the tweet content, are employed to obtain a more profound understanding of the stated emotion and the material shared or created by users (Fujiwara et al., n.d.) 2023). Sentiment analysis is used in this context to clarify the public opinions and views of politicians, including their perspectives on specific subjects. Consequently, this improves the availability of precise and top-notch information for voters. Sentiment analysis involves a broad and varied range of insights and interpretations, including social problems, political opinions, and consumer feelings.

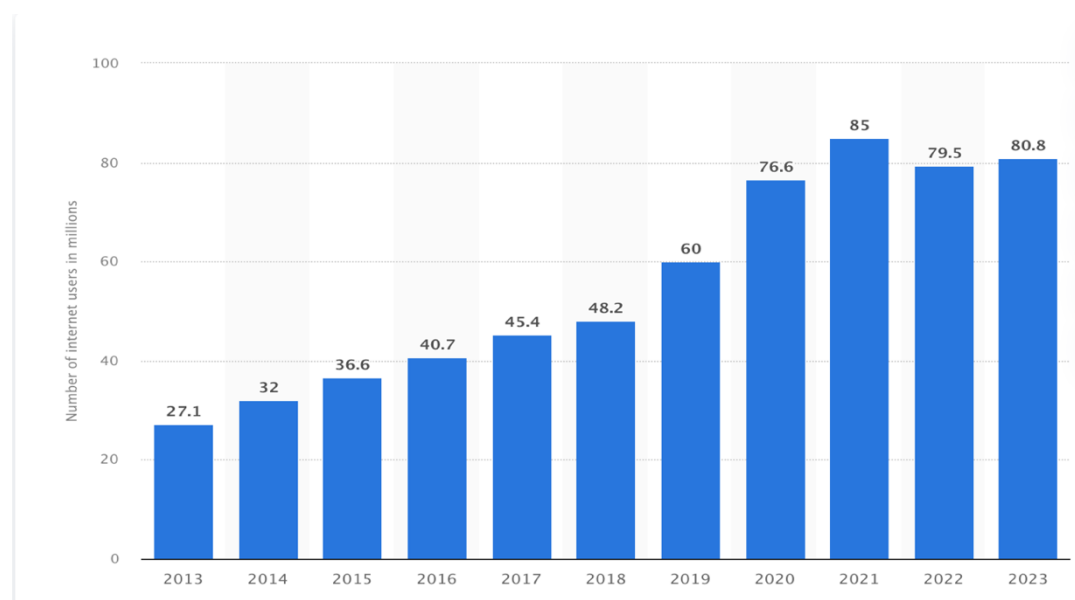
Social media involvement has been more popular in Arabic nations, as individuals perceive it as a vital platform for freely expressing their thoughts. Facebook is very prominent in Saudi Arabia, being widely acknowledged across many platforms. As of January 2020, Egypt has a notable 42.4 million Facebook members, or 40.4% of its total population (“Facebook users in Egypt - January 2020,” 2020). The significant influence of Twitter as a social media platform was particularly apparent during the Arab Spring upheavals in nations such as Syria, Yemen, Tunisia, and Egypt. Especially during this time of increased speech and popular uprising, people openly aired their opinions using Twitter. Through the utilization of informal, interconnected communication channels, demonstrators effectively organized movements, coordinated protests, and heightened awareness among regional and worldwide audiences. The source cited is (Shearlaw, 2016). In order to emphasize the importance of this type of social activism, (Alhindi et al., 2012) observed that there were around 40-45 tweets per minute in Egypt alone on January 25, 2011. These tweets played a crucial role in initiating and organizing the social movement that set the foundation for the Arab Spring.

The number of Arabic users on Twitter has experienced a significant surge of more than 100% since 2017. Based on the use statistics supplied by Crowd Analyzer in 2019, the number of Twitter users actively engaging in Saudi Arabia exceeded 11 million during the period from 2018 to 2019. This surpassed the popularity of both Egypt and the UAE, as seen in Figure 1. Currently, the population of Saudi Arabia is 33.85 million, with 23 million people, or for 68% of the entire population, actively participating on the platform.

In January 2023, Egypt's internet user base reached over 81 million persons, showing an increase from the 79.5 million users in the previous year. This is a significant increase over the 27 million internet users recorded a decade ago, as seen in Figure 1.

Figure 1

Number of Internet Users in Egypt



2.2 Arabic Sentiment Analyses

Arabic is of great significance to millions of individuals in Arabic-speaking nations, since it serves as both the official language and a language profoundly intertwined with the religious practices of Muslims. Arabic, being a Semitic language, consists of 28 letters. The world

population consists of around 5% native Arabic speakers, distributed over 22 Arabic-speaking nations, totalling around 400 million individuals as of 2022. Based on United Nations forecasts, the population is expected to increase to over 589 million by 2050 (“Population Levels, Trends and Policies in the Arab Region: Challenges and Opportunities | United Nations Development Programme,” 2013). Furthermore, Arabic has shown a significant increase in popularity on the internet and is currently acknowledged as the fourth most often utilized language (Awajan and Mohamad, 2018). Figure 2 visually depicts the Arab world.

Nevertheless, the number of natural language processing research dedicated to the Arabic language may not be commensurate with the widespread use of the language (Mozetič et al., 2018). In addition, the development of sentiment analysis techniques for the Arabic language is not as advanced as those for the English. However, there is a growing body of research focusing on Arabic sentiment analysis, as evidenced by studies such as (Jin, 2017) Figure 3 (El-Masri et al., 2017).

Figure 2

Arab World Map

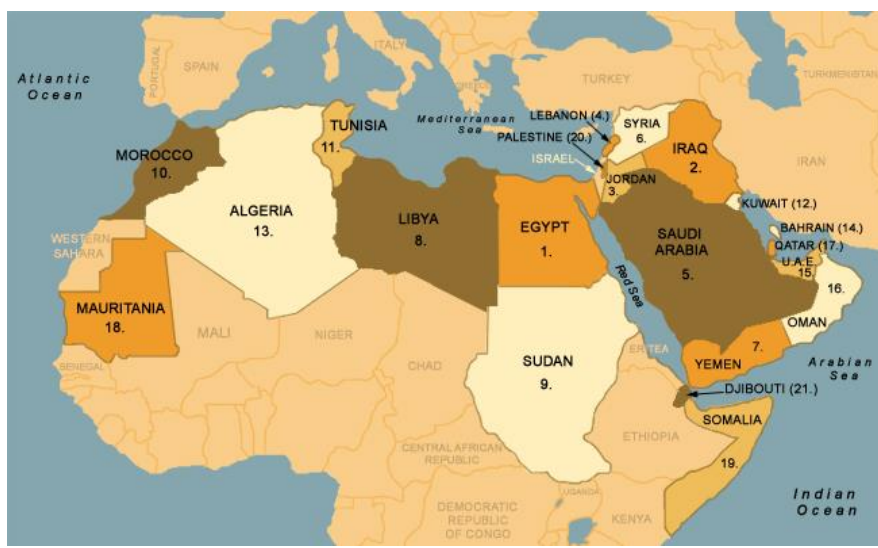
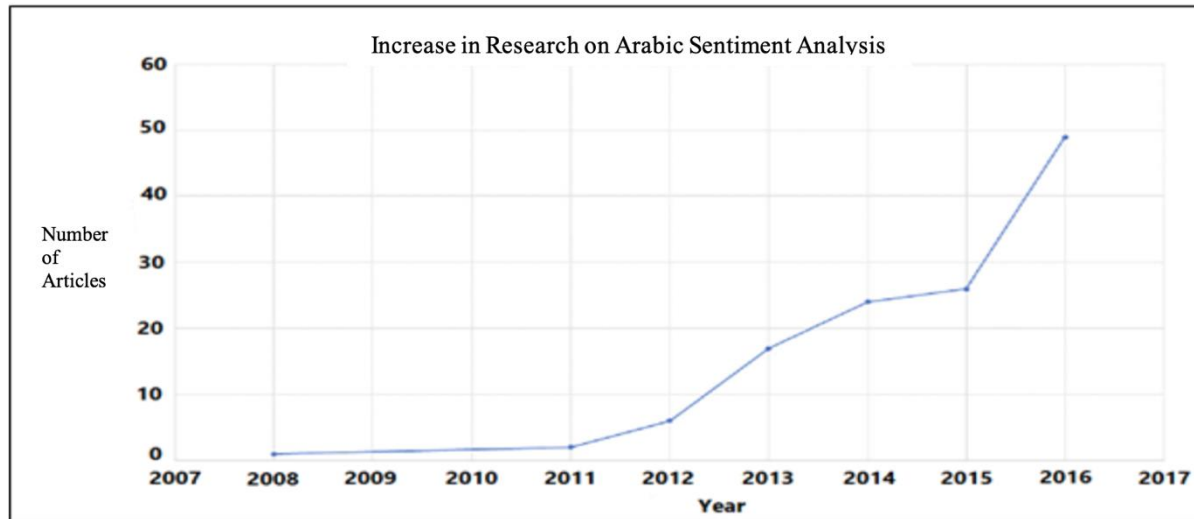


Figure 3

The Increasing Trend in Sentiment Analysis Research for the Arabic Language (El-Masri et al., 2017)



2.2.1 Arabic Language Processing Characteristic

The Arabic language exhibits a rich array of formal and informal terminologies, both in its written and spoken forms across different Arab regions. This linguistic diversity leads to the presence of texts composed in numerous dialects, adding complexity to sentiment analysis tasks (Guellil et al., 2021). Furthermore, Arabic, being a Semitic language, adheres to a right-to-left writing system and exhibits a cursive visual aspect resulting from the interconnectedness of letters to create words. Moreover, Arabic lacks separate lowercase and uppercase letters. Language may be categorized into three main groups: colloquial or dialectal Arabic, modern standard Arabic (MSA), and Classical Arabic.

1. Colloquial or dialectal Arabic

Colloquial or dialectal Arabic refers to the spoken form of the language used by Arabic speakers in their daily activities and interactions. Colloquial Arabic has regional variances that are impacted by disparities across various nations and areas. Conversely, modern standard

Arabic (MSA) is perceived as the standardized version of language that is consistently employed in all Arabic-speaking countries.

2. modern standard Arabic (MSA):

Modern Standard Arabic, originating from classical Arabic, functions as the standardized variant of the language employed for both written and spoken communication in all Arabic-speaking countries.

3. Classical Arabic:

The language used for reciting the Quran is commonly referred to as Quranic Arabic. Its origins stretch back to the Hijaz area in Saudi Arabia over a millennium ago. Classical Arabic comprises written documents, such as ancient poetry, that originated before the pre-Islamic Era (Mohammed, 2016).

2.2.2 Challenges in Arabic Language Processing

Examining words in Arabic phrases may be complex, particularly in activities such as sentiment analysis, data mining, and information retrieval. The process involves several stages, including as stemming, lemmatization, and the elimination of stop words. This necessitates a more comprehensive lemmatization or stemming process that surpasses the three primary variant stemmer approaches often used for Arabic text. Some of the methods used for stemming in Arabic text processing include the root-extractor statistical stemmer approaches, the light stem-dependent method by Larkey, the Arabic stemmer created by the Information Science Research Institute (ISRI), and the root-dependent method developed by Khoja (Yousif and Elkabani, 2015).

The multitude of spellings and typographic layouts in Arabic calligraphy results in a notable degree of ambiguity. The diverse Arabic writing traditions and flexible layout contribute

to the multiplicity of named entities, which can be expressed in many fashions. Furthermore, Arabic does not possess a defined lettering convention, unlike English. As an illustration, it is deemed permissible to represent the word "google" as either "جوجل" or "غوغل". Similarly, the term "Europe" can be translated as "أوروبا" or "اوروبا". Arabic words can be potentially ambiguous, as they might have numerous meanings. For instance, the named object "جمعة/ Friday" might have dual references, representing both a day of the week and a masculine name. Diacritization is essential for accurately communicating the intended meaning of a word. For example, "دَرَسَ)" can represent the past verb "studied" or the phrase "دَرْس" which means "lesson" (Abdelrahman et al., 2010). Nevertheless, a significant number of Arabic textual materials do not have diacritization, which presents difficulties in accurately parsing sentiment analysis tasks.

Moreover, the Arabic language is susceptible to frequent spelling mistakes. For example, the term "المدرسة" could be erroneously written as "المدرسه" by utilizing the letter "ه". Similarly, the word "يقرأ" which is spelled with the letter "أ" may be mistakenly transcribed as "يقرء" using the letter "ء" (Alsayadi and ElKorany, 2016)). In addition, Arabic demonstrates several inflections for word roots based on the context, as seen by instances such as "العب", "العبة", and "ملعب". Moreover, the Arabic language has terms like "لكن" that may express both positive and negative attitudes at the same time (Alhumoud et al., 2015). Moreover, Arabic poses several intricacies for language processing and analysis. The language has sophisticated semantics, grammatical structures, and morphologies, which arise from its complex derivational and inflectional properties. Consequently, English-oriented natural language processing systems have difficulties in fulfilling the processing requirements of the Arabic language (Alansary et al.,

2013). These constraints and challenges may occur while embarking on initiatives pertaining to Arabic sentiment analysis.

- **Data Collection**

The acquisition of Arabic data is a notable obstacle due to the limited availability of accessible options for collecting such data. Frequently, data collecting relies on internet forums or social media sites. Nevertheless, the gathered data may encompass a wide range of components, such as Arabizi (the use of Latin character for Arabic words) or different dialects. Consequently, it suggests that the sentiment analysis system might be trained on a comparatively restricted dataset. Hence, this can lead to diminished precision or potentially erroneous results (El-Masri et al., 2017).

- **Restricted Availability of Toolkits, Datasets, and Corpora**

The efficacy of sentiment analysis relies on the quantity of accessible data (El-Masri et al., 2017). However, unlike English, Arabic does not have abundant resources like easily accessible sentiment analysis toolkits, tagged text corpora, and large datasets. These materials are frequently either restricted in access or limited in availability. The data cleaning, processing, and aggregation procedures are influenced by these limitations, which might result in time-consuming tasks and potentially produce erroneous results. This requires more exertion and careful focus on language subtleties, especially when sentiment analysis methodologies depend on these techniques (Alsayadi and ElKorany, 2016).

- **Recognizing Named Entities**

Precisely recognizing Arabic nouns is of utmost importance in sentiment analysis. It is crucial to note that differentiating between Arabic nouns and adjectives can be somewhat difficult. In contrast to English, which commonly capitalizes nouns, Arabic lacks this visual

indicator, hence complicating the identification of nouns. The differentiation is crucial in identifying the individual who is articulating a specific viewpoint (El-Masri et al., 2017).

- **Issues pertaining to inaccuracies in transcriptions and spelling on social media platforms**

Social media services, such as Twitter, frequently feature phrases that undergo transliteration or misspelling. As an example, the Arabic translation of the word "mobile phone" is written phonetically as "موبايل" instead of its accurate form "محمول او جوال". Furthermore, tweets often contain hashtags and mentions of names, which might add unwanted interference to the dataset (El-Masri et al., 2017).

- **Analysis of Negation**

Arabic dialects contain negations that may not be present in modern standard Arabic (MSA), and they can be conveyed in various ways. The presence of several meanings associated with negation words complicates the task of accurately detecting negation. This complexity increases the likelihood of mistakes, as exemplified by the phrase "مش كدة", which translates to 'not' (El-Masri et al., 2017).

- **Identifying spam reviews**

Spurious or deceptive reviews, also referred to as spam reviews, have the intention of deceiving readers and can exhibit either excessively favourable or negative sentiments. Companies may employ this strategy to enhance or undermine the reputation of their products. There has been a scarcity of research specifically aimed at detecting fraudulent reviews written in Arabic (El-Masri et al., 2017).

- **Co-reference resolution**

The task of resolving co-references is a notable difficulty in the interpretation of emotion in Arabic. When there are differing perspectives, it might be difficult to determine the specific target of the feeling. For example, "وجبة مع الأرز مع اللحمة جميلة بس السلطة مالحة شوية واللحمة ناشفة قوي" in the sentence "The rice meal is delicious, but the salad is somewhat salty and the meat is very dry," where the presence of both positive and negative attitudes, along with the absence of well-organized sentences and the utilization of dialect, creates a challenge in identifying the intended polarity target (El-Masri et al., 2017).

- **Restricted Arabic Research for Dialectal Language**

Arabic sentence parsing is complex because of the language's sophisticated morphology, which is characterized by rich derivational and inflectional elements. Although parsers for modern standard Arabic (MSA) exist, there is a noticeable scarcity of parsers especially tailored for colloquial Arabic. The scarcity of resources is due to the lack of standardization and significant differences in lexical, morphological, and phonological features between colloquial Arabic and modern standard Arabic (MSA). The development of part-of-speech taggers and morphological analyzers for colloquial Arabic is hindered by these limitations (El-Masri et al., 2017).

- **Managing Inverted Emoticons**

Due to Arabic's right-to-left writing system, dealing with emoticons presents a significant obstacle. This might potentially result in the misreading or misplacement of emoticons, leading to contradicting feelings expressed in a tweet. Moreover, the utilization of emoticons for categorizing unsupervised data presents a difficulty as a result of this discrepancy in directionality. Moreover, tweets can express sarcasm by employing inverted emoticons, and

utilizing wrong ones can significantly alter the intended message of a statement (El-Masri et al., 2017).

- **Examining Linguistic Characteristics in Arabic Texts**

Analyzing dialectal or informal Arabic is a considerable obstacle for emotion analyzers because of its absence of organization and uniformity. It deviates from modern standard Arabic (MSA) in syntax, morphology, and pronunciation, hence complicating activities such as POS tagging and morphology analysis. Furthermore, it is important to note that stop words and negations may differ in dialects when compared to modern standard Arabic (MSA). Furthermore, it might be difficult to create lexicons for several Arabic dialects due to the lack of similar lexical possibilities for specific ideas across different dialects. Some reviewers also integrate both modern standard Arabic (MSA) and dialectal Arabic in a single sentence, particularly when citing Quranic passages or poetry (El-Masri et al., 2017).

- **Discerning Sarcasm**

Identifying sarcasm is a difficulty since it might lead to the incorrect categorization of emotions. There is a scarcity of academic research, particularly in the context of Arabic phrases (El-Masri et al., 2017).

- **Limited Sentiment Lexicons**

Arabic has a restricted quantity of emotion lexicons that are available to the public, unlike English. Although there exists a comprehensive Arabic attitude lexicon known as SANA, which encompasses all dialects and genres, its main emphasis is on two specific dialects: Levantine and Egyptian. In addition, the utilization of SANA in subjectivity and sentiment analysis tasks has not been extensive (El-Masri et al., 2017).

- **Analyzing Idiomatic Expressions and Composite Phrases**

Arabic social media material often includes colloquial idioms and complex sentences. The interpretations of these phrases can vary across different dialects and may have contrasting connotations in diverse Arabic-speaking areas. Furthermore, the words and phrases undergo changes in accordance with usage patterns, resulting in the constant development of novel expressions. For instance, the expression "a monkey is a deer in the eye of its mother" conveys a negative attitude. Due to the large number of idiomatic phrases in Arabic, it is difficult to determine the polarity connected with each one (El-Masri et al., 2017).

- **Managing Arabizi Usage (Franco-Arabic)**

Arabizi is a contemporary linguistic phenomenon observed on social media platforms, wherein Arabic words are written using Latin letters. Moreover, several internet authors alternate between English and Arabic, rendering it intricate to ascertain whether a word is written in English or Arabizi. The study efforts have not adequately addressed this particular element of Arabizi (El-Masri et al., 2017).

2.3.3 Analyzing Related Works in Sentiment Analysis of Arabic

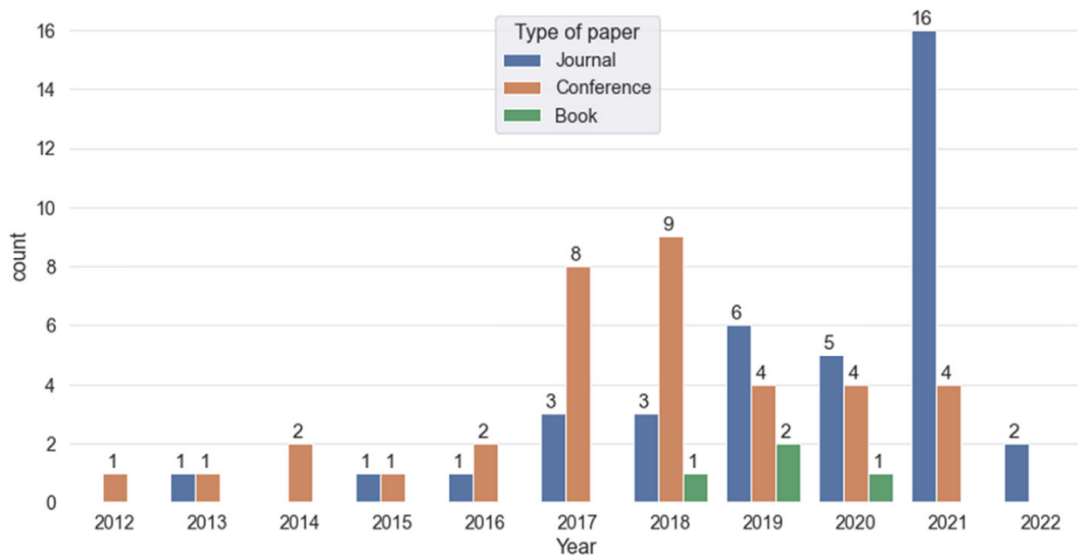
In recent years, significant advancements have been made in the field of natural language processing (NLP) and sentiment analysis, particularly in handling diverse languages and dialects. Meziane et al. (2020) provide an extensive survey on Arabic machine translation, highlighting the latest trends and challenges, which are crucial for understanding the complexities involved in processing Arabic language data (Meziane, 2020). This aligns with our approach of leveraging Deep Learning Models to enhance sentiment analysis in Arabic dialects.

Significant and impactful articles, which receive a considerable amount of citations, are crucial for furthering research on a certain subject (Mäntylä et al., 2018). The Systematic Literature Review (SLR) entailed procuring reputable journals and conferences. The review

concentrated on publications published between 2012 and the present. Figure 4 demonstrates that research papers provide around 80% of the primary materials in the field of Arabic sentiment analysis from 2012 to 2022, while reviews and journal articles collectively contribute roughly 18%. Book chapters, in comparison, make up just 2% of the overall sources, suggesting that they are used less frequently (Matrane et al., 2023).

Figure 4

Type of Research Works by the Time



The researchers (Abdul-Mageed et al., 2014) developed SAMAR, a supervised machine learning system designed for analyzing the subjectivity and sentiment of Arabic communications on social networks at the phrase level. Their contribution included four topics: the impact of platform-specific characteristics on performance; strategies for handling Arabic dialects; the applicability of standard features used for English to Arabic; and the most effective way to convey lexical information. In addition, they examined the sentiment analysis and subjective characteristics of four different platforms: web discussion forums, Wikipedia Talk Pages, Twitter, and synchronous chat. The languages utilized on these platforms exhibit variation, as

well as divergent intended purposes. Although users on other platforms occasionally utilized both modern standard Arabic (MSA) and dialectal Arabic, it is dialectal Arabic that predominates as the primary spoken language on the chat platform (Abdul-Mageed et al., 2012).

The findings indicate that including either a lemma or a lexeme as morphological information is advantageous. Additionally, using two speech tag (POS characteristics) components specific to Arabic, namely the extended reduced tag set (ERTS) and the reduced tag set (RTS), has been proven to enhance subjectivity and sentiment analysis of Arabic writing on social networks. However, the outcomes of their study indicate that distinct solutions are required for every job and platform. A definitive determination of the optimal mix of various information types for tasks involving sentiment analysis vs subjectivity, as well as for particular corpora, cannot be achieved by a singular, unified approach. Lexeme was shown to be more suitable for sentiment analysis, whereas the expanded POS tag set, along with lemma information, is more fit for subjectivity analysis (Abdul-Mageed et al., 2012). The DARDASHA dataset, consisting of 2,798 chat messages gathered from Egyptian chat rooms, was subjected to sentiment analysis using SAMAR. The test yielded a maximum accuracy result of 70.30%, indicating an improvement compared to the baseline. The test utilized morphological characteristics, lemmas, an extended reduced tag set, a polarity lexicon, and a unique feature.

In addition, the researchers in (Mourad and Darwish, 2013) introduced a corpus that was annotated manually. The corpus was created by annotating newswire items sentence by sentence. The dataset consisted of 2,855 sentences. In addition, the researchers utilized the annotated data to showcase an automated method for subjectivity and sentiment analysis. The researchers examined the impact of various levels of preprocessing frameworks on the job of categorizing subjectivity and sentiment analysis. The outcomes were quantified using the F-measure criterion.

They have shown that, while considering the subjective test, the maximum outcome of 71.54% F was achieved by combining stemming with morphology and adjective qualities. Nevertheless, the highest score achieved in the emotion test for the combination of stemming with morphology and adjective features was 95.52% F (Mourad and Darwish, 2013).

2.3 Speech Emotion Recognition

The topic of Speech Emotion Recognition (SER) study has had rapid growth since its establishment in the early 2000s (Schuller, 2018). The domain of speech-based human emotion identification is growing in tandem with the fascination in voice-based user interfaces. This section covers the techniques utilized in SER, as well as the limitations and challenges faced by previous research (Schuller, 2018).

2.3.1 Speech Emotion Recognition Features

The use of the emotion representation method determines whether a Speech Emotion Recognition (SER) job is treated as a regression or classification issue. Tasks in the SER (Speech Emotion Recognition) domain that employ categorical emotion representation are often categorized as such, whereas tasks that utilize dimensional representation are sometimes classed as regression tasks. In the initial stages of SER development, researchers analyzed speech data to determine fundamental attributes and sought to establish the connection between features and emotion representation using traditional machine-learning methods (Anagnostopoulos et al., 2012). In that era, it was challenging to discern a captivating collection of attributes, usually known as feature engineering. The scope of calculated features has broadened from basic high-level attributes to intricate low-level qualities. Literary characteristics are classified into two main groups based on their temporal structure, which is defined over a longer time span, similar to an utterance.

1) Segmental calculations, also known as spectral calculations, are performed on shorter time intervals.

2) Supra-segmental or prosodic features are calculated across a longer time period, such as an entire speech.

The literature use many spectral parameters, including frame energy, Log-Filer Power Coefficients (LFPCs), and Mel-Frequency Cepstral Coefficients (MFCCs). Pitch, shimmer, jitter, and speaking tempo are some examples of prosodic characteristics. Researchers have utilized statistical characteristics, which are calculated, alongside prosodic and spectral aspects (Anagnostopoulos et al., 2012; Swain et al., 2018). Recent research have utilized unprocessed audio data, known as an end-to-end solution, together with deep learning techniques, as demonstrated by (Trigeorgis et al., 2016); (Feng et al., 2018) and (Tzirakis et al., 2018).

2.3.2 Algorithms & Techniques of Speech Emotion Recognition

Machine learning (ML) has greatly influenced the advancement of recognition techniques from the beginning of SER. Commonly employed machine learning techniques for SER problems encompass Gaussian Mixture Models (GMMs), k-Nearest Neighbour Classifiers (kNNs), Hidden Markov Models (HMMs), Support Vector Machines (SVMs), Decision Trees (DTs), and Artificial Neural Networks (ANNs). Recently, the primary emphasis of research in SER has focused on algorithms that depend on deep learning (DL). Current speech emotion recognition (SER) methods have employed many deep learning (DL) techniques including autoencoders (AEs), deep belief networks (DBNs), recurrent neural networks (RNNs), auto convolutional neural networks (CNNs), autoencoders (DNNs), and long-short-term memory (LSTM) networks. Due to the presence of both benefits and drawbacks in each technique, there is no universally best strategy for all search engine results work.

Machine learning (ML) has greatly influenced the recognition approach since the inception of SER. Gaussian Mixture Models (GMMs), k-Nearest Neighbour Classifiers (kNNs), Hidden Markov Models (HMMs), Support Vector Machines (SVMs), Decision Trees (DTs), and Artificial Neural Networks (ANNs) are some machine learning techniques employed in Speech Emotion Recognition (SER) problems. Recently, there has been a significant focus on deep learning (DL) approaches in studies on speech emotion recognition (SER). Autoencoders (AEs), deep belief networks (DBNs), recurrent neural networks (RNNs), auto convolutional neural networks (CNNs), autoencoders (DNNs), and long-short-term memory (LSTM) networks are some of the deep learning (DL) approaches employed in recent speech emotion recognition (SER) solutions. There is no universally optimal method for all SER jobs due to the inherent pros and cons associated with each one. SVM-based classifiers have demonstrated superior performance compared to many other machine learning algorithms in the SER literature, including as logistic regression, k-nearest neighbour, random forest, and gaussian mixture model (Rybka and Janicki, 2013); (Kuchibhotla et al., 2016) (Chen et al., 2012). The study done by (Tian et al., 2015) shown that when a little quantity of data was used for training, the support vector machine (SVM) performed better than their combined long short-term memory and recurrent neural network (LSTM-RNN) model. When a significant quantity of training data was given, the LSTM-RNN model demonstrated superior performance compared to SVM. The authors (Chen et al., 2012) (Rybka and Janicki, 2013) have found that Support Vector Machines (SVM) provide superior performance compared to Artificial Neural Networks (ANN). Several studies have found that Artificial Neural Networks (ANN) outperform Support Vector Machines (SVM) in terms of performance (Staroniewicz, 2009). The resilience of a model is determined by the magnitude of its discriminative feature space, as demonstrated by this attribute of Support

Vector Machines (SVM). Kuchibhotla et al., (2016) found that their feature selection strategy significantly improved the accuracy of Support Vector Machine (SVM) in a research investigating feature selection strategies. This result establishes proof of the association between the discriminative characteristics of the feature space and the SVM model. Recently, Support Vector Machines (SVM) have frequently been employed in research as a classifier for automatically learnt features via a representation technique. The SVM takes the learnt representation as its input in this approach. This strategy has been examined in the literature by (Stolar et al., 2017); (Luo et al., 2018) and (Heracleous et al., 2019). In general, Support Vector Machines (SVM) are extensively used in Speech Emotion Recognition (SER) because they have low data requirements, outperform other machine learning methods, and can effectively exploit the distinguishing characteristics of high-dimensional data. Nevertheless, one significant drawback of SVM is its incapacity to capture sequential information as a result of its deficiency in contextual comprehension (Tian et al., 2015). However, voice data has a sequential nature. Temporal variations in voice data can provide valuable information on linguistic and paralinguistic signals that may be linked to emotions. While contextual learning is absent, feature engineering should be enough to clarify the performance of Support Vector Machines (SVM) on Speech Emotion Recognition (SER).

Artificial Neural Networks (ANN) is a machine learning technology that seeks to imitate the learning mechanism of the human brain, which has been the subject of extensive controversy. An artificial neuron serves as the basic component of an artificial neural network (ANN), mimicking the structure of a genuine brain neuron (Mcculloch and Pitts, 1943). A deep Neural Network (DNN) is a multi-layered network of artificial neurons. The formation topology of an ANN can vary significantly, with each topology possessing distinct characteristics. This section

only focuses on the utilization of unadulterated, completely interconnected iterations of Deep Neural Networks (DNN) and Artificial Neural Networks (ANN) in Speech Emotion Recognition (SER) challenges. The Artificial Neural Network (ANN) has played a significant role in the field of Speech Emotion Recognition (SER) from its inception (Nicholson et al., 1999); (Hozjan and Kačič, 2003) and (Petrushin, 2000). Artificial neural networks (ANNs) have been recognized as somewhat successful in acquiring knowledge about non-linear interactions (Anagnostopoulos et al., 2012). Acquiring knowledge of non-linear interactions was seen beneficial at that time. As to the study conducted by (Abo absa et al., 2018). Deep Neural Networks (DNN) have the ability to acquire knowledge about a diverse range of categories without compromising on efficiency. However, both pure artificial neural networks (ANNs) and deep neural networks (DNNs) do not possess the capability to acquire contextual knowledge, similar to support vector machines (SVM) as mentioned by (Mao et al., n.d.) and (Zhang et al., 2011). In addition, networks that are completely linked are more prone to overfitting. A convolutional neural network (CNN) refers to a particular deep neural network (DNN) architecture that incorporates regularization techniques between layers. Regularizations are incorporated into the pooling and convolutional layers of the DNN. Convolutional layers perform convolution on their input layer by multiplying it with a smaller layer known as the convolutional kernel. The refined function employs pooling layers to decrease the input dimension. As per the findings of (Krizhevsky et al., 2012), CNN has made substantial advancements in tasks related to recognizing pictures and patterns due to its ability to gain abstract local knowledge. CNNs are commonly employed in the SER domain, utilizing spectrogram and cepstrum inputs such as MFCC, due to their image-like behaviour (Badshah et al., 2017). Similar to Support Vector Machines (SVM) and Artificial Neural Networks (ANN), Convolutional Neural Network (CNN) models have shown their ability to extract geographical

information from the data. However, they lack the capability to extract temporal information (Ma et al., 2018). Occasionally, researchers have integrated the CNN model with other models capable of acquiring temporal properties as a potential answer. The Gated Recurrent Unit (Syed et al., 2018), the Long-Short Term Memory (LSTM) (Yang and Hirschberg, 2018), and the Recurrent Neural Network (RNN) (Ma et al., 2018) are all examples of these models.

Recurrent Neural Networks (RNNs) and Long-Short Term Memory (LSTM) networks are two types of DNN topologies that include the inherent ability to acquire temporal information. RNNs were introduced by (Rumelhart and McClelland, 1987), while LSTM networks were developed by (Hochreiter and Schmidhuber, 1997). The speech signal's properties encompass both spatial and temporal information, as previously mentioned. To acquire proficiency in extracting emotional cues from vocal expressions, a framework equipped with temporal learning capabilities would be advantageous. Combining a CNN model with an RNN or LSTM would result in a model that can effectively learn both temporal and spatial dimensions. Most neural network topologies in speech emotion recognition (SER) that utilize a temporal model are often associated with a convolutional model (Etienne et al., 2018).

2.3.3 Representation Learning Methods for Recognizing Emotions in Speech

During the feature engineering phase of representation learning pipelines, a substantial level of human involvement is necessary to convert raw inputs into a representation. Feature engineering is a time-consuming procedure that heavily depends on the knowledge of human domain specialists. Due to the abundance of data and computational power available, there is a growing desire to assign the task of feature engineering to computers. Considering the prevalence of the desire for perfect characteristics in the field, similar to the scarcity of extensive and realistic datasets, it is not surprising that a new approach has emerged. This approach involves

learning features from data, as highlighted in a recent review article by (Schuller, 2018) which emphasizes the need for an automated method of feature learning for SER.

As per the findings of (Bengio et al., 2014) learning approaches has the capability to extract valuable features from a raw signal and utilize those attributes in a machine learning model. Researchers in the speech community have used representation learning approaches for tasks including data denoising, speech emotion identification, and automatic speech recognition. The study employs many deep neural network (DNN) architectures for the purpose of representation learning, such as the Generative Adversarial Network (Li et al., 2021), Autoencoder (AE) (Latif et al., 2020), Convolutional Neural Network (CNN) (Hema and Garcia Marquez, 2023), and RNN (Lee and Tashev, n.d.). Studies in the literature on software engineering research have demonstrated that employing representation learning, as opposed to conventional methodologies, leads to superior results.

CNN networks are commonly used in representation learning to forecast the spatial correlation of the input data. Usually, recurrent architectures like as RNNs or LSTMs are employed to model the temporal dependencies in the data. Information in speech signals is conveyed in both the spatial and temporal domains through the use of combination models, such as CNN-(Luo et al., 2018) or CNN-LSTM (Trigeorgis et al., 2016), which are employed for representation learning. However, researchers in SER sometimes calculate spectral data at a basic level and employ the representation learning approach to analyze spectral properties. In such scenarios, researchers exclusively utilize Convolutional Neural Networks (CNN) for training representations due to the image-like nature of spectral properties and their partial integration of temporal information.

An autoencoder consists of two main components, namely the encoder and decoder, which are combined sequentially to form a unique deep neural network architecture. The encoder component of an AE network frequently encodes a compact representation during the training phase. The decoder section utilizes the output representation of the encoder to generate a signal that is directly equivalent to the input signal. A following procedure utilizes the encoder component of the AE network. This technique enables researchers to train representations without labelled data. (Ghosh et al., 2016) state that autoencoders are utilized in the SER domain as well.

Another approach to obtaining shared representations for several tasks using a unified input data format is known as multi-task learning. As to the study conducted by (Lotfian and Busso, 2018), this technique entails training a model to do several tasks using the input data, rather than teaching it to perform a single task. For example, a Deep Neural Network (DNN) may be trained to predict the speaker's gender and emotions by analyzing voice patterns. The underlying network in this approach captures features associated with gender and emotions. When employing this approach, the model is less prone to overfitting a single task, as opposed to training two separate networks for each job.

Representation learning-based solutions for speech emotion recognition (SER) often demonstrate better generalization performance compared to other methods (Ghosh et al., 2016). Moreover, certain research has demonstrated a substantial improvement in the duration required for training a novel model.

Representation learning has the advantage of being able to utilize previously acquired representations for other tasks without altering the input data. This approach is commonly known

as transfer learning. One instance of this occurrence in the domain of voice processing is the utilization of automatic speech recognition models to augment the speech recognition job.

Representation learning has the advantage of being able to utilize the vast quantity of unlabeled data available for training purposes. Although there is a large number of speech data available for speech emotion identification, there is only a limited quantity of annotated data that includes emotion labels. Consequently, by utilizing speech data to enhance the accuracy of a representation, one may later train the acquired representation with a less amount of emotional data for categorization (Ghosh et al., 2016).

2.3.4 Arabic Speech Emotion Recognition

Speech emotion recognition systems vary based on the language, dataset characteristics, features extractions, and the implemented classifiers that recognize the emotion from the spoken words. Multiple research have been undertaken to identify emotions from speech in various languages, but only a limited number of them have been carried out specifically for Arabic speech. The researchers in (Hifny and Ali, 2019) utilized the KSUEmotions dataset (Meftah et al., 2021), which is recognized as one of the most extensive Arabic contributions to the corpus of Arabic performed emotions. The study employed two distinct models to discern emotions. The initial model integrated a combination of convolutional neural networks (CNN), bi-directional long short-term memory (BLSTM), and deep neural networks (DNN) to create the attention-based CNN-LSTM-DNN model. On the other hand, the second model just employed the deep CNN model. The findings indicated that the attention-based CNN-BLSTM-DNN model outperformed the CNN model, achieving an accuracy of 87.2% compared to 85.0% respectively. Another study utilized the identical dataset (Meftah et al., 2020) and employed Support Vector Machines (SVM) and k-Nearest Neighbors (KNN) algorithms to categorize emotions. The

outcomes of many trials shown that KNN achieved a superior performance of 75.49% compared to SVM's 68.75%. The emotion of sadness achieved the best accuracy in this sample, followed by pleasure and finally surprise. A different research study utilized Sentiment Analysis and Emotion Recognition (SER) to identify emotions in the Egyptian dialect of Arabic (Abdel-Hamid, 2020). The dataset obtained was a dataset that consisted of partially natural data. Prosodic, spectral, and wavelet characteristics were retrieved and utilized as inputs. Both the SVM and KNN algorithms were employed as classifiers. The researchers performed many tests, including both speaker-dependent and speaker-independent ones. They categorized the outcomes based on each classifier, extracted feature, gender, actor, and emotion.

Further research (Aljuhani et al., 2021) examines the process of emotion identification using Arabic Saudi dialect. The mel-frequency cepstral coefficient (MFCC) and mel spectrogram were retrieved as spectral characteristics. Subsequently, the classification techniques support vector machine (SVM), multi-layer perceptron (MLP), and k-nearest neighbor (KNN) were employed. The experimental results indicated that the Support Vector Machine (SVM) achieved the highest accuracy of 77.14% in Arabic speech emotion identification.

Table 1 presents the results obtained from much research cited in this area of the literature review:

Table 1

Arabic Speech Emotion in Literature

Reference	Dialect	Classifier	Accuracy %
(Aljuhani et al., 2021)	Saudi Arabic Dialect	SVM	77.14
		KNN	68.57
		MLP	71.43

(Hifny and Ali, 2019)	MSA	CNN-BLSTM-DNN	87.2
		CNN	85
(Meftah et al., 2020)	MSA	KNN	75.4
		SVM	68.7
(Abdel-Hamid, 2020)	Egyptian Arabic	SVM	66.8
	Dialect	KNN	61.7
(Li et al., 2015)	English dataset	CNN	68
	RAVDESS		

To highlight the limitations of previous works in Arabic Speech Emotion Recognition (SER) and how your research addresses these gaps, we can approach it as follows:

2.4 Speech Recognition Deep Learning and Hybrid Models

This section discusses the implementation of machine learning techniques employed in this study. To facilitate the acceptance of each classifier, it offers a detailed description of each machine learning technique and underscores its importance. The chapter ends by assessing and contrasting the machine learning classifiers using metrics such as precision, recall, F1-score weighted averages, and average accuracy results.

2.4.1 Tuning the Parameters

Hyperparameter sets are integral components of machine learning algorithms employed to accomplish the classification objective. The algorithm library or software package contains the default settings for these hyperparameters. However, during the tuning process, the default hyperparameter values may be adjusted to provide enhanced classification performance and results. Several approaches are available for hyper-parameter tuning, including the classic grid

search, random search, advanced iterative techniques such as Bayesian optimization, and iterated F-racing (Probst et al., 2019). Furthermore, users are responsible for specifying the desired ranges and values of the hyperparameters that require adjustment, as well as selecting a suitable tuning approach. In certain cases, hyperparameter values can be retained at their default settings if they yield satisfactory outcomes across different situations. Alternatively, in different scenarios, each dataset is employed to manually adjust the hyperparameters. The selection of appropriate values for the hyperparameters is a process of trial-and-error, drawing on past experience, and taking into account recommendations from scientific literature. Improper hyperparameter tuning can have a detrimental impact on the classifier's performance and diminish its output (Probst et al., 2019).

The two most often employed strategies for hyperparameter optimization are grid search and manual search. However, a study conducted by (Bergstra and Bengio, 2012) shown that random search outperformed grid search in optimizing hyperparameters, both in theory and in experiments. The empirical results were generated by comparing them with the findings of previous research that employed grid search and manual search methods to construct deep belief networks and neural networks. According to the researchers, they were able to generate models that required very little computing time and achieved either superior or similar performance. This was determined by comparing their model to neural networks that were exclusively developed via grid search. In addition, they argued that random search is more pragmatic than grid search as it allows researchers to flexibly include fresh (or successful) trials or exclude unsuccessful ones from the collection, which is not possible with grid search. Moreover, the researchers have determined that only a restricted set of hyperparameters were found to be efficient for the majority of the datasets. However, various hyperparameters were shown to be important for

different datasets (Bergstra and Bengio, 2012). Given that this study employed random search to fine-tune the hyperparameters of the classifiers, it is clear that grid search is an inadequate method for selecting classifiers when dealing with unseen datasets.

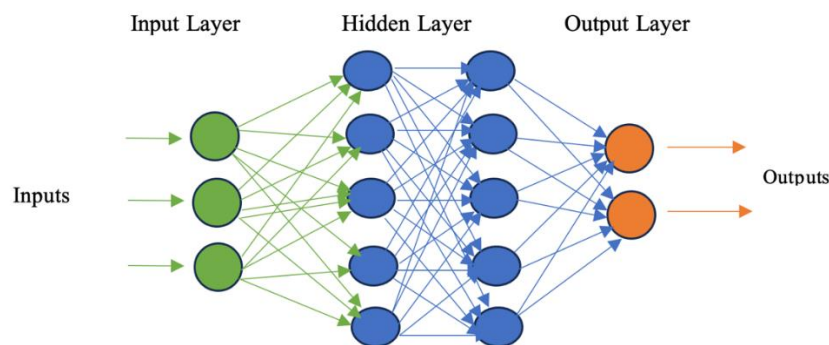
2.5 Deep Learning Approaches

Artificial Neural Networks have high efficacy in the field of Machine Learning. Neural networks are commonly used for processing datasets that include text, audio, and pictures. Diverse types of neural networks are utilized for distinct tasks. Recurrent neural networks, particularly long short-term memory neural networks, are utilized to forecast sequences of words. On the other hand, convolution neural networks are applied to classify pictures. In this blog, we will establish the fundamental components of CNN (Convolutional Neural Network) as outlined by (Gurney, 1997). Figure 5 illustrates the presence of three distinct layers in a typical neural network.

Input Layer: This layer serves as the input for our model. The total number of features in our data, or the number of pixels in the case of an image, is equal to the number of neurons in this layer.

Figure 5

Neural Network Block Diagram



Hidden Layer: The hidden layer gets the input that was previously provided to the input layer. The presence of several hidden layers is determined by the quantity of data and the structure of our model. The number of neurons in each hidden layer is variable, typically exceeding the number of features. The network exhibits nonlinearity when it incorporates learnable biases and an activation function into the output of each layer. This output is obtained by multiplying the output of the previous layer with the learnable weights of that layer in a matrix.

The output layer utilizes a logistic function, such as sigmoid or softmax, to process the data from the hidden layer. This transformation assigns a probability score to each class based on the output.

Feedforward is the procedure of acquiring the output from each layer of the model subsequent to the input data being provided to it. Subsequently, an error function is employed to measure the magnitude of the error; common examples of error functions encompass square loss error and cross-entropy. The performance of the network is evaluated using the error function. Subsequently, we calculate the derivatives in order to traverse backwards within the model. The objective of this step, referred to as backpropagation, is to minimize the loss to the greatest extent feasible. The sentiment analysis employed the subsequent deep learning techniques.

2.5.1 Convolutional Neural Networks

The convolutional neural network (CNN) is a very effective deep learning technique for computer vision and image recognition tasks. CNNs, or Convolutional Neural Networks, draw inspiration from the brain's visual information processing mechanism. They are specifically developed to autonomously detect patterns and characteristics inside images (Albawi et al., 2017).

CNNs, or Convolutional Neural Networks, are indeed a mathematical formulation that stimulates the neural networks of the brain. Some systems just utilize the two primary layers, namely the input and output layers, while others can incorporate the three basic levels, including the input, output, and hidden layer. Several research in the past several years have explored NLP with sentiment analysis, yielding promising outcomes. The solitary model was evidently surpassed by the CNN and LSTM algorithms. (Naive Bayes, Maximum Entropy, and Support Vector Machines). CNN employed a solitary convolutional layer to depict the grammar as an assemblage of words (Badshah et al., 2017).

A key feature of Convolutional Neural Networks (CNNs) is their capacity to acquire hierarchical representations of the input data. The lower layers of the network acquire basic elements such as edges and textures, whilst the upper layers acquire more intricate attributes such as objects and shapes. CNNs excel in various image and video identification tasks because to their hierarchical architecture. Nevertheless, the process of training a Convolutional Neural Network (CNN) necessitates a substantial amount of annotated data and can be demanding in terms of resources and time. Moreover, because to their designation as "black boxes," comprehending the process by which CNNs generate their predictions might provide a challenge (Albawi et al., 2017). As a result, there have been attempts to create methods for examining and displaying the patterns that Convolutional Neural Networks (CNNs) acquire. Figure 6 depicts the CNN Architecture.

Figure 6

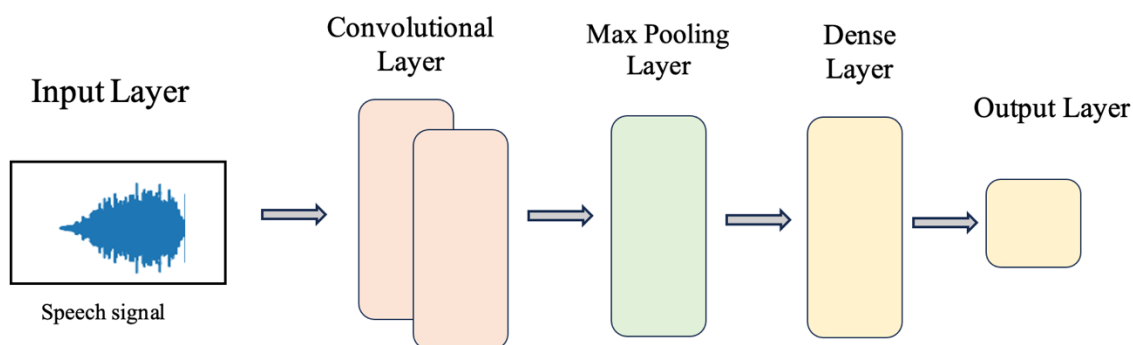
CNN Framework



Nevertheless, the process of training a Convolutional Neural Network (CNN) necessitates a substantial amount of annotated data and can be demanding in terms of resources and time. Moreover, because to their designation as "black boxes," comprehending the reasoning behind the predictions made by CNNs might provide a challenge (Albawi et al., 2017). Therefore, endeavors have been made to provide methods for the examination and depiction of the patterns that Convolutional Neural Networks (CNNs) acquire. Figure 7 displays the CNN Architecture.

Figure 7

Simple Architecture of CNN



- **The topics covered include layers, classification, and feature learning.**

A Convolutional Neural Network (CNN) consists of many hidden layers positioned between the input and output layers.

Convolutional: Convolution applies a sequence of filters on the input images, with each filter detecting a distinct characteristic and activating accordingly.

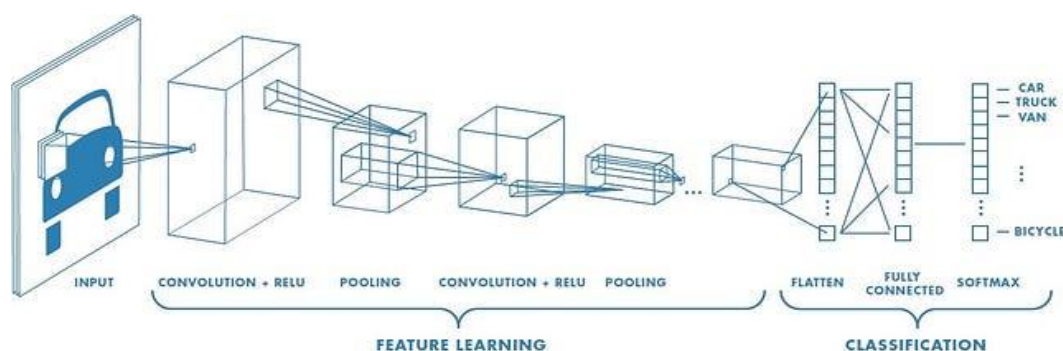
The Rectified Linear Unit (ReLU) maintains positive values and sets negative values to zero, facilitating faster and more effective training. The term "activation" is often used to describe the continuation of just the active qualities onto the following layer.

Pooling: Pooling simplifies the output and decreases the amount of parameters that the network needs to learn by implementing nonlinear downsampling.

These processes use several layers, ranging from tens to hundreds, which are trained to detect certain sets of characteristics through repetitive iterations. An example instance of a network with many convolutional layers. Each training image undergoes many resolutions of filters, and the output of each convolved image is used as the input for the next layer. Figure 8 depicts the architecture of a Convolutional Neural Network (CNN) (“What Is a Convolutional Neural Network?,” n.d.).

Figure 8

CNN Architecture



A Convolutional Neural Network (CNN) utilizes shared weights and bias values, which remain consistent across all hidden neurons inside a certain layer. This distinguishes it from a conventional neural network.

This suggests that all buried neurons in different regions of the picture are detecting the same feature, such as an edge or blob. Consequently, the network is able to withstand item displacement inside a picture. A network that has been trained to recognize automobiles, for example, will have the ability to do so regardless of the car's position within the image.

Convolutional Neural Networks (CNNs) have emerged as a potent tool in several domains of artificial intelligence, such as Automatic Speech Recognition (ASR). Convolutional Neural Networks (CNNs) are a specialized form of artificial neural networks that excel at

analyzing data arranged in a grid-like structure, such as pictures and spectrograms. This makes them highly suitable for jobs that require understanding spatial relationships. Automatic Speech Recognition (ASR), a technology that tries to transform spoken language into written text, has experienced significant progress with the use of Convolutional Neural Networks (CNNs).

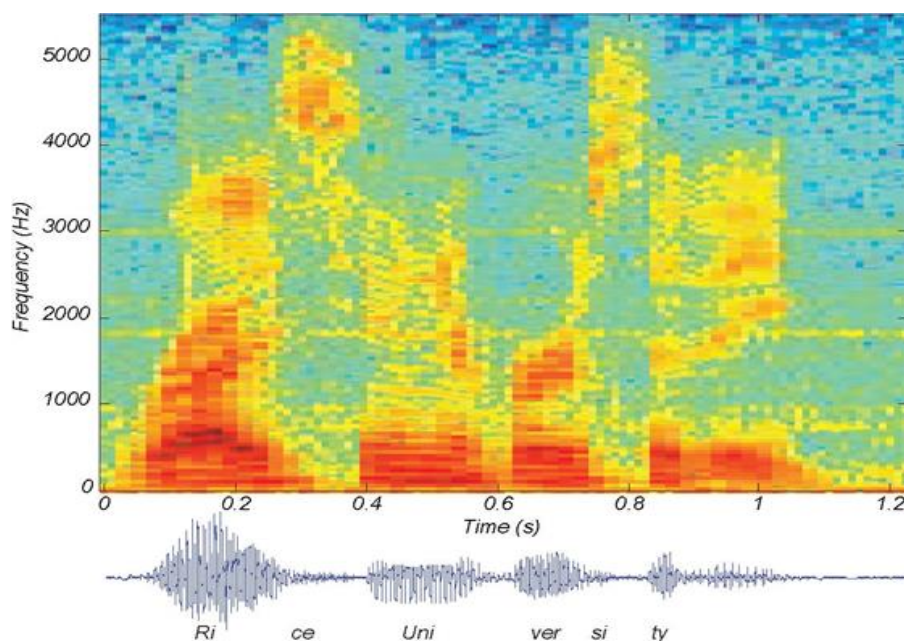
Convolutional Neural Networks (CNNs) in Automatic Speech Recognition (ASR):

Extraction of features:

Convolutional neural networks (CNNs) excel in autonomously acquiring hierarchical representations of data. Within the field of Automatic Speech Recognition (ASR), the spectrogram of an audio signal, which depicts the distribution of frequencies across time, can be seen as a picture. Convolutional Neural Networks (CNNs) have the ability to effectively extract pertinent features from spectrograms, therefore capturing crucial patterns that are fundamental for comprehending speech. Figure 9 displays a spectrogram of a spoken signal.

Figure 9

Spectrogram of Speech Signal



Temporal Dependencies:

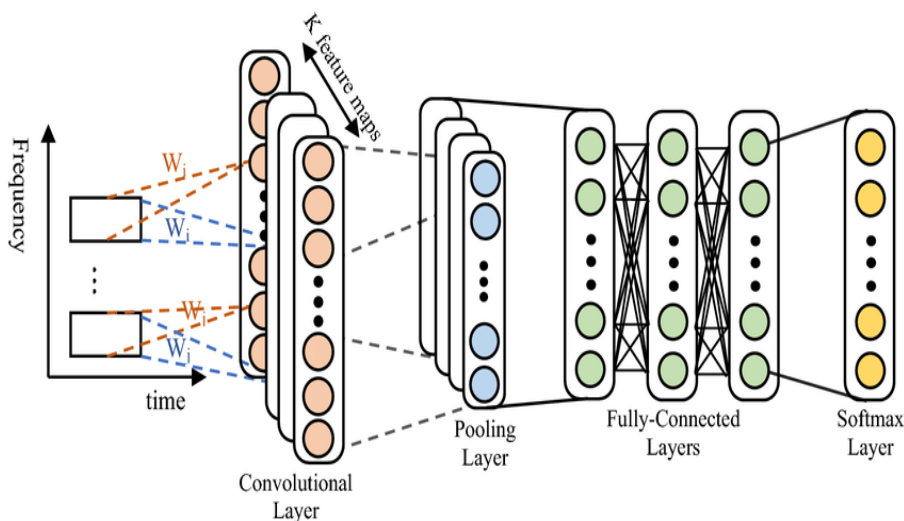
Although CNNs are primarily intended for spatial hierarchies, the essential temporal dependencies in voice recognition are handled by employing techniques such as 1D convolutions or integrating CNNs with recurrent neural networks (RNNs). CNNs are able to accurately represent both the frequency content and temporal patterns seen in voice signals. Figure 17 depicts the Convolutional Neural Network (CNN) architecture for Automatic Speech Recognition (ASR) (“Convolutional Neural Network Architecture for Speech Recognition,” n.d.).

Robustness to Variability:

Convolutional neural networks (CNNs) enhance the resilience of automatic speech recognition (ASR) systems by extracting consistent representations of speech information. These systems possess the ability to manage deviations in pronunciation, accent, and ambient noise, rendering them highly beneficial for practical use cases.

Figure 10

Inserting CNN Architecture for Speech Recognition



2.5.2 A Bidirectional LSTM

In order to understand how Bi-LSTM works, it is necessary to first understand the LSTM network and its unit cell. LSTM stands for long short-term memory. The introduction of LSTM networks was initially made by Hochreiter and Schmidhuber in 1997, as documented in their publication (Hochreiter and Schmidhuber, 1997). These recurrent neural networks are the most commonly employed ones.

Recurrent neural networks are renowned for their superior ability to process sequential input. However, there are instances where it is necessary to store the results of previous data as well. There are distinct semantic variations between the assertions "I possess the ability to engage in cricket" and "I have the intention to engage in cricket." It is evident that a solitary word has the ability to alter the significance of a sentence, therefore emphasizing the need of monitoring preceding words. Nevertheless, a fundamental Recurrent Neural Network (RNN) does not possess this type of memory. In order to address this difficulty, it is imperative that we acquire knowledge about LSTM.

The LSTM unit consists of three gates:

- The input gate

The input gate, also known as the second sigmoid function, receives the current state, $x(t)$, and the previous hidden state, h_{t-1} . The values of x_t and h_{t-1} range between 0 and 1, with 1 being insignificant and 0 being significant. Furthermore, the tanh function will compute the information pertaining to both the current and hidden states. The hyperbolic tangent function will generate an output within the range of -1 to 1, hence facilitating network control. Point-by-point multiplication of the output values generated by the activation functions is now feasible (Efanov et al., 2022).

- The forget gate

The forget gate is responsible for determining the data that should be excluded and the data that should be kept for further processing. The sigmoid function computes the information derived from the previous hidden state (h_{t-1}) and the current input X_t . After passing through the sigmoid function, the values are transformed into a range from 0 to 1. These values indicate whether the previous output's part is necessary, with values closer to 1 suggesting a higher requirement.

- The output gate.

The subsequent concealed state's value is found by utilizing the output gate. This state encapsulates data regarding previous inputs. At first, the third sigmoid function receives the values of the hidden state, including both the current and previous states. Subsequently, the hyperbolic tangent function is employed on the recently generated cell state derived from the cell state. The multiplicity of these two outputs is achieved in a step-by-step manner.

Ultimately, both the newly hidden state and the newly formed cell state are transferred to the subsequent stage.

An analysis of error propagation in current Recurrent Neural Networks (RNNs) revealed that extended time delays are impractical due to the phenomenon of backpropagated error, which either undergoes explosive growth or decays exponentially. This insight acted as the catalyst for the development of the Long Short Term Memory architecture.

An LSTM layer consists of memory chunks that are coupled in a recurrent manner. Consider these blocks as a differentiable counterpart to the memory chips used in digital computers. Each unit has three multiplicative components (input, output, and forget gates) that

simulate the write, read, and reset functions of the memory cells, together with interconnected memory cells (Etienne et al., 2018).

To summarize, the forget gate determines whatever relevant information from previous phases is necessary. The output gates finalize the subsequent concealed state, while the input gates ascertain whatever relevant data may be provided from the preceding stage.

- The architecture used is Bi-LSTM.

A bidirectional LSTM, or Bi-LSTM, is a sequence model that consists of two LSTM layers. One LSTM layer processes the input in the forward direction, while the other processes it in the backward manner. It is commonly utilized in occupations that include Natural Language Processing (NLP). The rationale behind this approach is that the model may enhance its knowledge of the relationship between sequences (e.g., comprehending the words that come before and after in a sentence) by analyzing input in both forward and backward directions.

The training sequence is delivered in both forward and backward directions to two distinct LSTM networks, which are linked to the same output layer, enabling the bidirectional LSTM networks to function. This implies that the Bi-LSTM retains comprehensive and sequential information about all preceding and subsequent points for each point in a given sequence.

To clarify, we encode the sequence by utilizing both forward and backward directions. This involves concatenating the output of both the forward and backward LSTM at each time step. The encoded representation of a word now comprehends the words that come before and after it.

The bidirectional LSTM architecture is formed by utilizing two unidirectional LSTMs to process the sequence in both the forward and backward directions. This design incorporates two

separate LSTM networks, each with a unique function. One LSTM network processes the token sequence in its original order, while the second LSTM network processes the token sequence in the reverse order. The final result of each of these LSTM networks is the addition of the two probability vectors that are produced as output. Equation 11 is one method of illustrating it:

$$p_t = p_t^f + p_t^b \quad (13)$$

where

p_t : the network's final probability vector.

p_t^f : the forward LSTM network's probability vector.

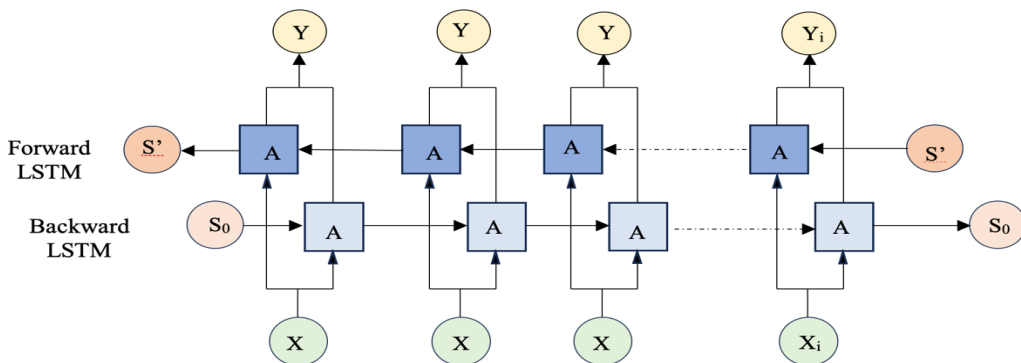
p_t^b : The backward LSTM network's probability vector.

The basic architecture of the Bi-LSTM is illustrated in Figure 11, with A and A' representing LSTM nodes, X_i denoting the input token, and Y_i representing the output token. The final output of Y_i is composed of both A and A' LSTM nodes.

Now, we will examine the implementation of a review system that utilizes Bi-LSTM layers in Python, employing the TensorFlow framework. We will do sentiment analysis using the IMDB movie review dataset. Initially, we will construct the network and train it to discern if a review is positive or negative.

Figure 11

BI-LSTM Layer Architecture



Bidirectional Long Short-Term Memory (BI-LSTM) networks have demonstrated significant efficacy in sequence-to-sequence tasks, rendering them extremely suitable for Automatic Speech Recognition (ASR). ASR is the process of transforming spoken language into written text. This work is improved by collecting both previous and upcoming context. The BI-LSTM, which is an extension of regular LSTMs, is particularly effective at capturing long-term interdependence in sequential data.

The BI-LSTM model consists of two LSTM layers, one that processes the input sequence in the forward direction and another that processes it in the reverse direction. The bidirectional nature of the network enables it to collect dependencies in both forward and backward directions, resulting in a more thorough comprehension of the sequential data (Efanov et al., 2022).

BI-LSTM networks have emerged as a fundamental component in ASR by efficiently collecting contextual information, hence improving the model's capacity to reliably detect and transcribe spoken English. The bidirectional flow of information in BI-LSTM is extremely helpful for accurately representing the complexities of speech patterns.

2.5.3 Gated Recurrent Unit Networks

The Gated Recurrent Unit (GRU) is an alternative to Long Short-Term Memory (LSTM) networks. It is a type of recurrent neural network (RNN) that was initially introduced by (Bahdanau et al., n.d.). GRU has the ability to handle sequential data, such as time-series data, text, and speech, similar to LSTM.

The functioning of GRU is based on the selective modification of the hidden state of the network at each time step via gating mechanisms. Gating strategies are utilized to manage the

flow of information into and out of the network. The GRU consists of two gating mechanisms: the update gate and the reset gate (Shewalkar et al., 2019).

The extent to which the previous hidden state is forgotten is governed by the reset gate, whereas the extent to which the new input is used to update the hidden state is defined by the update gate. The GRU computes its output based on the updated hidden state.

The equations used to determine the reset gate, update gate, and hidden state of a GRU are as follows:

$$\text{Reset gate: } r_t = \text{sigmoid}(W_r [h_{t-1}, x_t]) \quad (14)$$

$$\text{Update gate: } z_t = \text{sigmoid}(W_z [h_{t-1}, x_t]) \quad (15)$$

$$\text{Candidate hidden state: } h_t' = \tanh(W_h [r_t h_{t-1}, x_t]) \quad (16)$$

$$\text{Hidden state: } h_t = (1 - z_t) h_{t-1} + z_t h_t' \quad (17)$$

where x_t represents the input at time step t , h_{t-1} is the previous hidden layer, and h_t is the current hidden layer. Learnable weight matrices W_r , W_z , and W_h are also comprised.

In essence, GRU networks are a type of recurrent neural network (RNN) that effectively capture sequential input by selectively modifying the hidden state at each time step via gating mechanisms. They have shown effectiveness in several natural language processing applications, including as speech recognition, machine translation, and language modelling.

GRU is equipped with only three gates and does not retain an internal cell state. The hidden state of the Gated Recurrent Unit (GRU) encompasses the information stored in the Internal Cell State of a Long Short-Term Memory (LSTM) recurrent unit. This Gated Recurrent Unit is the recipient of the pooled information. The subsequent passage delineates the many gates of a Gated Recurrent Unit (GRU).

Revise the Gate(z): It determines the quantity of historical knowledge that has to be preserved for future generations. It is similar to the output gate of an LSTM recurrent unit.

Resetting the Gate(r) involves determining the extent to which prior events should be intentionally disregarded. It is analogous to the collaborative functioning of the Forget Gate and Input Gate in an LSTM recurrent unit.

Present Memory Gate (h_t): It is often overlooked in a typical discussion of a Gated Recurrent Unit Network. The Reset Gate incorporates it, similar to how the Input Modulation Gate is a constituent of the Input Gate. Its purpose is to ensure that the input has a mean of zero and introduce certain non-linear characteristics. The incorporation of this element into the Reset gate also functions to diminish the impact of previous data on the information now being broadcast into the future.

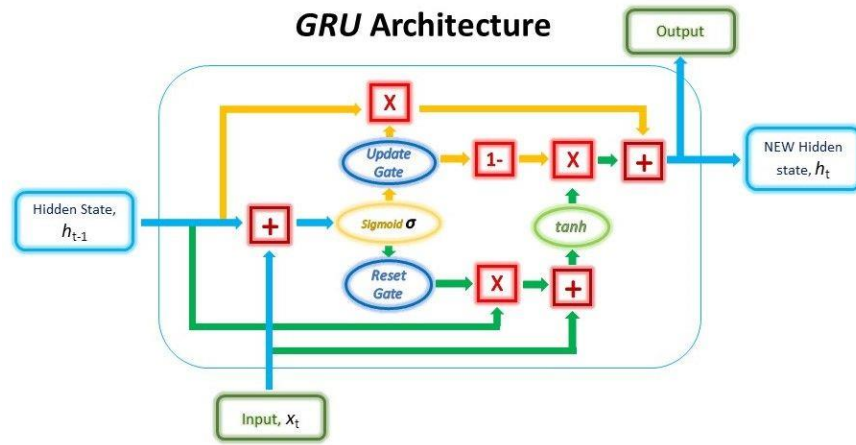
The core operation of a Gated Recurrent Unit (GRU) Network is similar to that of a standard Recurrent Neural Network (RNN) in terms of their representation. However, the key difference lies in the internal mechanism of each recurrent unit. GRU networks incorporate gates that regulate the flow of information from the current input and the previous hidden state (Shewalkar et al., 2019).

- A GRU operation

Consider two vectors as inputs: the present input and the previous hidden state. The GRU Architecture is seen in Figure 12.

Apply the suitable activation function to each member of the parameterized vectors individually, based on the gates. Below is the supplied list of gates, along with the corresponding activation function that must be utilized for each gate (Cho et al., 2014).

Revision: Sigmoid Function for Gate Update

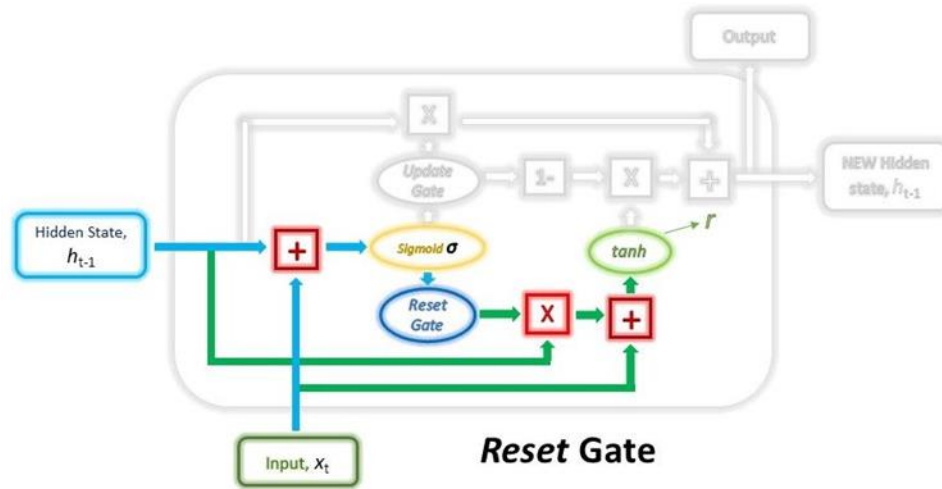
Figure 12*GRU Architecture*

Reset Gate: Sigmoid Activation Function

The calculation procedure for the Current Memory Gate exhibits a little deviation.

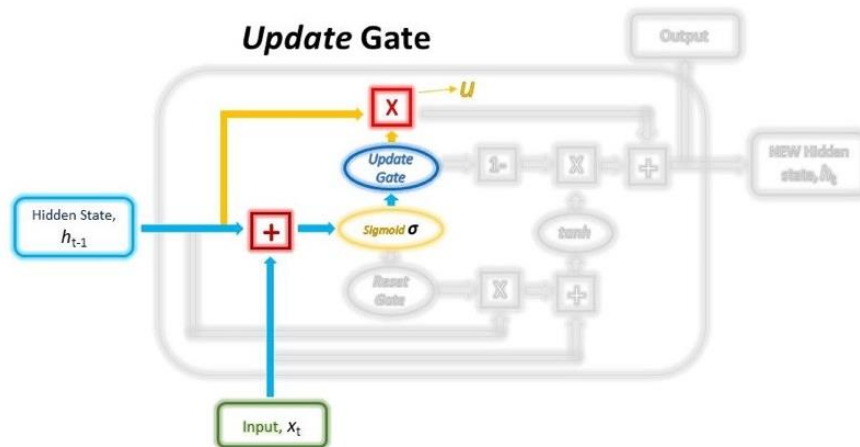
Initially, the state vector that was previously hidden and the Hadamard product of the Reset Gate are calculated. Subsequently, the vector is parameterized and appended to the pre-existing input vector that has already been parameterized (GRUs) - Coding Ninjas,” n.d.). In Figure 13, the reset gate is responsible for identifying and discarding prior information. It uses the same formula as the update gate in Equation 17.

$$r_t = \sigma(W^{(r)}x_t + U^{(r)}h_{t-1}) \quad (17)$$

Figure 13*Reset Gate*

The update gate as in Figure 14, regulates the retention of information from previous phases in order to manage long-term memory (Gangwar and Vadlamani, 2022). The equation employed in the update gate is as in Equation 18:

$$z_t = \sigma(W^{(z)}x_t + U^{(z)}h_{t-1}) \quad (18)$$

Figure 14*Updated Gate*

z_t denotes the output of the update gate at step t . x_t represents the current input, and it undergoes multiplication by its weight, $W^{(z)}$. The the previous $t-1$ steps information is stored in h_{t-1} , with $U^{(z)}$ representing the corresponding weight. After the addition process, the sigmoid activation function is implemented.

In order to calculate the final result of the Gated Recurrent Unit (GRU), we utilize the information acquired from the two gates. The reset gate is utilized to establish a fresh memory location that preserves information from previous instances (Cho et al., 2014). The computation is represented in Equation 19 as:

$$h'_t = \tanh(Wx_t + r_t \odot Uh_{t-1}) \quad (19)$$

We multiply the current input, x_t , with its weight (W), and h_{t-1} with weight U . Afterward, we calculate the Hadamard product, which refers to the element-wise multiplication, between the output of the reset gate (r_t) and Uh_{t-1} . Subsequently, we calculate the total and subsequently apply the hyperbolic tangent activation function.

For the final output (h_t) of the Gated Recurrent Unit (GRU), we utilize the update gate and h'_t , previously calculated. The h_t calculation is as in Equation 20:

$$h_t = z_t \odot h_{t-1} + (1 - z_t) \odot h'_t \quad (20)$$

The output of the Gated Recurrent Unit (GRU) is gained by taking the Hadamard product of the update gate (z_t) and h_{t-1} , and similarly, the Hadamard product of $1 - z_t$ and h'_t . The final output is then determined by adding these results.

GRU efficiently resolves the issue of the vanishing gradient problem by preserving relevant information and propagating it to succeeding stages. With proper training, GRU has the capacity to achieve exceptional performance.

2.5.4 Artificial Neural Network

Artificial Neural Networks (ANNs) are computational models that imitate the operation of biological systems made up of "neurons." They are inspired by the central nervous systems of animals. They demonstrate proficiency in machine learning and pattern recognition, appearing as linked networks of "neurons" that calculate values based on inputs.

A neural network is a machine learning technique that is based on the structure and function of a human neuron. The human brain, comprised of many neurons, processes data by electrical and chemical impulses. Neurons are connected by specialized structures known as synapses, which enable the passage of messages. Artificial Neural Networks (ANNs) are created by emulating vast quantities of linked neurons.

Artificial Neural Networks (ANNs) imitate the cognitive process of the human brain in processing information. They comprise a multitude of interconnected processing units working together to process and extract significant outcomes from data. Neural networks are very adaptable and may be used for various tasks, including both classification and regression of continuous target characteristics (Sharma, 2021).

Neural networks are widely used in several fields, including economics, forensics, and data mining for pattern detection. In addition to categorization, they are utilized for analyzing extensive datasets following thorough training. Studies suggest that the neural network software sector is expected to expand at a Compound Annual Growth Rate (CAGR) of 33.5% from 2019 to 2026 ("Compound Annual Growth Rate (CAGR) Formula and Calculation," n.d.).

A neural network generally comprises three layers:

1. Input layer: The units in this layer represent the unprocessed data that is used as input for the network.

2. Hidden layer(s): The behaviour of each hidden unit is influenced by the actions of the input units and the weights assigned to the connections between the input and hidden units. There might potentially exist one or several concealed layers.

3. The output layer is affected by the activity of the hidden units and the weights connecting the hidden and output units.

- *The structure of an artificial neural network:*

An Artificial Neural Network is commonly organized into layers, consisting of linked nodes, each with an activation function (Gupta, n.d.). The network typically has three main layers:

a. *Input layer:* The input layer's function is to accept the values of explanatory attributes for each observation. The number of input nodes in this layer usually corresponds to the number of explanatory variables. The input layer serves the purpose of introducing patterns to the network and then sending this information to one or more hidden layers. The nodes in the input layer are inert, simply replicating incoming values to their many outputs. The input layer duplicates each value and transmits it to all hidden nodes.

b. *Hidden layer:* The hidden layers in the network perform predetermined operations on the input values. Hidden nodes or input nodes are sources of incoming arcs for each node. The hidden layer, which is responsible for the actual processing, employs a network of weighted connections. There is a possibility of the existence of several concealed levels. The values that are input into a hidden node are multiplied by pre-determined weights, which together contribute to the calculation of a single numerical value.

c. *Output layer:* The output layer is connected to hidden layers, which in turn receive connections from either other hidden levels or the input layer. This layer generates an output

value that corresponds to the prediction of the response variable. In categorization settings, it is common to have only one output node. The active nodes in the output layer cooperate to process the data and provide the ultimate output values.

The efficacy of the neural network in altering data relies on the meticulous choice of weights, setting it apart from standard information processing approaches.

2.6 Hybrid Deep Learning Approaches

Hybrid deep learning has arisen as a result of the ongoing development in artificial intelligence and machine learning, as researchers strive to create more robust and adaptable models. Hybrid deep learning is a significant change in approach that combines the advantages of several neural network structures to provide solutions that are more resilient, adaptable, and efficient .

Deep learning has transformed several fields, such as computer vision and natural language processing, by utilizing its hierarchical feature representation and ability to autonomously learn complex patterns. Nevertheless, some architectural designs may include intrinsic constraints when employed for specific functions. Hybrid deep learning overcomes these constraints by combining the characteristics of several neural network designs.

This introduction delves into the notion of hybrid deep learning, examining its motivations and the promise it possesses for addressing intricate challenges in several disciplines. Hybrid models seek to enhance performance and flexibility by combining several designs, such as Convolutional Neural Networks (CNNs), Recurrent Neural Networks (RNNs), and others, in order to exploit the distinctive strengths of each component (Sarker, 2021).

- **Reasons for using Hybrid Deep Learning:**

1. *Thorough Feature Extraction:* Different deep learning architectures may excel in extracting specific types of features. CNNs excel at collecting spatial information, but RNNs are specifically developed for sequential data. Hybrid models integrate these skills to create a more thorough feature extraction, which is essential for comprehending intricate patterns in varied datasets.

2. *Tackling Task-Specific Challenges:* Some activities may require specific and tailored techniques. Hybrid deep learning enables academics and practitioners to customize models to effectively tackle specific issues related to a particular activity. Through the integration of components that are optimized for various parts of a task, these models have the ability to offer more refined and efficient solutions.

3. *Enhanced Generalization:* The fusion of several designs leads to enhanced generalization. Hybrid models has the ability to adjust to a wider array of input variables and environmental circumstances, rendering them more resilient and dependable in practical scenarios.

- *Applications of Hybrid Deep Learning:*

1. *Computer Vision:* In the field of computer vision, hybrid models can provide improved capabilities for object detection, scene understanding, and picture production by effectively combining spatial and sequential pattern analysis (Mahony et al., 2020).

2. *Natural Language Processing:* In natural language processing tasks such as sentiment analysis or machine translation, hybrid Deep Learning Models may integrate the contextual comprehension of recurrent neural networks (RNNs) with the hierarchical feature extraction capabilities of convolutional neural networks (CNNs) to provide more precise and contextually-aware predictions (Otter et al., 2021).

3. *Speech Recognition*: Speech recognition involves the integration of architectures such as Bidirectional Long Short-Term Memory (LSTM) with Convolutional Neural Networks (CNNs). This combination allows for the capturing of both temporal dependencies and spectral information, resulting in enhanced accuracy when transcribing spoken words (Zhang et al., 2018).

4. *Healthcare*: Hybrid models have potential uses in the healthcare industry, particularly in processing varied data types like medical pictures and patient records. These models leverage a blend of specialized architectures to properly handle such data (Miotto et al., 2018).

As we explore the field of hybrid deep learning, we set out on a quest to discover uncharted territories in artificial intelligence. Hybrid models combine the capabilities of several neural network designs to bring about a new age of adaptable, efficient, and high-performing solutions in diverse fields.

This thesis explores the use of hybrid Deep Learning Models in automated speech recognition (ASR). These models combine distinct deep learning architectures to use the characteristics of each, resulting in improved performance of ASR systems. The hybrid models include the CNN-Bi-LSTM hybrid model, the GRU-Bi-LSTM hybrid model, and the CNN-GRU hybrid model.

2.6.1 CNN-Bi-LSTM Hybrid Model

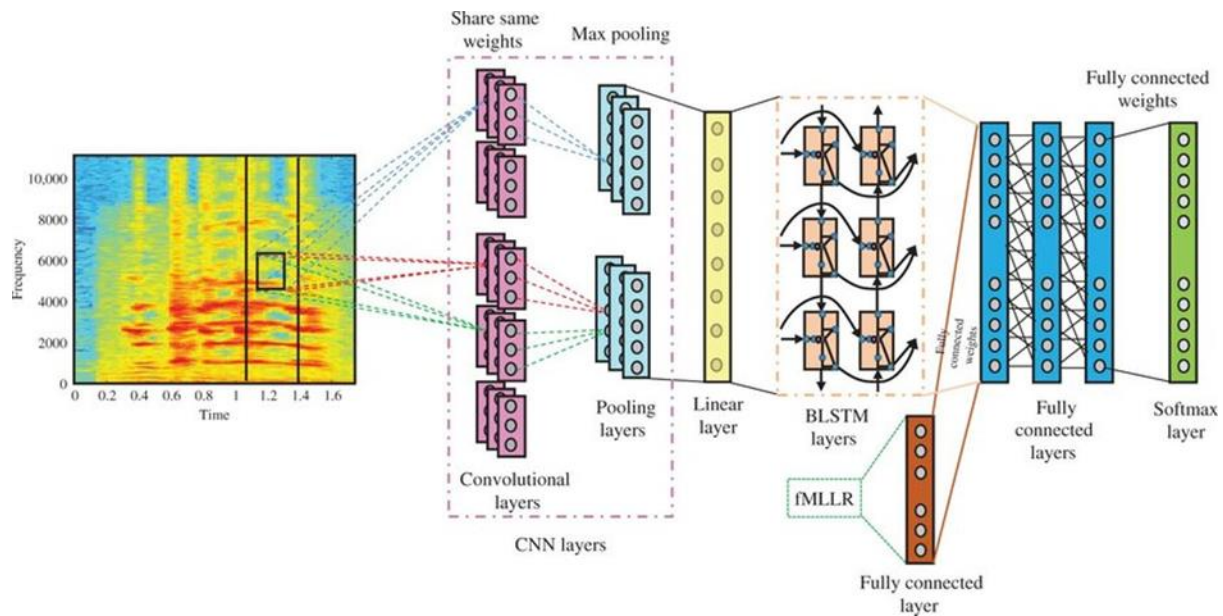
The area of speech recognition technology has made significant progress, and the CNN-Bi-LSTM hybrid model is a clear example of the novel methods driving these developments. The integration of Bidirectional Long Short-Term Memory (Bi-LSTM) and Convolutional Neural Network (CNN) architectures in this model has shown outstanding capabilities in Automatic

Speech Recognition (ASR). This essay examines the design, operation, and importance of the CNN-Bi-LSTM model in the context of (ASR).

- **CNN-Bi-LSTM Model Architecture**

The CNN-Bi-LSTM model synergistically integrates the advantages of both Bi-LSTM and CNN architectures. The Bidirectional Long Short-Term Memory (LSTM) component effectively captures contextual dependencies by analyzing input sequences in both the forward and backward directions. The bidirectional nature of the model enables it to comprehend the subtleties of spoken language by taking into account information from both preceding and subsequent contexts.

At the same time, the Convolutional Neural Network component, known for its ability to extract spatial features, examines spectrogram representations of voice sounds. The Convolutional Neural Network (CNN) effectively records specific patterns within a certain locality, hence improving the model's capability to distinguish pertinent auditory characteristics. The fusion of these two architectures yields a hybrid model that demonstrates exceptional proficiency in acquiring complex temporal and spatial representations that are vital for Automatic Speech Recognition (ASR). Figure 15 depicts the CNN-BiLSTM architecture for Automatic Speech Recognition (ASR) as proposed by (Passricha and Aggarwal, 2020).

Figure 15*CNN-BI-LSTM Architecture*

- *Functioning of the CNN-Bi-LSTM Model in ASR*

The operation of the CNN-Bi-LSTM model in ASR entails several steps. Firstly, the model takes in audio signals and converts them into spectrogram representations. The CNN component then carries out spatial feature extraction, emphasizing important patterns in the spectrogram. At the same time, the Bi-LSTM component captures long-term dependencies and contextual information. Finally, the model's last layers combine these extracted features to forecast the corresponding transcription, effectively transforming spoken language into text.

- *Importance in Automatic Speech Recognition (ASR) Applications:*

The CNN-Bi-LSTM hybrid model is highly important in ASR applications due to its enhanced accuracy and resilience in comparison to conventional models. The capacity to accurately capture both the time-based and spatial characteristics is in accordance with the intricate nature of spoken language. This attribute renders it highly beneficial in a wide range of

applications, spanning from voice assistants to transcription services and speech-activated gadgets (Mishra et al., 2023) .

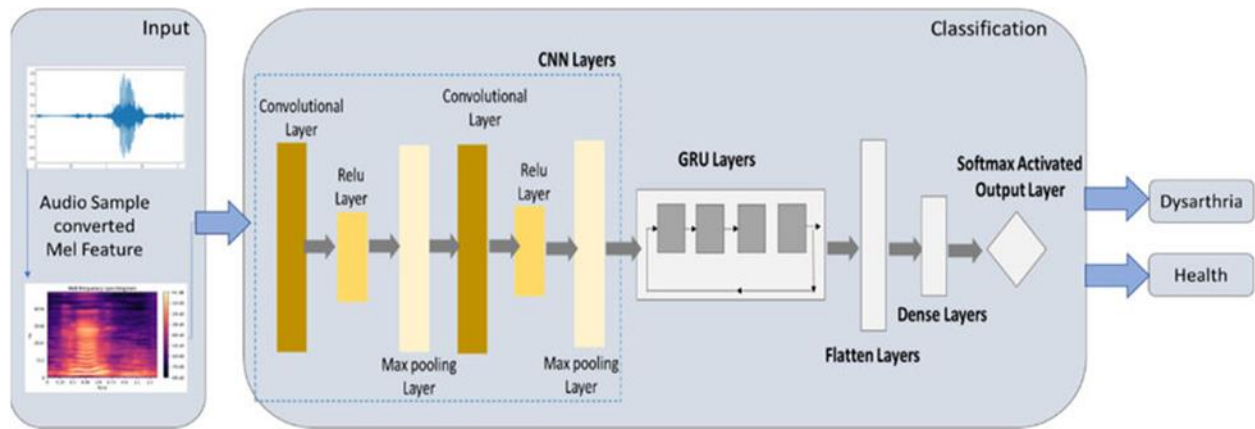
To summarize, the CNN-Bi-LSTM hybrid model is a significant breakthrough in the field of Automatic Speech Recognition. This model demonstrates the effectiveness of combining Bidirectional Long Short-Term Memory and Convolutional Neural Network architectures to address the challenges of spoken language. The CNN-Bi-LSTM model represents a significant advancement in technology, offering improved precision and adaptability in the field of voice recognition as it continues to develop.

2.6.2 CNN-GRU Hybrid Model

The CNN-GRU Hybrid Model is an advanced fusion technique that achieves exceptional performance in Automatic Speech Recognition. The CNN-GRU model is a revolutionary design in the field of Automatic Speech Recognition (ASR) that effectively integrates Convolutional Neural Networks (CNNs) with Gated Recurrent Units (GRUs). This essay examines the structure, operation, and the influential effect of the CNN-GRU model in improving the precision and effectiveness of ASR systems.

- *The architecture of the CNN-GRU model is as follows:*

The CNN-GRU model combines the advantages of CNNs and GRUs to form a hybrid architecture that effectively captures spatial and temporal characteristics in voice data. The CNN module analyzes input spectrogram representations and extracts spatial characteristics using convolutional layers. The extracted characteristics are subsequently inputted into the GRU units, which effectively capture sequential dependencies and contextual information. Figure 16 visually illustrates the layers and links in the CNN-GRU architecture, as presented by (Shih et al., 2022).

Figure 16*CNN-GRU Architecture*

- *Operation of the CNN-GRU Model In Arabic Speech Emotion Recognition:*

The CNN-GRU model functions by initially converting unprocessed audio inputs into spectrogram representations. The CNN analyzes these representations, emphasizing spatial aspects that are essential for comprehending speech. The resulting data is subsequently inputted into the GRU units, which possess the capability to flexibly adjust to the sequential environment. This enables them to effectively capture long-term connections and improve the model's capacity to comprehend the subtleties of spoken language. The integrated design guarantees thorough feature extraction and contextual comprehension.

- **Significance in ASR Applications:**

The CNN-GRU model is highly significant in ASR applications, providing a flexible solution for transcription services, voice assistants, and other speech-activated technologies. The model's capacity to capture both spatial and temporal characteristics makes it a strong and flexible choice for a wide range of real-world applications.

To summarize, the CNN-GRU model is a state-of-the-art development in the field of Automatic Speech Recognition. This hybrid architecture effectively tackles crucial issues in ASR

by combining the spatial feature extraction capabilities of CNNs with the contextual understanding and sequential processing of GRUs. The CNN-GRU model exemplifies the efficacy of fusion in developing highly precise and efficient speech recognition systems, showcasing the impact of technological advancements.

2.6.3 GRU-BI-LSTM Hybrid Model

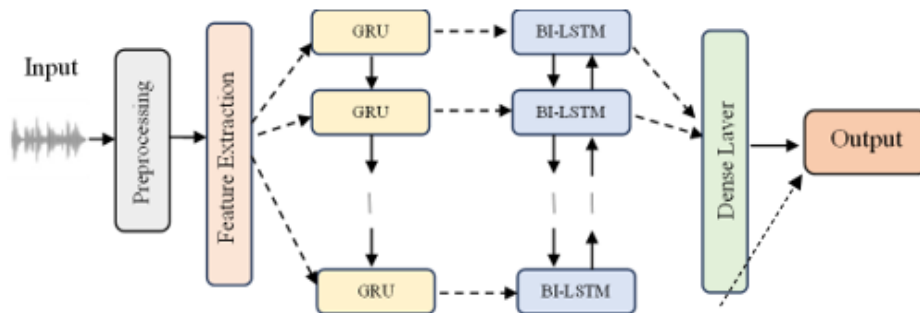
The GRU-BI-LSTM model is an innovative design in the domain of Automatic Speech Recognition (ASR), combining the advantages of Gated Recurrent Units (GRUs) and Bidirectional Long Short-Term Memory (Bi-LSTM) networks. This essay explores the structure, operation, and significant influence of the GRU-BiLSTM model in improving the precision and effectiveness of ASR systems (Shen et al., 2022).

- **GRU-BiLSTM Model Architecture:**

The GRU-BiLSTM model combines two robust versions of recurrent neural networks (RNNs): GRUs and BiLSTMs. Every GRU unit consists of reset and update gates, which determine the information to be forgotten and updated, accordingly. Concurrently, the BiLSTM, which is an enhanced version of the conventional LSTM, analyzes input sequences in both the forward and reverse orientations. Figure 17 depicts the hierarchical structure and interconnections inside the GRU-BiLSTM architecture.

Figure 17

GRU-BI-LSTM Architecture



The GRU-BiLSTM model effectively captures sequential dependencies and contextual information inside speech signals, making it very proficient in ASR. Within the ASR pipeline, the initial step involves converting raw audio inputs into spectrogram representations. The GRU-BiLSTM model takes these representations as input. Within the model, the GRU units have the ability to selectively retain or forget information, while the BiLSTM analyzes the sequential context in both forward and backward directions. The model exhibits dynamic adaptability to fluctuations in speech patterns, hence enhancing the precision of transcriptions (Shen et al., 2022).

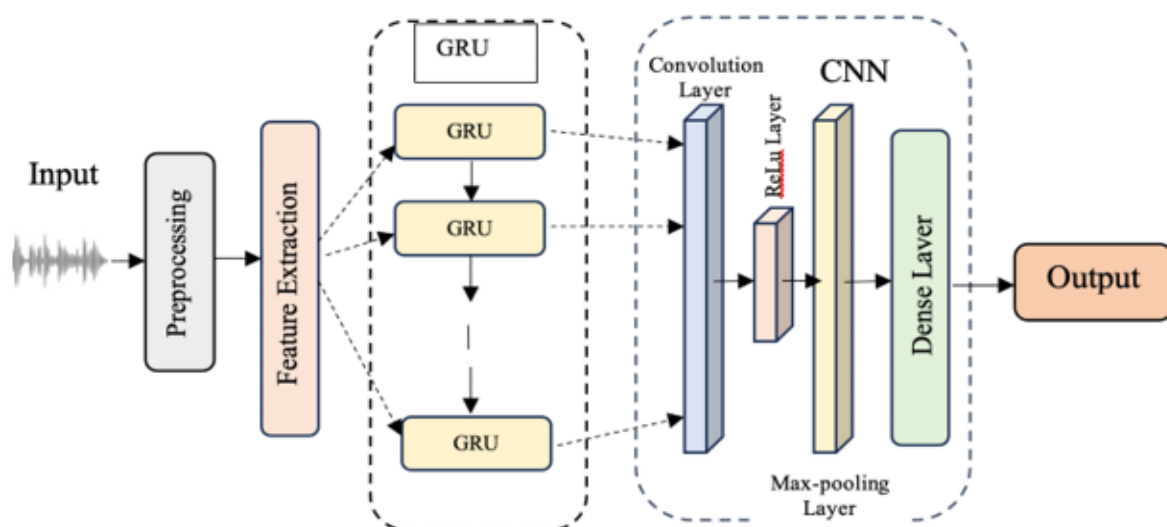
2.6.4 GRU-CNN Hybrid Model

- **The GRU-CNN architecture**

The GRU-CNN architecture is a hybrid model that combines Gated Recurrent Units (GRU) and Convolutional Neural Networks (CNN) to improve performance in Automatic Speech Recognition (ASR) (Cho et al., 2014). Figure 18 shows the GRU-CNN architecture. The architecture generally comprises of two primary components:

Figure 18

GRU-CNN Architecture



1. **Gated Recurrent Unit (GRU):** The GRU is a kind of recurrent neural network (RNN) that is particularly effective at capturing the sequential dependencies present in the input data. The network has gated techniques to regulate the information flow, enabling it to better capture long-term dependencies compared to conventional RNNs.

2. **Convolutional Neural Network (CNN):** CNNs are particularly adept at extracting spatial characteristics and patterns from input data. Within the field of Automatic Speech Recognition (ASR), Convolutional Neural Networks (CNNs) are utilized to capture localized connections within auditory characteristics, detecting significant patterns in both the temporal and spectral domains.

- **Benefits of using GRU-CNN in Automatic Speech Recognition (ASR):**

1. The GRU-CNN model effectively analyzes the temporal and spatial aspects by using the advantages of both architectures. This allows it to capture the sequential relationships and spatial patterns seen in voice signals (Girirajan and Pandian, 2022).

2. **Enhanced Generalization:** The hybrid model frequently demonstrates a superior ability to generalize to unfamiliar data, hence mitigating overfitting in comparison to employing separate designs.

3. **Improved Feature Learning:** The model's capacity to extract features at both sequential and spatial levels boosts its ability to comprehend and depict intricate patterns in speech.

2.6.5 GRU-ANN Hybrid Model

The architecture of the GRU-ANN (Gated Recurrent Unit - Artificial Neural Network) is as follows:

The GRU-ANN architecture is a hybrid model that combines Gated Recurrent Units (GRU), a specific form of recurrent neural network (RNN), with an Artificial Neural Network

(ANN). This combination is specifically engineered to optimize the functionality of Automatic Speech Recognition (ASR) by capitalizing on the advantages of both systems (Cho et al., 2014).

The architecture generally comprises the subsequent components:

1. The Gated Recurrent Unit (GRU) is utilized to capture sequential dependencies in the input voice data. The network incorporates gating techniques that regulate the information flow, enabling it to capture long-term relationships with more efficacy compared to conventional RNNs.

2. The Artificial Neural Network (ANN) component is responsible for the processing and analysis of characteristics that have been retrieved from the voice stream. The model is composed of numerous layers of linked nodes, allowing it to acquire intricate patterns and correlations within the data.

- **Benefits of using GRU-ANN in Automatic Speech Recognition (ASR):**

1. *Effective Sequence Learning:* The Gated Recurrent Unit (GRU) is very proficient in capturing sequential relationships, which is essential for comprehending speech patterns. On the other hand, Artificial Neural Networks (ANN) play a significant role in acquiring hierarchical representations.

2. *Enhanced Feature Representation:* The integration of GRU and ANN enables the model to acquire intricate and informative representations, hence improving its capacity to distinguish between various speech patterns.

3. *Mitigated Overfitting:* The hybrid model exhibits more resilience and lower susceptibility to overfitting in comparison to employing separate architectures, since it possesses the ability to effectively generalize to unfamiliar data.

2.7 Conclusion

The chapter presents an overview of various Deep Learning Models for Arabic speech recognition, including CNN, Bi-LSTM, GRU, ANN, and their hybrid variations (CNN-Bi-LSTM, GRU-Bi-LSTM, CNN-GRU). It describes the architecture, functionality, and relevance of these models for Automatic Speech Recognition (ASR). The differences between Egyptian Arabic and modern standard Arabic (MSA) are explored in terms of linguistic aspects (phonological, morphological, syntactic, lexical, semantic, and pragmatic-cultural differences), highlighting challenges in navigating between the two.

Additionally, the chapter delves into sentiment analysis, focusing on its classification, phases, tasks, and challenges in Arabic, particularly in social media contexts. It also outlines feature selection techniques and examines sentiment analysis methods used in academic research. The final section discusses emotion recognition in speech, addressing various features, algorithms, and representation learning techniques used to detect emotions, as well as the difficulties and limitations of this field.

Chapter 3: Literature Review

The proliferation of voice assistants in recent years may be attributed to the widespread popularity of Google Home, Amazon Echo, Siri, Cortana, and other similar models. The most prominent uses of automatic speech recognition (ASR) are as follows. These applications convert spoken words into written text by initially utilizing a sample of spoken audio in a certain language. Therefore, these algorithms are commonly known as Speech-to-Text algorithms. For this reason, they are sometimes known as speech-to-text algorithms. The traditional ASR system consists of modules for language modelling, pronunciation lexicon, and acoustic modelling. Collectively, these systems facilitate many capabilities such as vocal pattern recognition, speaker authentication, dialect identification, and additional functionalities, all derived from a solitary speech signal.

3.1 Emotions in Human Being

Emotions refer to intricate mental states that impact changes in behavior, encompassing both psychological and physical aspects. Pleasant emotions are crucial for human survival since they have the potential to enhance productivity in the workplace, whilst negative emotions can lead to health issues (Shu et al., 2019). Given that emotions serve as fundamental indicators of individuals' internal moods, the investigation of emotions has broadened to encompass other academic disciplines. Scientists conducting research in the realm of computer science are exploring methods to accurately detect, reproduce, and elicit emotional responses (Wadley et al., 2022), psychology (Moskowitz et al., 2021) neuroscience (Ashhad et al., 2022), medicine (Laput et al., 2016), and education (Simonton et al., 2021).

3.1.1 Emotions Depiction

Currently, the investigation of human emotion encompasses several disciplines such as computer science, affective computing, and psychology. Multiple depictions of emotions have been employed in the realm of research, as evidenced by the works (Ashhad et al., 2022, 2022; Ekman, 1992; Gu et al., 2016). There are two main branches of these representations: category and dimensional representation. Emotions, such as pleasure, sadness, and anger, are categorized and represented using a single label in the category representation paradigm. Although there are other category representations of emotions ((Jack et al., 2014) and Gu et al., 2016), all of them may be traced back to basic emotions as proposed by Ekman (1992). (Rubin and Talarico, 2009) propose that the dimensional representation model conceptualizes emotions as a singular point inside a multi-dimensional space.

The circumplex model is a commonly employed and easily understandable dimensional framework for emotions, consisting of two dimensions: valence (positive vs negative) and arousal (high versus low). There is currently no definitive agreement on the choice of representation model. Consequently, as stated by (Buechel and Hahn, 2016), scholars and professionals choose the approach that they consider most relevant to their specific use case. The preferences and understanding of the end users, however, are essential factors to consider when designing a system for the general public. Two further representations of emotions are the constructive model proposed by (Barrett, 2006) and the evaluation model of emotions developed by (Scherer et al., 2001). Nevertheless, those are not well regarded in the domain of emotional computing.

3.2 Arabic Speech Emotion Recognition Overview

Developing a dependable Arabic speech recognition system necessitates interdisciplinary collaboration. In order to address the challenges posed by language complexities, the

development of a robust Arabic Automatic Speech Recognition (ASR) system requires the integration of many Natural Language Processing (NLP) components, in addition to managing the traditional acoustic pipeline. The challenges inherent to Arabic speech recognition can be summarized as follows:

1. Arabic is a language that has a lot of word forms, and the use of word parts is often used to reduce the occurrence of out-of-vocabulary words, as explained in section 2.4.
2. Arabic contains diacritical marks representing short vowels, which are often disregarded in written language, as seen in section 2.2.
3. Insufficient labelled data is accessible for Arabic research.

There exist several variations of Arabic, each characterized by distinct vocabulary and sound.

Prior research on Arabic speech recognition has been conducted before this thesis. In this document, we want to present a concise overview of previous studies conducted on the topic of Arabic automated voice recognition. Multiple research laboratories, totaling more than eight, resolved the second and first concerns mentioned earlier during the 2002 Johns Hopkins Summer Workshop on Arabic ASR (Kirchhoff et al., 2003). In addition, they have conducted study on the distinctions between formal and dialectal Arabic. Significant enhancements in word error rate were seen on the LDC Arabic Call Home data¹, along with a proposed novel approach for automatic vowel restoration, morphology-based language modelling, and integration of out-of-corpus language model data.

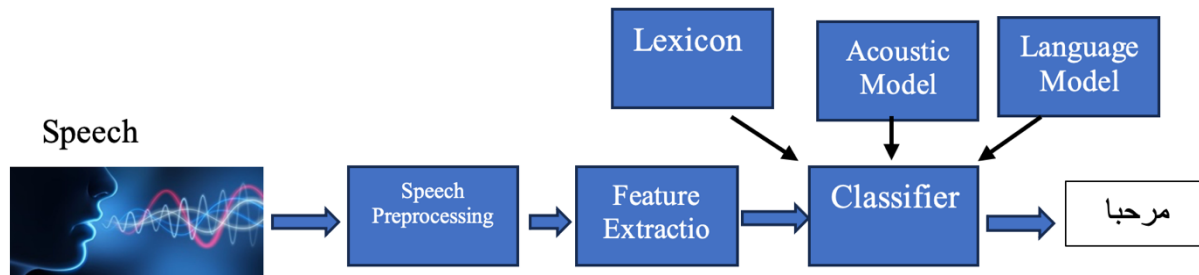
3.3 Arabic Emotion Speech Recognition Architecture

Figure 5 illustrates that the ASR system consists of two primary components: the front end and the classifier. The front-end block is utilized to derive the spectrum representation of the

speech waveform. The name is Melanie Rhythm. The most often utilized characteristics are Cepstral Coefficients (MFCC). The classifier block analyzes the incoming audio data to identify the most accurate word sequence match, using the lexicon, language model, and acoustic model.

Figure 19

ASR Architecture



Speech data can be encoded as an observation 'O' or as a sequence of speech vectors. To ascertain the most probable sequence of words, one must solve Equation 3, commonly known as the fundamental equation of statistical voice recognition (“The Handbook of Computational Linguistics and Natural Language Processing | Wiley,” n.d.).

$$W_h = \operatorname{argmax}_w P(w|o) \quad (3)$$

where:

- o : a series of observations or speech vectors which describes the input voice data.
- W_h : the process of translating audio from acoustic speech into a matching word sequence.
- argmax_w : is the search space and how the vocabulary functions.

In the field of voice recognition, the decomposition of Equation (3) into the probability $P(o|w)$, the prior $P(w)$, and the denominator $P(o)$ is achieved using the Bayes rule. It is important to note that the denominator, which is independent of the word sequence and does not affect the word sequence search, is nonetheless a fundamental part of the process. As a result, $P(o)$ is

excluded (“The Handbook of Computational Linguistics and Natural Language Processing | Wiley,” n.d.) as explained in Eq. (4)

$$W_h = \operatorname{argmax}_w \frac{P(O|W)P(w)}{P(o)} \quad (4)$$

$$W_h = \operatorname{argmax}_w P(w|o)P(o)$$

The linguistic model is represented by the symbol $P(w)$, but the probability of the acoustic model is commonly represented as $P(o|w)$. The language model probability $P(w)$ may be calculated independently and using different corpora, as it is not dependent on acoustics.

Equation 4 demonstrates the partitioning of the speech recognition system development process into two fundamental modules: acoustic modelling and linguistic modelling. The objective of acoustic modelling is to develop a model that can effectively interpret speech signals when there is an observation vector present. The raw waveform undergoes front-end signal processing to produce the acoustic observation vector, denoted as $o = [o_1, o_2, \dots, o_T]$. An ideal scenario would include this vector remaining unchanged in the presence of non-speech related factors, such as background noise, speaker characteristics, and changes in pronunciation. However, it is important to note that not all of the variability can be standardized at the feature processing step. Consequently, the acoustic models will need to divide the burden. Conversely, language models should strive to predict the prior distribution of the word sequence w before detecting speech signals. In classical language models, the distribution of each word is thought to depend on the previous $n-k$ words. This dependency is established by the frequency of n -grams. The variable k represents the order or history of the language model. The final section of this chapter provides an analysis of the latest advancements in Arabic speech recognition technology. The subsequent sections provide a synopsis of a standard speech recognition system.

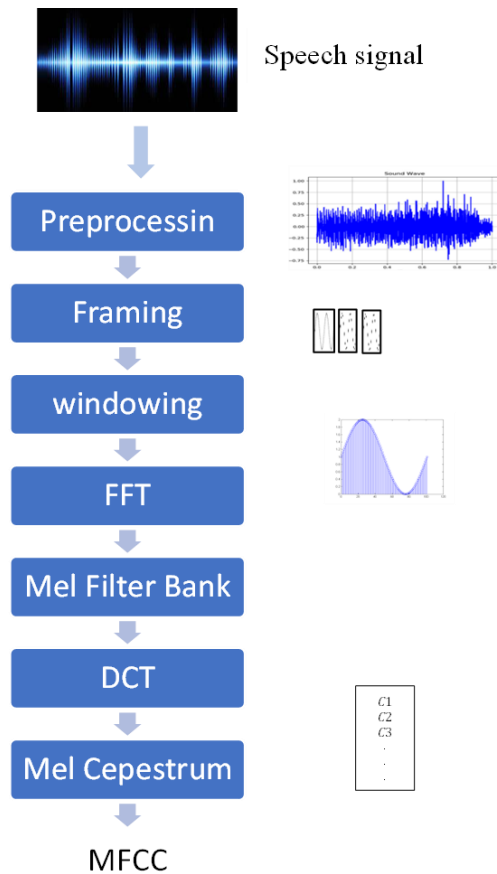
3.4 Feature Extraction

Feature extraction is a crucial component of the speech recognition process as it collects significant information from sample speech. A waveform in continuous time and magnitude format is received. The primary goal of the front end of signal processing is to extract the acoustic attributes that will be used in acoustic modelling. This is achieved by sampling the raw acoustic waveform and converting it into feature vectors. The voice recognition system will utilize a compact acoustic feature representation that aims to preserve the maximum amount of signal information while minimizing the impact of speaker and environmental acoustic variations. Furthermore, in addition to doing spectrum analysis, it usually captures a frame of the speech signal every 16-32 milliseconds and refreshes it every 8-16 milliseconds (Hermansky, 1990) and (Xie and Liu, 2006). The regular frontend utilizes mathematical components such as Discrete Cosine Transformation (DCT), Logarithm Calculation (LOG), Fast Fourier Transformation (FFT), and Linear Discriminate Analysis (LDA) to find features in speech. The Mel Frequency Cepstral Coefficients (MFCC) method, which utilizes the Mel scale and operates in the frequency domain, is widely recognized as the most common approach (Xie and Liu, 2006). The speech signal is subsequently divided into discrete time intervals, each containing a variable number of samples. Most systems utilize frame overlapping to ensure a smooth transition from one frame to the next. Subsequently, a Hamming window is employed on every time frame to eliminate any abrupt changes at the boundaries. Subsequently, to isolate the frequency components of a signal in the temporal domain, the Fast Fourier Transformation (FFT) is computed for each frame. Once the frame has undergone Fourier transformation, the logarithmic Mel-Scaled filter bank is employed. The actual frequency is converted to the frequency perceived by humans using the mel scale. The calculation of the Discrete Cosine

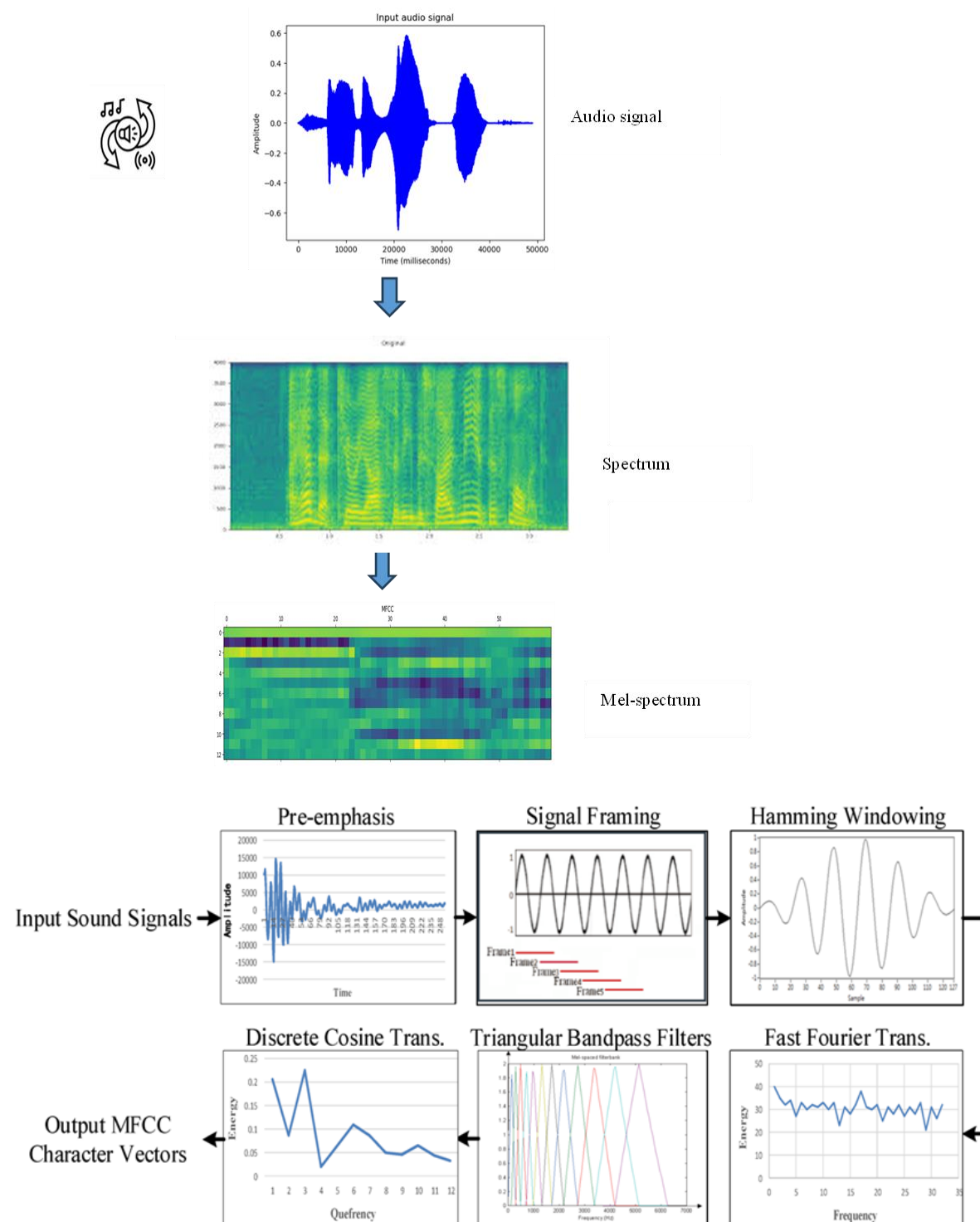
Transformation (DCT) of the outputs from the filter bank is the last step to be performed. Figure 21 depicts the complete process of extracting MFCCs.

Figure 20

Feature Extraction Diagram



Each speech frame is analyzed to calculate a collection of Mel-frequency cepstral coefficients (MFCC). The acoustic vector, which is a set of coefficients, is extremely useful for further analysis and processing in speech recognition since it accurately represents the phonetically important parts of speech. The process of extracting MFCC characteristic vectors is illustrated in Figure 22 (Tsai et al., 2019). Figure 23 depicts the conversion of a speech signal from a .wav file to a Mel-spectrum.

Figure 21*Transferring Audio to Mel-Spectrum*

3.5 Neural Network Acoustic Model

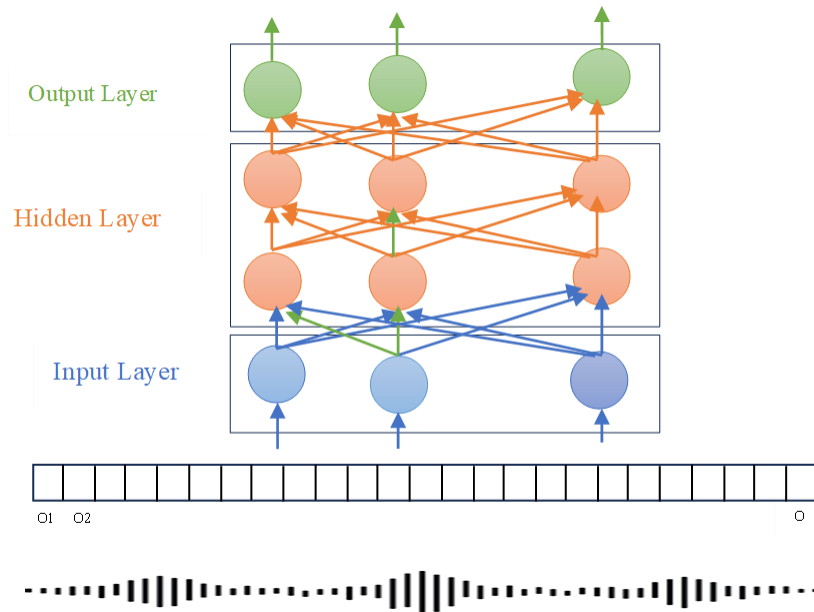
Since the middle of the 1980s, speech recognition for acoustic modelling has drawn a lot of attention in neural network-based systems. Phoneme identification (Waibel et al., 1989) was one of the earliest applications that proved successful. The recognition of single words was then included (Bottou et al., 1989). Additionally, continuous voice recognition has shown promising results (Haffner, 1992), but with limited vocabulary. Hidden Markov Model / Artificial Neural Network (HMM/ANN) hybrid technique was developed concurrently (Bourlard and Wellekens, 1990), (Bourlard and Morgan, 1994) and (Renals et al., 1994).

The objective of acoustic modelling is to assign the correct label to each frame, typically at intervals of 10 milliseconds. The neural network analyzes the probability of each phone based on the input feature data, treating it as a problem of estimating probabilities. In contrast, the Gaussian mixture model represents the acoustics by applying generative probability. Over two decades ago, researchers employed a hybrid methodology to investigate the utilization of multilayer perceptron in the field of voice recognition. Achieving favorable outcomes was possible through the anticipation of HMM states by utilizing a solitary hidden layer of a neural network to analyze acoustic coefficients inside certain time intervals (Bourlard and Morgan, 1994). Nevertheless, the available hardware and learning method were inadequate to effectively train a neural network with several hidden layers on a substantial amount of data. This summation is highly commendable (Bourlard and Morgan, 1989). Recent improvements in computer technology and machine learning algorithms have led to the development of more efficient approaches for training deep neural networks (DNNs) with several hidden layers and a large output layer. These techniques are specifically designed to enhance the performance of speech recognition systems that require a huge vocabulary. The primary objective of this is to

manage the multitude of hidden Markov model (HMM) states that arise from representing each phone with multiple tri-phone HMMs, which consider the neighboring phones. These triphone HMMs can include a large number of interconnected states, known as context-dependent states, even if some of the states are linked together (Hinton et al., 2012).

Based on the provided input data, we may utilize the neural network's output to gauge the probability of the phone's arrival. Typically, phone prediction involves using a window of λ frames context, where λ is usually between 3 and 4, rather than just a single frame. Figure 10 shows a context consisting of three frames to the right and three frames to the left for the network's central frame input, referred to as O_t . Furthermore, it showcases a feed-forward neural network including a solitary input layer capable of handling waveform samples derived from an MFCC, PLP, or filter bank. A Mel-frequency cepstral coefficient (MFCC) feature vector consists of 39 dimensions per frame. The feature vector consists of 273 Mel-frequency cepstral coefficient (MFCC) inputs, obtained by multiplying 7 frames (3 frames of left context, the center phone, and 3 frames of right context) by 39. Following are one output layer and two hidden layers. The hidden layers assist in constructing comprehensive representations capable of accommodating various changes in feature vector coefficients, such as differences in accent and speech pace.

Back-propagation is employed to calculate the required gradients, whilst gradient descent is utilized to estimate the weights and biases at each layer. For a classification task in a neural network, there are several error functions available, such as mean square error (MSE).

Figure 22*Transferring Audio to Mel-Spectrum*

- **Deep Neural Network**

Deep neural networks (DNNs) emerged from the development of hybrid HMM/ANN systems, which proved to be effective. In recent times, there has been a notable increase in computational capacity. These neural network architectures have several hidden layers. The activation function is indicated as $h(\cdot)$, while the weight matrix of layer n is represented by M_n . Research has shown that this approach surpasses the traditional MLP with a single hidden layer in speech recognition tasks (Hinton et al., 2012). However, it is well acknowledged that training these types of networks may be difficult, especially when there is a scarcity of data. Pre-training techniques have been developed to address this issue. These approaches frequently depend on utilizing unsupervised generative models to provide "good" intermediate representations, as shown in Equation 5:

$$Y_{out} = h(M_n h(M_{n-1} \dots h(M_1 x)) \quad (5)$$

A training session on identification may begin with these representations. approaches such as noisy auto-encoder approaches and greedy layer-wise training have been proposed. Deep Belief Networks (DBNs) are widely employed in the field of voice processing.

The objective of this pretraining technique is to maximize the likelihood in order to optimize the combined probability of the data and labels. The framework of constrained Boltzmann machines forms the basis of the system. Additional methods for regularization have been suggested, such as the implementation of the dropout methodology. For each training update, a fraction of the weights are randomly set to zero as the foundation of this approach. This approach enhances the network's ability to generalize, resulting in less interdependence across neurons.

- **Recurrent Neural Networks**

A type of neural network that utilizes a directed graph to establish connections between its units, also known as neurons, is referred to as a recurrent neural network (RNN) (Elman, 1990). In other words, an RNN-based model can classify an example at a certain time by leveraging the predictions made on previous instances. An alternative proposal is to develop bi-directional recurrent neural networks (RNNs), as proposed by (Schuster and Paliwal, 1997). These networks have two RNNs, each responsible for processing information in a specific direction. This indicates that the predictor has the ability to provide projections in both forward and backward directions at any given instant.

Recurrent neural network systems have been suggested for voice recognition by utilizing hidden Markov models (HMMs). The paper describes a recurrent neural network technique termed alpha-net (Schuster and Paliwal, 1997) within the HMM framework. (Robinson, 1994) presented a recurrent network that calculates its output based on the current input and a hidden

state variable that is influenced by all prior inputs. The main limitation of these models is the phenomenon of the fading gradient, which hinders the ability to capture long-range context. Research conducted by (Mubarak et al., 2021) has shown that long-short-term memory networks (LSTM) can effectively address this problem.

3.6 Modeling the Language

Statistical language models represent word sequences as probability distributions. The n-gram logistic regression model is the most basic approach for calculating the likelihood of a sequence of words. A 2-gram, or bigram, refers to a series of two words, like "Arabic Speech." Similarly, a 3-gram, or trigram, refers to a sequence of three words, such "Automatic Speech recognition" (Jurafsky and Martin, 2008). An n-gram refers to a sequence of n words. Equation (6) presents a formula for the n-gram..

$$p(w) = \prod_{k=1}^K p(W_k | w_{k-1}, w_{k-2}, \dots, w_{k-n+1}) \quad (6)$$

The equation has two primary hyper-parameters: K, representing the amount of words in W, and n, indicating the order of the language model (three for the trigram language model and two for the bigram language model). The calculation of LM should ideally be applicable to any sequence. Increased values of n in an n-gram can lead to sparsity. Hence, the occurrence of zero probability in speech recognition applications may be attributed to data sparsity, where the number of available values (n) often falls between the range of two to four. Maximum likelihood estimation, or MLE, is a precise and direct method for calculating probabilities. In order to obtain the maximum likelihood estimate (MLE) for the parameters of an n-gram model, we collect frequency data from a corpus and then normalize it to a range of 0 to 1. The bigram probability of a word w_k , given a prior word w_{k-1} , may be computed using the approach described in Equation (7):

$$P(w_K | w_{K-1}) = \frac{C(w_{K-1}w_K)}{C(w_{K-1})} \quad (7)$$

Regarding the MLE n-gram parameter estimate general case as in Eq. (8):

$$P(w_K | w_{K-1}, w_{K-2}, \dots, w_{K-n+1}) = \frac{C(w_{K-n+1}w_K)}{C(w_{K-n+1})} \quad (8)$$

Equation (12) computes the probability of an n-gram by dividing the observed frequency of a prefix by the recorded frequency of a certain sequence, similar to Equation (11). The term "relative frequency" is used to denote this ratio. Some word sequences in the training text may have extremely low or nonexistent frequency. In order to prevent the language model from assigning a probability of zero to unknown word sequences, it is necessary to distribute a portion of the probability mass from more common word sequences to these unseen word sequences. This phenomenon is commonly referred to as discounting or smoothing. To address this issue, there are generally two primary methods: (i) In the context of language modelling, back-off or interpolation refers to a technique introduced by (Katz, 1987). This technique involves assigning probabilities to unseen word tokens based on the chance that they belong to a lower-order n-gram. The model distributes the probability mass unevenly across these unseen word tokens; and, (ii) In the counting approach proposed by (Kneser and Ney, 1995), the smoothing technique involves assigning a certain amount of probability distribution to n-gram sequences that were not seen during the training process. In addition, researchers have constructed more advanced language models based on recurrent neural networks, as demonstrated by the work (Mikolov, n.d.) and (Lecorvé and Motlicek, 2012).

3.7 Decoding

During the process of decoding, the language and acoustic models are utilized to identify the recognition hypothesis that best aligns with the models. To determine the most probable

sequence of words, one must utilize the Viterbi method (Barker et al., 2005). Due to the impracticality of doing a comprehensive breadth search, the beam search method (Chiu et al., 2018) is commonly used to efficiently identify the word sequence by means of trimming.

The process for beam search is as outlined below. In the initial decoding step, we provide the decoder with a unique start of sentence character, $\langle \text{sos} \rangle$. Afterwards, the decoder generates a distribution, denoted as $P(y_1 | \langle \text{sos} \rangle, c_0)$, over the G graphemes that make up the output vocabulary. We construct a partial transcript for each of the top k most probable graphemes and include it in our beam. During the decoding stage that follows, we enlarge each of the k partial transcripts by each of the G graphemes, resulting in a combined total of kG possibilities. Next, we select only the k most likely partial transcripts from this set of extensions. Once the distinctive $\langle \text{sos} \rangle$ token marking the end of a sentence is detected, the transcript is considered finished and removed from the beam. This process will continue until all unfinished transcripts have been completed.

3.8 Evaluation Metrics

When assessing the effectiveness of a classification model, it is typical to use measures such as accuracy, precision, recall, and F1 score. These measures are particularly relevant for evaluating models that predict binary outcomes, such as differentiating between spam and non-spam, fraud and non-fraud, or positive and negative (Shewalkar et al., 2019).

Accuracy is a fundamental and often used measure that calculates the ratio of accurately predicted occurrences to the whole dataset. It functions as a metric of the model's ability to accurately classify occurrences into their assigned categories. The accuracy is calculated using the Equation 9:

$$\text{Accuracy} = \frac{\text{True Positives} + \text{true negatives}}{\text{True Positives} + \text{False Positives} + \text{true negatives} + \text{false negatives}} \quad (9)$$

Precision is a quantitative measure that evaluates the proportion of correctly predicted positive instances to all instances predicted as positive by the model. The term "signifies" indicates the level of accuracy of the model in predicting positive cases, and it is determined by applying a specific Equation 10:

$$\text{Precision} = \frac{\text{True Positives}}{\text{True Positives} + \text{False Positives}} \quad (10)$$

Recall, a different measure, quantifies the proportion of correctly identified positive cases compared to the total number of positive examples in the dataset. The metric measures the model's ability to accurately detect positive cases and is computed as Equation 11:

$$\text{Recall} = \frac{\text{True Positives}}{\text{True Positives} + \text{False neative}} \quad (11)$$

The **F1 score**, which is calculated as the harmonic mean of accuracy and recall, achieves a balanced evaluation between the two measures. Both accuracy and recall are equally important. The F1 score is calculated by Equation 12:

$$\text{F1 score} = \frac{\text{True Positives}}{\text{True Positives} + \text{False Positives}} \quad (12)$$

A confusion matrix is a tool used in machine learning and statistical classification to evaluate the performance of a classification algorithm. It offers a detailed overview of the system's performance. It is especially advantageous for addressing binary or multiclass classification issues. The confusion matrix provides a concise summary of the classification process, categorizing the anticipated and actual class labels into different groups.

The following are the essential elements of a confusion matrix:

Fundamental Concepts:

1. True Positives (TP): Instances in which the model accurately predicts the positive class.

2. True Negatives (TN): Instances where the model accurately predicts the absence of the positive class.

3. False Positives (FP) - Type I Error: Instances in which the model wrongly predicts the positive class.

4. False Negatives (FN) - Type II Error: Instances in which the model inaccurately forecasts the negative class.

Configuration of the Confusion Matrix: A confusion matrix is structured as a 2x2 matrix for binary classification and as a NxN matrix for multiclass classification, where N represents the total number of classes.

Binary Classification: Table 2 represents the binary classification for the confusion matrix.

Table 2

Binary Classification

	Predicted Positive	Predicted Negative
Actual Positive	True Positives	False Negatives
Actual Negative	False Positives	True Negatives

A confusion matrix is a potent instrument for assessing the effectiveness of a classification model by offering a comprehensive breakdown of the expected and actual class labels. It serves as the foundation for calculating numerous performance measures that help evaluate the model's success in various categorization areas.

Performance metrics obtained from the confusion matrix:

1. Accuracy

2. Precision
3. Recall
4. F1-Score

3.9 Learning Curve in Deep Learning Models

Within the realm of Deep Learning Models, a learning curve is a visual depiction that showcases the performance of the model, usually in relation to its training and validation datasets, during many epochs or iterations throughout the training procedure. The learning curve offers useful insights regarding the model's proficiency in acquiring and adjusting to new information over a period of time (Cao et al., 2016).

- **The essential elements of a learning curve comprise:**

1. Loss during training and loss during validation:

- The training loss refers to the amount of error seen on the training dataset at each epoch. Typically, it will gradually decrease as the model improves its ability to accurately represent the training data.

- Validation Loss: Represents the discrepancy on an independent validation dataset that the model has not been exposed to during the training process. It aids in evaluating the model's ability to apply learned knowledge to new data. Ideally, the value should drop. However, if it begins to increase, it might be a sign of overfitting.

2. Training Accuracy and Validation Accuracy: - Training Accuracy: Refers to the ratio of accurately categorized occurrences in the training dataset. The value should progressively rise over successive epochs.

- Validation Accuracy: A metric that is akin to training accuracy, but it is calculated using a separate validation dataset. It assesses the model's ability to generalize to unfamiliar data.

3. Overfitting and Underfitting: - Overfitting arises when the model gets excessively intricate, precisely fitting the training data and exhibiting poor performance on novel data.

- Underfitting occurs when the model is excessively simplistic, resulting in an inability to accurately capture the fundamental patterns within the data.

4. Convergence: - A learning curve aids in determining when a model has reached convergence, suggesting that more training may not substantially enhance performance.

Interpreting a learning curve entails evaluating the model's efficacy in learning, its ability to generalize to new data, and determining the need for alterations, such as modifying hyperparameters or model design.

A learning curve is an essential tool for monitoring and comprehending the dynamics of training Deep Learning Models. It allows practitioners to make educated decisions to improve the performance of the model.

3.10 Conclusion

We provide a concise overview of the numerous functionalities included in modern voice recognition systems. A comprehensive analysis of the literature pertaining to ASR architecture was presented. The language model encompassed the front-end functionality. The process of acoustic modelling involves the utilization of a neural network acoustic model to represent and analyze sound patterns. This model is then used to create a language model and perform decoding tasks. The evaluation measures were elucidated, including many aspects. We presented a comprehensive introduction of Arabic voice recognition.

Chapter 4: Dataset Corpus

Arabic social media platforms display a blend of two linguistic variations: (MSA) and (DA) (Sadat et al., 2014). Although there are certain similarities across these language variations, there are also notable differences that might impact (SA) applications.

Arabic speakers frequently utilize Dialect Arabic or "colloquial Arabic" in their everyday interactions, a widely prevalent custom observed in several countries and areas. This category encompasses Egyptian Arabic, Maghrebi Arabic, Gulf Arabic, and several additional dialects. Internet and social media users often employ Dialect Arabic, primarily in written format, which displays regional variances. This thesis focuses on the Egyptian dialect and modern standard Arabic (MSA) language.

The Arabic language has complex morphology and elaborate sentence patterns, as recognized by grammarians. Arabic has a complex morphology, which includes the use of templatic and affixational morphemes, detailed morphological rules, and a comprehensive system of parts.

Arabic employs a significant amount of inflections, mostly because of the incorporation of relational phrases and pronouns. The intricacy of Arabic morphology stems from over 10,000 roots that serve as the foundation for nouns and verbs. Arabic morphology encompasses 120 distinct patterns that highlight the importance of 5,000 fundamental roots. Arabic has a variable word order, enabling the strategic placement of words at the beginning of a sentence to emphasize them. During the parsing process, the syntactic analyzer scrutinizes the input tokens provided by the lexical analyzer and strives to determine the sentence structure according to the norms of Arabic grammar (Shaalán et al., 2019).

4.1 Data Collection

The process of collecting data for emotion speech recognition entails assembling a varied dataset that encompasses a wide spectrum of emotional expressions in spoken language (Alluhaidan et al., 2023). The process of data collecting generally involves the following steps:

1. *Data Source Selection*: Determine the suitable sources for gathering emotional speech data. Examples of potential sources for data collection might encompass openly accessible datasets, data obtained from controlled trials, or authentic recordings derived from platforms such as social media, interviews, or customer service calls.

2. *Moral considerations*: Ensure strict adherence to ethical norms and utmost regard for privacy during the data gathering process. Acquire consent from participants, and if relevant, anonymize or de-identify the gathered data to safeguard individuals' identities.

3. *Emotional Expression Diversity*: Strive to incorporate a broad range of emotional expressions into the dataset, encompassing emotions such as joy, sorrow, rage, astonishment, apprehension, and neutral states. This variety aids in the training of the model to accurately identify a wide range of emotions.

4. Ensure *speaker diversity* by incorporating a wide range of individuals with varying ages, genders, accents, and linguistic backgrounds. The inclusion of several demographic groups enhances the resilience and applicability of the emotion recognition model.

5. *Recording settings*: Take into account diverse recording settings, including distinct surroundings, ambient sounds, and recording equipment. This guarantees that the model is proficient in managing the variances experienced in real-life situations.

6. *Emotion Annotation*: Manually assign suitable emotion labels to the gathered voice data. Human annotators can analyze the recordings and assign labels based on the conveyed

emotions. The annotated data will be utilized for both training and assessing the emotion recognition model.

7. *Quality Assurance*: Conduct quality assessments on the gathered data to verify its precision and pertinence. Eliminate any extraneous or subpar recordings that might have a negative impact on the model's training.

8. *Dataset Partitioning*: Segregate the acquired dataset into distinct subsets for training, validation, and testing purposes. The training set is utilized for model training, the validation set aids in refining hyperparameters and mitigating overfitting, and the test set evaluates the model's performance on unfamiliar data.

9. *Documentation*: Maintain comprehensive records on the process of gathering data, encompassing details about the individuals involved, the conditions under which the data was recorded, and any unique factors that might potentially affect the use of the dataset.

10. *Ongoing Updates*: If relevant, it is advisable to constantly update the dataset by adding fresh recordings. This will ensure that the model stays effective and can adjust to changing patterns of emotional expression.

By adhering to these procedures, the gathered dataset transforms into a significant asset for training and assessing emotion speech recognition models.

4.1.1 YouTube as a Source of Emotion-Rich Data

YouTube, being a massive archive of Egyptian multimedia material, offers an unrivalled resource for gathering varied and authentic data for emotion recognition. By utilizing Egyptian Arabic-language YouTube videos, a wide range of emotional responses may be documented, showcasing authentic situations and cultural subtleties that are unique to Arabic speakers.

Gender Diversity in Speech: Acknowledging the importance of gender diversity in the presentation of emotions, the dataset includes speakers of both male and female genders. This inclusion recognizes the inherent diversity in the expression of emotions across genders, guaranteeing that the emotion detection system is attuned to different speech patterns and intonations.

Audio Features: The dataset focuses mostly on audio features that have been taken from the YouTube videos. The variables mentioned include speech patterns, intonations, and other auditory characteristics that are crucial for recognizing emotions (Liu et al., 1997). Examining audio data enables a comprehensive comprehension of emotional manifestation, catching subtle intricacies that may exist in speech but are not perceptible through visual signals.

- **Emotion Categories:**

The dataset includes seven main emotion categories, providing a thorough depiction of the whole range of emotions. The range of emotions may encompass happiness, sorrow, rage, love, fear, disgust, and natural responses (Schröder et al., 2007). The model is taught to classify emotions into several categories, which enhances its ability to identify a broad variety of emotional states. This makes the model more resilient and adaptable for use in practical situations.

Challenges and Considerations

Gathering data for emotion identification, particularly from public platforms such as YouTube, presents a range of difficulties. The issues encompass guaranteeing the privacy and permission of persons showcased in the films, rectifying biases in the dataset, and handling the probable fluctuations in recording settings. Thorough deliberation is devoted to ethical norms in order to ethically manage these problems.

To summarize, the method of collecting data for Arabic emotion identification from YouTube videos is a careful and thorough effort with the goal of building a varied and inclusive dataset. The dataset establishes the groundwork for creating a strong and culturally aware emotion detection system designed specifically for Arabic speakers. It achieves this by encompassing a wide range of emotions, including those expressed by both male and female speakers, and by emphasizing acoustic characteristics. This undertaking not only adds to progress in affective computing but also demonstrates a dedication to inclusiveness and precision in comprehending the emotional terrain of Arabic speakers.

4.2 Data Annotation

In the fields of artificial intelligence and machine learning, data annotation is an essential procedure. Below are the main topics in the data annotation:

- **Emotion Categorization Specialist:**

The effectiveness of emotion recognition algorithms greatly depends on the proficiency that tasked with annotating and classifying emotions in the dataset. For our situation, we want the expertise of a recognized professional in emotion classification with a background in psychology, linguistics, or a comparable discipline, as stated by (Seirafi et al., 2013). This expert has a sophisticated comprehension of cross-cultural emotional displays, guaranteeing precise and culturally aware annotations for Arabic speakers within the Education Department at British University in Dubai (BUID).

- **Procedure for Emotional Category Annotation:**

1. Preparation and Training: Prior to beginning the annotation process, the emotion classification specialist undertakes comprehensive training (Alluhaidan et al., 2023). This encompasses being acquainted with the chosen emotion categories, instructions for annotating,

and exposure to a diverse range of emotional expressions in Arabic speech. The training materials may consist of audiovisual examples and books that focus on emotional expression patterns.

2. Emotion categories are specified with clarity and precision to provide guidance for the annotation process. The selection of these categories is based on well-established psychological frameworks, which include a wide variety of emotions including happiness, romance, melancholy, rage, fear, and contempt. The professional acquaints themselves with these categories in order to guarantee uniformity and precision in annotating.

3. Annotation Guidelines: The specialist is given comprehensive guidelines that specify the requirements for each emotion type. These rules may encompass verbal clues, tone of voice, pitch fluctuations, and other acoustic characteristics that are pertinent to the portrayal of emotions. The objective is to provide a uniform methodology for annotation, reducing the influence of personal opinions and guaranteeing consensus among annotators.

4. Multi-Modal Analysis: In addition to auditory clues, the process of annotation may encompass multi-modal analysis, taking into account visual signals derived from video recordings. Facial expressions, body language, and environmental information together contribute to a comprehensive comprehension of emotional states. The specialist is skilled in utilizing both aural and visual signals to achieve precise classification.

5. The annotation process follows an iterative approach, involving ongoing feedback exchanges between the specialized and supervisory teams. Periodic meetings and discussions are conducted to resolve uncertainties, elucidate instructions, and enhance the annotation procedure. By employing an iterative technique, the emotion classification process is guaranteed to achieve a commendable degree of precision and dependability.

6. **Quality Control and Inter-Annotator Agreement:** In order to ensure the accuracy of annotations, a portion of the dataset is reserved for testing inter-annotator agreement. Several experts may individually classify emotions within this specific group, and their annotations are then compared. Consistency tests and discussions are performed to address inconsistencies and improve the dependability of the annotated dataset.

7. **Documentation and Reporting:**

Detailed documentation is consistently maintained during the annotation process. This encompasses documentation of annotating criteria, any alterations implemented throughout the procedure, and observations regarding encountered difficulties. A comprehensive report is created, which outlines the methods and results of the annotation process. This report offers transparency and valuable insights into the emotional classification of the dataset.

By meticulously following these processes and enlisting the expertise of a certified professional in emotion classification, the dataset attains a high level of accuracy and cultural sensitivity in identifying and classifying emotions in Arabic speech. The meticulous procedure of annotation significantly contributes to the advancement of resilient emotion detection models specifically designed for the complexities of the Arabic-speaking community.

4.2.1 Annotation Process for Egyptian Dialect Speech Recognition

The annotation process for the Egyptian dialect speech recognition task involved several key stages: annotator selection, data segmentation, emotional category assignment, and inter-annotator agreement assessment. The aim was to create a high-quality dataset that accurately represents emotional expressions in Egyptian Arabic speech.

Annotator Selection and Description

The selection of annotators was critical to ensuring the accuracy and cultural relevance of the dataset. The annotators were chosen based on the following criteria:

1. Native Egyptian Arabic Speakers: All annotators were native speakers of the Egyptian dialect, ensuring a deep understanding of local linguistic nuances and emotional expressions.
2. Linguistic and Phonetic Expertise: Annotators had backgrounds in linguistics and phonetics, with experience in dialectal variations of Arabic. This expertise was necessary for correctly identifying and categorizing emotions.
3. Experience in Speech Annotation: Annotators were required to have prior experience in speech or language annotation to ensure familiarity with the annotation tools and methodologies used in the project.

Annotation Procedure

The process was divided into the following steps:

1. Data Segmentation: The audio files were segmented into utterances, each representing a distinct emotional expression. Segments were carefully reviewed to avoid overlaps and ensure clarity.
2. Emotion Categorization: Each segment was labelled with one of the predefined emotional categories: happiness, sadness, anger, disgust, fear, romance, and neutrality. Annotators relied on both prosodic and phonetic features, such as intonation and speech rate, to identify emotions. Table 3 illustrates the emotion categories.

Table 3*The Emotions Categories*

Emotion Type	Example in Arabic
Happiness	النهاردة انا سعيدة قوي اني شفتكم
Romantic	انا بحبك وما اقدرش أعيش من غيرك
No Emotions	الخلطة دي ممتازة للشعر
Anger	انت بنادم مابتفهمش
Sadness	انا قلبي بيتقطع عليه
Fear	انا خيفة قوي
Disgust	ايه القرف دا

3. Consensus Building: Regular meetings were held between annotators to discuss difficult or ambiguous cases. This ensured a consistent understanding and application of emotional labels across the dataset.

Inter-Annotator Agreement

To validate the reliability of the annotation process, the Inter-Annotator Agreement (IAA) was calculated using Cohen's Kappa Statistic. Cohen's Kappa is a robust statistical measure that accounts for agreement occurring by chance (Schuller et al., 2009). It is useful in speech emotion recognition, where subtle differences in emotional expressions may lead to varying interpretations.

Cohen's Kappa Formula: The statistic is defined as:

$$k = \frac{P_0 - P_e}{1 - P_e}$$

where:

- P_0 is the observed agreement among annotators.
- P_e is the expected agreement by chance.
- Interpretation of Kappa Values:
 - < 0.20 : Slight agreement.
 - $0.21\text{--}0.40$: Fair agreement.
 - $0.41\text{--}0.60$: Moderate agreement.
 - $0.61\text{--}0.80$: Substantial agreement.
 - $0.81\text{--}1.00$: Almost perfect agreement.
- Result: The inter-annotator agreement for the dataset was calculated, yielding a Cohen's Kappa value of 0.75, indicating substantial agreement. This high level of agreement validates the consistency and reliability of the annotated labels (El Ayadi et al., 2011).

The careful selection of annotators and a structured annotation process, supported by a strong inter-annotator agreement, ensures that the resulting Egyptian Arabic speech emotion dataset is both culturally authentic and reliable for training Deep Learning Models.

These references provide foundational support for the methods used in the annotation process, Cohen's Kappa statistics, and existing frameworks for emotion recognition in speech processing.

4.2.2 Annotation

Annotation is of utmost significance in machine learning and Deep Learning Models, since they largely depend on labelled data to identify patterns and make predictions or classifications on fresh, unseen data. Without precise and thoroughly annotated training data, models may not be able to reach their highest level of performance, resulting in incorrect outcomes.

Annotation is crucial in the domain of emotion voice recognition, as it serves as a vital step in training machine learning and Deep Learning Models. Sentiment analysis with emotion identification models relies primarily on labelled data to identify patterns and connections in voice signals that correlate to various emotional states. Annotated datasets provide as a basis for these models to acquire and apply their comprehension of different emotional signals.

Annotators apply labels, such as happiness, sadness, disgust, fear, romantic, anger and neutral, to segments of voice data, therefore generating a reliable reference for training the models. The precision and depth of these annotations have a direct influence on the model's capacity to precisely detect and categorize emotions in speech. Thoroughly annotated datasets improve the model's performance, guaranteeing strong and dependable emotion identification skills when confronted with unfamiliar voice samples.

This study employed emotion-type abbreviations to measure the specific kind of emotion. The abbreviations were later transformed manually into relevant emotion categories, including happy, sadness, fear, disgust, anger, romance, and neutral. The process of transformation is indicated in Table 4. The collection consists of labelled speech occurrences, with each instance allocated based on certain emotional signals.

Table 4

Emotions with their Labels

Emotion Type	Labels	Numerical Representation
Anger	AN	0
Disgust	DI	1
Fear	FR	2
Happiness	HA	3

Neutral	NO	4
Romantic	RO	5
Sadness	SA	6

The first step after collecting the dataset entailed identifying and understanding the audio recordings. The naming structure of each dataset was distinct, and the emotion label was derived from the file names for later categorization. Table 5 provides an illustrative demonstration of how to label the filenames. The study of the datasets involved the employment of wave plots, which displayed randomly picked audio files. The purpose of these plots was to visually depict the characteristics of the data being examined. In addition, the speech data's quality was evaluated by detecting any possible background noise in the audio recordings and assessing the level of clarity with which people could comprehend the transmitted emotions.

Table 5

Labeling the Filenames

Filenames
DI طرحة_DI.wav
NO بشرة_NO.wav
FR خوف_FR.wav
AN غضب_AN.wav
SA مبلغ_HA.wav
HA مبروك_HA.wav
RO حب_RO.wav

The annotation in this thesis pertains to the act of assigning a specific emotion category to each audio segment by labelling or tagging. The audio files were named using the acronyms of the emotion types together with the file's ID name. The MFCC_Data_PreProcess function derives the labels from the filenames by extracting the first two alphabet characters that represent the emotion type. The filenames are extracted and utilized as the label. The labels correspond to the emotional content of the audio segments and are added to the "labels" list in the data dictionary.

4.3 Arabic Dataset Corpus

Sentiment analysis entails classifying the emotional disposition conveyed in audio, encompassing many forms of speech such as conversations, songs, or other verbal expressions. The emotions in sentiment analysis are typically categorized into several groups, as outlined by (Duwairi, 2015). These groupings are widely used in research on Speech Emotion Recognition (SER) to analyze emotions. In this thesis, Figure 24 represents the frequent sentiment categories and the corresponding emotions linked with them that have been employed.

1. Positive Sentiment Emotions: happiness and romantic.

- Example of happiness: "انا سعيدة قوي النهاردة اني شفtekم"
- Example of romantic: "انا بحبك وما اقدرش أعيش من غيرك"

2. Neutral Sentiment Emotions: Indifference, objectivity, lack of strong emotion.

- Example: "الخلطة دي ممتازة للشعر"

3. Negative Sentiment Emotions: Anger, sadness, fear and disgust.

- Example of anger: "انت بنادم مابتفهمش"
- Example of sadness: "انا قلبي بيتقطع عليه"
- Example of fear: "انا خيفة قوي"

- Example of disgust: "ايه القرف دا"

Figure 23

Sentiment Analysis with Distributed Emotions

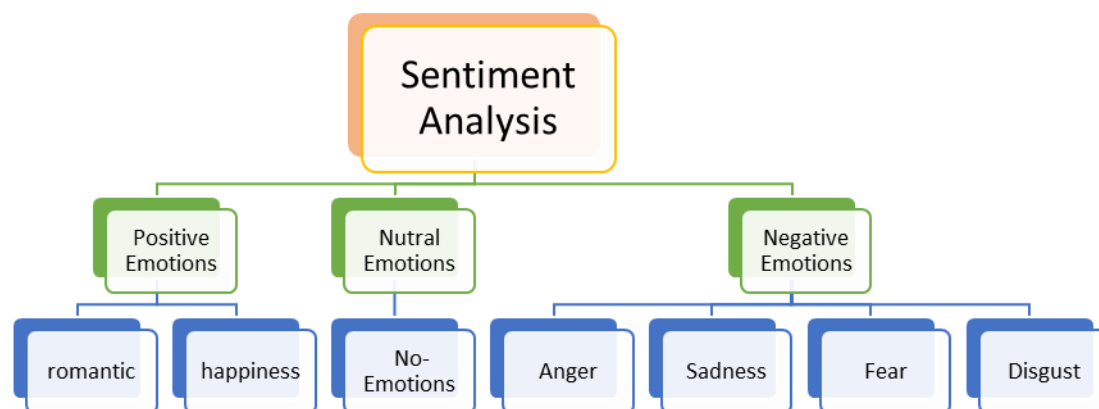


Table 6 displays the sentiment categories and the corresponding emotions linked to them (Abdul-Mageed et al., 2014).

This thesis gathered a dataset of emotional speech in the Egyptian dialect from YouTube videos. These videos encompass a range of programs, such as comedy shows, snippets, and short films, with a main objective of highlighting Egyptian culture and material that is important to the local audience.

The evaluation group analyzed the movies by carefully observing and evaluating them to find certain sequences that displayed clear emotional indicators. The objective was to evaluate the effectiveness of different machine learning algorithms in detecting emotions. Afterwards, the movies were processed to extract audio recordings, which were then divided into smaller segments. Each segment had a duration ranging from 1 to 50 seconds and was recorded in the mp3 format. The segmentation was determined by the inherent pauses in speech, guaranteeing that each segment included a whole phrase.

Sentiment analysis methods strive to categorize text into these specific classifications, so offering valuable insights into the emotional context of the material. These categories facilitate the comprehension of public sentiment, consumer input, and patterns in diverse textual data sources for corporations, scholars, and organizations.

Table 6

Sentiment Analysis with Related Emotions

Sentiment Analysis	Emotion Type	Example in Arabic	Translate to English
Positive Sentiment	Happiness	انا سعيدة قوي اني شفتكم النهاردة	I am glade meeting you today
	Romantic	انا بحبك وما اقدرش أعيش من غيرك	I love you and can't live without you
Neutral Sentiment	No Emotions	الخلطة دي ممتازة للشعر	This mixture is excellent for hair
Negative Sentiment	Anger	انت بنادم مابتفهمش	You regret that you don't understand
	Sadness	انا قلبي بيتقطع عليه	My heart breaks for him
	Fear	انا خيفة قوي	I am very afraid
	Disgust	ايه القرف دا	What a disgusting thing

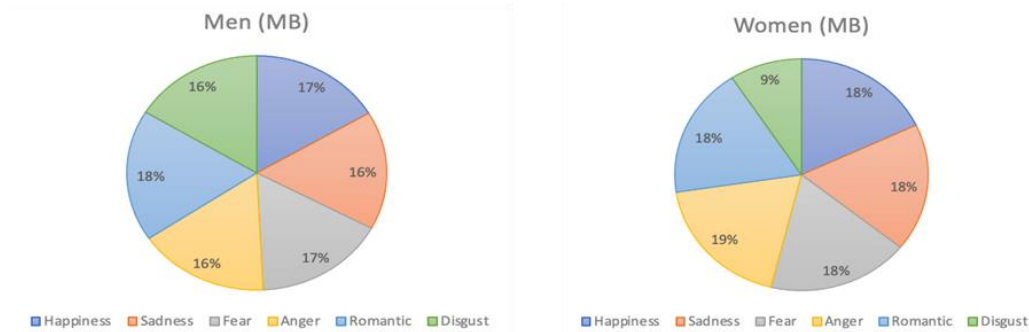
The audio data is sourced from YouTube channels, during a period of 11 years from March 2012 to September 2023 with a corpus of more than 350,000 words. The dataset comprises around 400 Arabic utterances with an Egyptian dialect. These statements are articulated by more than 200 male Arabic speakers and around 200 female speakers, ranging in age from 27 to 70 years, totaling over 40 hours. The dataset categorizes the collected emotions as anger, contempt, fear, romance, melancholy, and happy. Each category consists of around 1 gigabyte of audio files for both male and female speakers, resulting in a dataset size of approximately 15 gigabytes. Following the preprocessing phase and elimination of extraneous elements such as music, quiet, and background noise, table 7 presents a comprehensive analysis of emotions, including their respective magnitudes, for both males and females. Therefore, these emotions were employed as descriptors for the audio segments. As part of the data preparation stage before to classification, measures were implemented to remove any unwanted interference such as noise, background music, and intervals of stillness.

Table 7

Distribution of Emotions Among Male and Female Speakers

Emotion	Women (MB)	Men (MB)
Happiness	504.1	511.3
Sadness	499.5	506.9
Fear	508.8	507.7
Anger	524.5	506.5
Romantic	510	562
Disgust	259	505.3

Figures 24 displays the proportionate distribution of emotions within the corresponding categories for both male and female speech participants.

Figure 24*Emotion Distribution for Men and Women Speakers*

4.4 Dataset Splitting

In the field of Arabic voice recognition, the data is commonly categorized into three primary sets: training data, validation data, and testing data.

1. Training Data

Definition: Training data refers to a specific subset of the dataset that is utilized to train the Automatic Speech Recognition (ASR) model. The dataset has a substantial amount of audio samples that have been labelled, with each sample being matched with its matching transcription.

Purpose: The main objective of the training set is to familiarize the ASR model with a wide array of Arabic speech patterns and linguistic differences. During this phase, the model acquires the ability to associate audio characteristics with their corresponding text representations.

2. Validation Data

Definition: Validation data refers to a smaller portion of the dataset that is specifically reserved for the purpose of refining the ASR model throughout the training process. It is employed to evaluate the model's performance on unseen data and to tune hyperparameters. The user did not provide any text.

Objective: The purpose of the validation set is to mitigate overfitting by offering a distinct set of data that the model has not been explicitly trained on. It allows for the modification of model parameters to enhance generalization.

3. **Testing Data**

Definition: Testing data refers to a distinct subset of the dataset that the model has not been exposed to during its training or validation process. This subset is fully independent from the data used for training and validation. It is employed to assess the ultimate performance and capacity to generalize of the trained ASR model.

Purpose: The testing set impartially evaluates the model's efficacy in identifying speech patterns in unfamiliar and unobserved data. It assists in assessing the anticipated performance of the model in real-world situations.

- **Data Splitting Strategies**

It is customary to assign a larger proportion of the data to the training set (e.g., 70-80%), a smaller percentage to the validation set (e.g., 10-15%), and the remaining data to the testing set.

The user did not provide any text. Randomization guarantees that the data in each set accurately reflects the general distribution of speech patterns and linguistic variation.

Through meticulous management of these datasets, it is possible to train, validate, and evaluate a highly effective and resilient Arabic voice recognition model for real-world applications.

4.5 Conclusion

We provided a concise overview of the data corpus, delineating the data sources employed in the voice recognition algorithms. This section provides an in-depth analysis of data

annotation, with a specific emphasis on the Arabi dataset corpus. Moreover, this part offers a thorough elucidation of data splitting

Chapter 5: Methodology

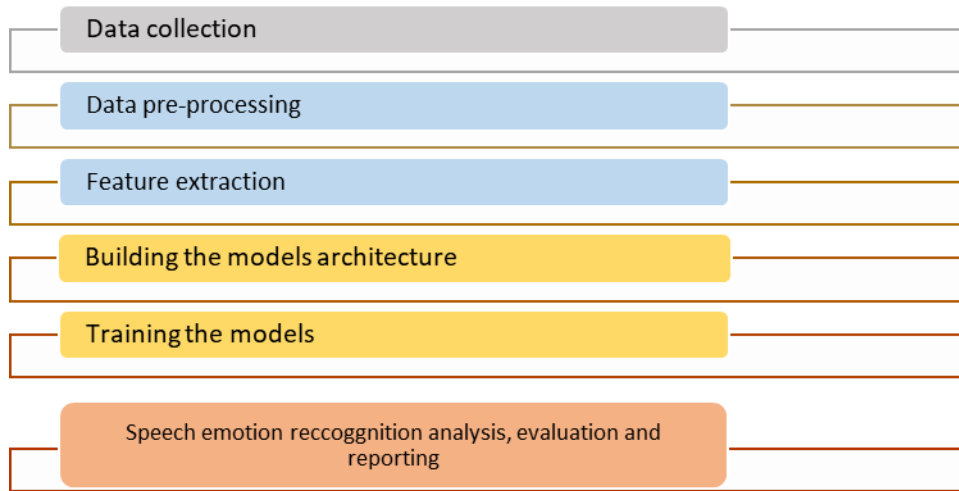
Deep learning models are being used to analyze the sentiment of dialect speech, which involves understanding the emotional content of spoken language. This area of artificial intelligence is important because it improves the way machines interact with humans by allowing them to recognize and respond to emotional cues in speech. When dealing with Arabic speech, the specific phonetic traits, many dialects, and cultural subtleties provide distinctive difficulties, requiring a particular method to accurately identify emotions.

The proposed methodology only utilizes a multilayer neural network for the purpose of Arabic Speech Emotion Recognition. The core aspect of the SER system is around the identification of emotions in speech, using the voice input given to the system. Afterwards, the multilayer neural network independently performs the process of extracting and selecting features.

This approach outlines the methodical process of creating a reliable sentiment analysis system for Arabic speakers, specifically focusing on dialect speech. Every stage, from data collection to model training and deployment, is carefully designed to specifically handle the inherent linguistic diversity and emotional expressions found in the Arabic language. The schematic in figure 25 illustrates the fundamental elements of the framework.

1. Data Collection

An inclusive and comprehensive collection of Arabic voice recordings is compiled, covering a wide range of emotions and include speakers from various areas and backgrounds. Refer to section 5 for further information.

Figure 25*Framework of the Speech Sentiment Recognition*

2. Data Preprocessing involves preparing and cleaning the data before it can be used for analysis or modelling.

The initial step in processing raw audio files involves several preprocessing approaches, such as converting the format, using data augmentation methods, and transforming them into Mel spectrograms or Mel-frequency cepstral coefficients (MFCC). These transformations are performed to extract significant characteristics that are useful for deep learning. Refer to section 4 for further information.

3. Architecture of the Model

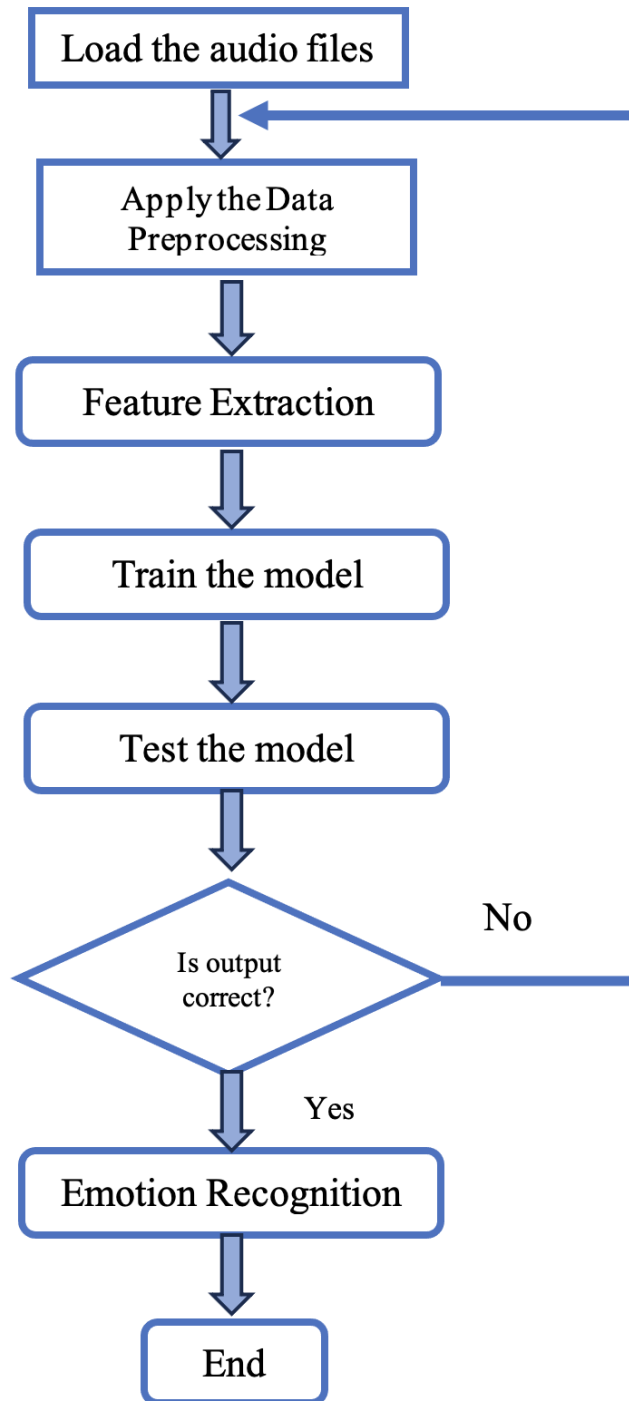
An appropriate deep learning architecture is selected, taking into consideration the time-based patterns of voice inputs. Popular models such as Convolutional Neural Networks (CNNs), bidirectional LSTM (Bi-LSTM), Gated Recurrent Unit (GRU), and hybrid architectures like CNN-BI-LSTM, GRU-BI-LSTM, and CNN-GRU are taken into account. Refer to section 7 for further information.

4. Training: The dataset is divided into training, validation, and testing subsets. The model undergoes training using suitable loss functions and optimization techniques, while hyperparameters are adjusted to achieve optimal performance.

5. Evaluation: The model's performance is thoroughly evaluated on the testing set using measures like as accuracy, precision, recall, and F1 score. An in-depth examination, using confusion matrices, provides valuable insights into the model's capabilities and limitations.

6. Fine-Tuning : Iterative optimization based on evaluation findings may entail tuning hyperparameters or altering the model architecture to improve accuracy and generalization.

7. Performance Analysis: Consistent surveillance and evaluation of the model's performance across various circumstances enable continuous refinement and adjustment based on user interactions. Refer to section 4 for further information.6. This approach aims to create a detailed structure that recognizes the complexities of Arabic speech. Its goal is to enable the creation of emotionally intelligent systems that can comprehend and respond to the diverse range of human emotions expressed via spoken Arabic. Figure 26 displays the flowchart outlining the process.

Figure 26*The Flowchart of the Methodology*

Below is a basic pseudocode outline for speech sentiment emotion recognition:

1. *Load the dataset*
2. *Preprocess the audio data:*
 - a. *Convert audio to a common sample rate (16KHz)*
 - b. *Split audio into segments*
 - c. *Extract relevant features (MFCCs)*
 - d. *Normalize or standardize the feature values*
3. *Label the emotional categories in the dataset (SA, HA, FR, DS, RO, AN, NON)*
4. *Split the dataset into training, validation and testing sets*
5. *Build the deep learning model:*
 - a. *Define the model architecture (Deep Learning Models)*
 - b. *Compile the model with an appropriate loss function and optimizer*
6. *Train the model on the training dataset:*
 - a. *Feed the preprocessed audio data to the model*
 - b. *Adjust model parameters to minimize the loss*
7. *Evaluate the model on the testing dataset:*
 - a. *Measure WER, accuracy, precision, recall, and F1 score*
 - b. *Visualize the performance metrics*

5.1 Data Collection

The dataset obtained from YouTube consists of around 400 spoken utterances in Arabic, specifically in an Egyptian dialect. This dataset includes contributions from a total of around 400 native Arabic speakers, comprising around 200 males and 200 females. The age range of the participants spans from 27 to 70 years. The dataset consists of emotions classified as anger,

disgust, fear, romance, melancholy, and happiness, with a total duration of around 30 hours. Each emotional category is linked to audio files of around 1 gigabyte for both male and female speakers, resulting in a total dataset size of roughly 15 gigabytes. The detected emotions were used as labels for the audio segments. Prior to categorization, a preprocessing stage was conducted to eliminate any unwanted elements such as noise, background music, and intervals of stillness from the data.

Data gathering is a crucial stage in the advancement of sentiment analysis with emotion detection algorithms, especially when targeting Egyptian Arabic speakers. The dataset in question is obtained from YouTube videos, providing a wide range of emotional expressions that are varied and abundant. This review explores the complexities of the data gathering process, highlighting the wide range of emotions involved, and recognizing the subtleties that arise when considering speakers of both genders. To obtain further information on the process of gathering data, please refer to chapter 3.

5.2 Data Preprocessing Methodology

Preprocessing is the first step in the development of an Automatic Speech Recognition (ASR) system, occurring before to later stages in the speech recognition process. The main purpose of this is to differentiate between signals that are produced with vocal cord vibrations and those that are not, and to create vectors that extract relevant characteristics. The voice signal is modified by preprocessing to improve its appropriateness for future feature extraction analysis.

5.2.1 Pre-Processing

Data preparation is an essential stage in Arabic Speech Recognition (ASR) that aims to improve the quality of incoming audio data and simplify the process of extracting useful features. The procedure entails many stages to guarantee that the speech signals are suitably prepared for

further analysis by ASR models. Consult Chapter 4 for more elucidation. The following are the applied steps of data preparation for Arabic voice recognition:

1. *Audio Loading and Conversion*: The first stage entails importing audio files containing spoken utterances in the .wav format, which is a widely used format for speech data. The audio files are parsed, and the waveform is transformed into a 2D NumPy array, which represents the amplitude or intensity of the sound over time. The value is affected by the sample frequency. For example, when using a sample rate of 44.1 kHz, the 2D Numpy array will consist of a singular row containing 44,100 numerical values, representing 1 second of audio. Furthermore, audio files might include either a single or dual channels.

2. *Data Augmentation*: Data augmentation techniques are employed to enhance the generalization and diversity of the dataset. This may involve adjusting the audio's spatial positioning to the left or right or modifying the pitch by a few degree. Augmentation facilitates the ASR model in acquiring knowledge from a wider spectrum of input variations.

3. *Mel Spectrograms*: The audio input undergoes a transformation process to generate Mel spectrograms. Spectrograms visually portray the audio signal by breaking it down into a collection of frequencies across time. Mel spectrograms specifically capture the frequency elements of the speech signal, which is crucial for extracting features for Automatic Speech Recognition (ASR).

4. *MFCC (Mel-Frequency Cepstral Coefficients)*: The voice signal is transformed into Mel-Frequency Cepstral Coefficients (MFCC) in addition to spectrograms. MFCCs are a condensed and crucial collection of coefficients obtained from the audio spectrum. MFCCs are very efficient at capturing the fundamental characteristics of human speech.

5. *Spectrogram Data Augmentation*: Implementing data augmentation techniques exclusively on spectrograms enhances the dataset even more. Methods such as SpecAugment, which utilize temporal and frequency masking, are frequently employed to enhance the resilience of models.

6. *Normalization and Standardization*: Normalization procedures are used to verify that the input characteristics have uniform scales. This is essential for the model to reach convergence throughout the training process. Standardization can be conducted to normalize the data by adjusting it to have a mean of zero and a variance of one.

7. *Categorization*: Every audio segment is categorized with its appropriate emotion to aid in supervised learning.

8. *Noise Reduction*: Preprocessing may entail the elimination of background noise, extraneous noises, or periods of quiet from the audio data. Noise reduction approaches optimize the signal-to-noise ratio, hence enhancing the model's capacity to accurately identify speech.

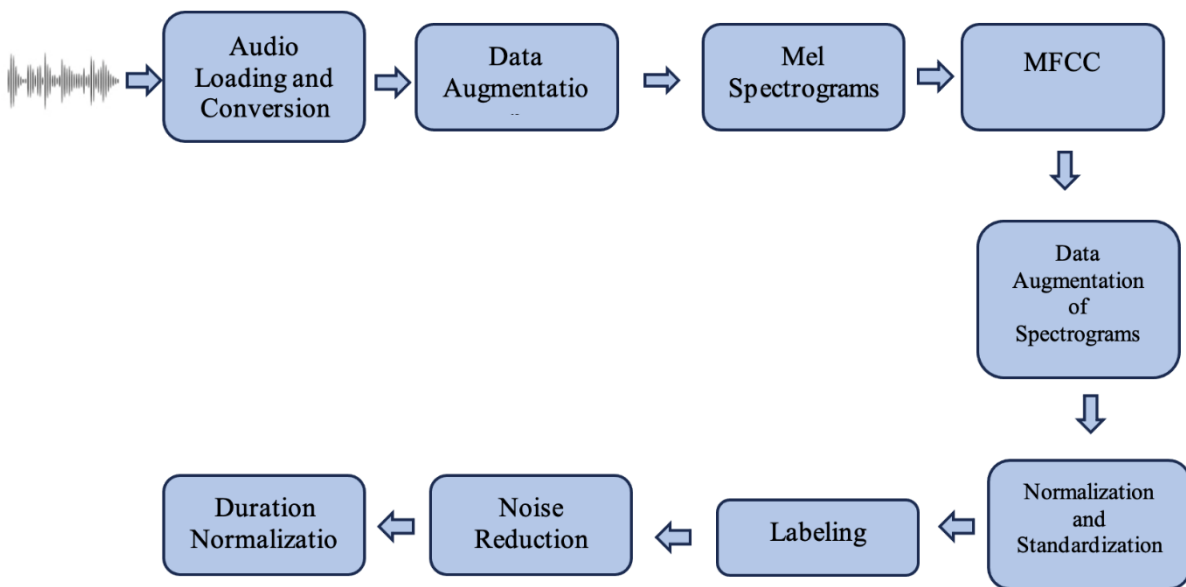
9. *Duration Normalization*: To enable successful learning from the data, it is necessary to normalize the duration of audio files, which may have various lengths, to a constant length. A typical strategy involves choosing an average time for all audio recordings, and preparing them for modelling. By using these preprocessing procedures, the input data for Arabic Speech Recognition is improved, enabling ASR models to efficiently acquire and identify speech patterns in Arabic language utterances. Figure 27 provides an overview of the Data Pre-Processing steps involved in Arabic Speech Recognition.

The preprocessing stage is the first step in the construction of an Automatic Speech Recognition (ASR) system, preceding further phases in speech recognition. The primary function of this is to distinguish between sounds produced with vocal cord vibration and those without,

eventually generating vectors that represent certain sound characteristics. The voice signal is modified during preprocessing to enhance its suitability for future feature extraction analysis. This chapter provides a comprehensive explanation of the procedures carried out during data preparation.

Figure 27

Data Pre-Processing for Arabic Speech Recognition



Pre-processing is highly important in both Natural Language Processing (NLP) and Sentiment Analysis (SA), as it plays a critical role in improving the performance of sentiment analysis models and refining the quality of obtained datasets. Typical pre-processing methods encompass text cleansing, standardization, tokenization, removal of stop words, and stemming. The precise procedures in the pre-processing phase may differ depending on the study's goals and the language being examined.

Regarding Arabic sentiment classification, specific research has investigated specialized pre-processing measures to assess their influence on effectiveness. The examination has focused on techniques such as the inclusion of emoticons and the use of individual stop lists. The findings of these research indicate that the implementation of customized stop lists and the substitution of emoticons with substantive words can improve the classification of Arabic sentiment. This underscores the need of customized pre-processing techniques in maximizing the results of sentiment analysis for the Arabic language. The data was stored as an mp3 file and subsequently converted to WAV files.

5.2.2 Data Splitting

Within the domain of Arabic voice recognition, the data is often divided into three main categories: training data, validation data, and testing data.

Training data and testing data are essential elements in the field of machine learning and data analysis. They are crucial for creating, evaluating, and verifying prediction models. As part of this inquiry, an essential first step was to train the model and transform datasets into a format that can be easily understood by machines, prior to undertaking sentiment analysis. Dataset annotation refers to the act of adding meaningful labels or explanations to a collection of speech, with the purpose of optimizing it for natural language processing (NLP) applications.

The annotation employed a comprehensive categorization method, consisting of seven separate categories to accurately define different emotions in Arabic speech: happy, sad, fear, disgust, anger, romance, and neutral. The inclusion of many classes in the annotation provided a more comprehensive understanding of the subtle aspects of speech emotion identification within the dataset.

The Arabic corpus collected for analysis was carefully divided into training, validation, and testing datasets, following a ratio of 60:10:30. The objective of this data partitioning strategy was to train the model using a wide range of examples, in order to evaluate its ability to properly generalize to new and unknown data. Figures 28 to 30 depict the equitable distribution of emotion categories throughout the training data, validation data, and testing data. The data is uniformly distributed across these sets, guaranteeing a representative sample for each emotion category.

To facilitate the training of the model with the data, it is necessary to synchronize the audio files and convert them into CSV (comma-separated values) format for input into the model. The CSV file consists of rows, each containing three values. The first value in each row is the file path of the audio filename. The second value represents the emotion type using abbreviations. The third value represents the numerical representation of the emotions. The procedure has been finalized, yielding a grand total of 2,875 records. Figure 31 presents a distribution of the dataset, showcasing the labelling data table.

Figure 28

Distribution of Emotions in Training Data

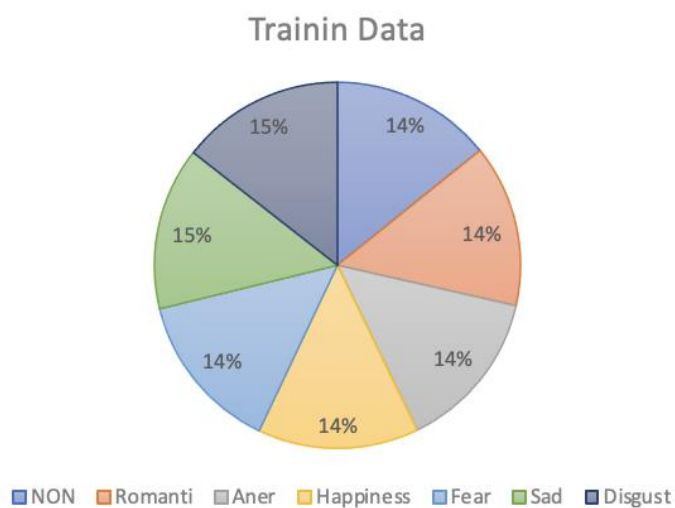


Figure 29
Distribution of Emotions in Validation Data

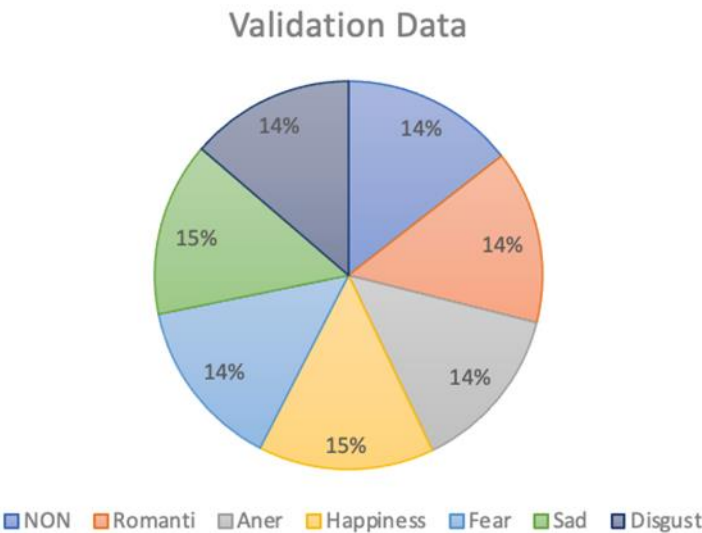


Figure 30
Distribution of Emotions in Validation Data

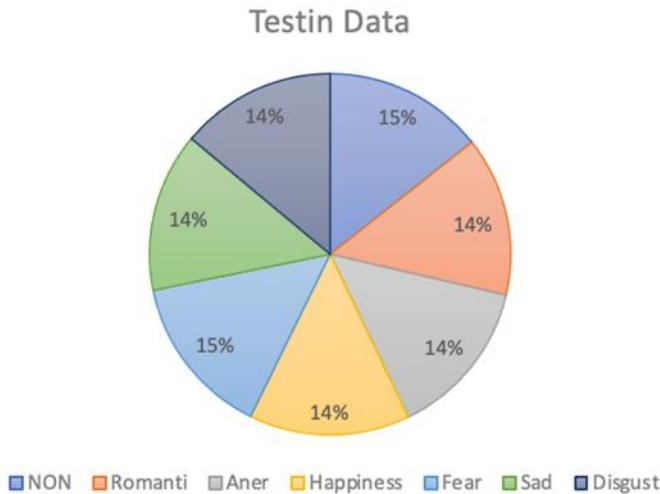


Figure 31*The Labeling Table*

	Speech	labels	labels_encoded
0	/content/drive/MyDrive/MEN/Speech/NO٢ سنت_NON.WAV	non	4
1	/content/drive/MyDrive/MEN/Speech/FR٢ عنبر_FRI...	fri	2
2	/content/drive/MyDrive/MEN/Speech/FR١ عمل_FRI.wav	fri	2
3	/content/drive/MyDrive/MEN/Speech/FR١ زين_FRI.wav	fri	2
4	/content/drive/MyDrive/MEN/Speech/FR١ جن_FRI.wav	fri	2

5.2.3 Feature Extraction

The voice signal comprises a multitude of elements that communicate emotional traits. MFCC, short for Mel-Frequency Cepstral Coefficients, is a method used to extract important features for speech analysis by collecting vital information from speech data segments. The Mel-frequency cepstral coefficients (MFCCs) of an audio signal typically comprise a succinct collection of characteristics, usually spanning from 10 to 20, that accurately depict the spectral envelope pattern. The calculation of MFCC characteristics entails the utilization of linearly spaced frequency filters for lower frequencies and logarithmically spaced frequency filters for higher frequencies. This design emulates the low-frequency sensitivity of the human ear (Mishra and Sharma, 2020). The following pictures depict the computation of the MFCC matrix for each audio sample. They showcase the raw signals of the emotions prior to preprocessing, along with the visualization of MFCC features using spectrograms for each emotion. Figure 32 (a and b) displays the signals for the raw data of Happiness and the spectrogram wave, respectively. Figure 33 displays two signals: Figure a represents the raw data signal, whereas Figure b represents the spectrogram wave signal for Sadness. The remaining emotions are represented by raw data and

spectrogram waveforms for happiness, respectively. Figures 34 to 38 depict the remaining emotions.

Figure 32

The Waveform of Happiness Emotion

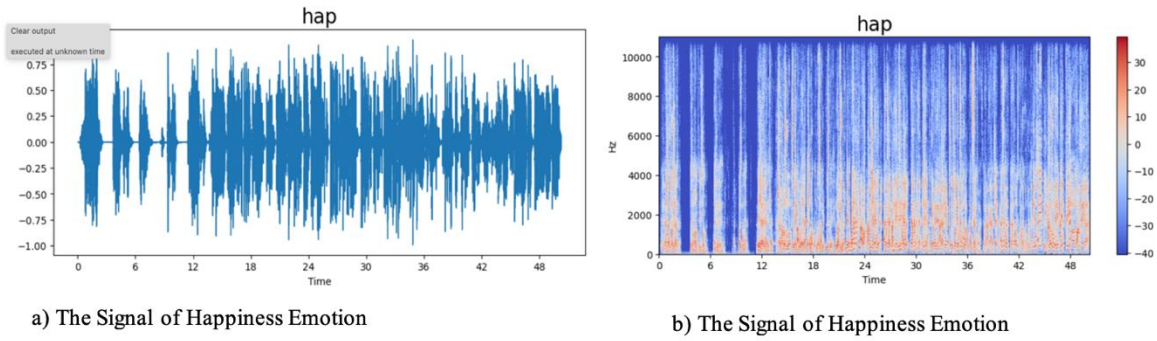


Figure 33

The Waveforms for the SAD Emotion

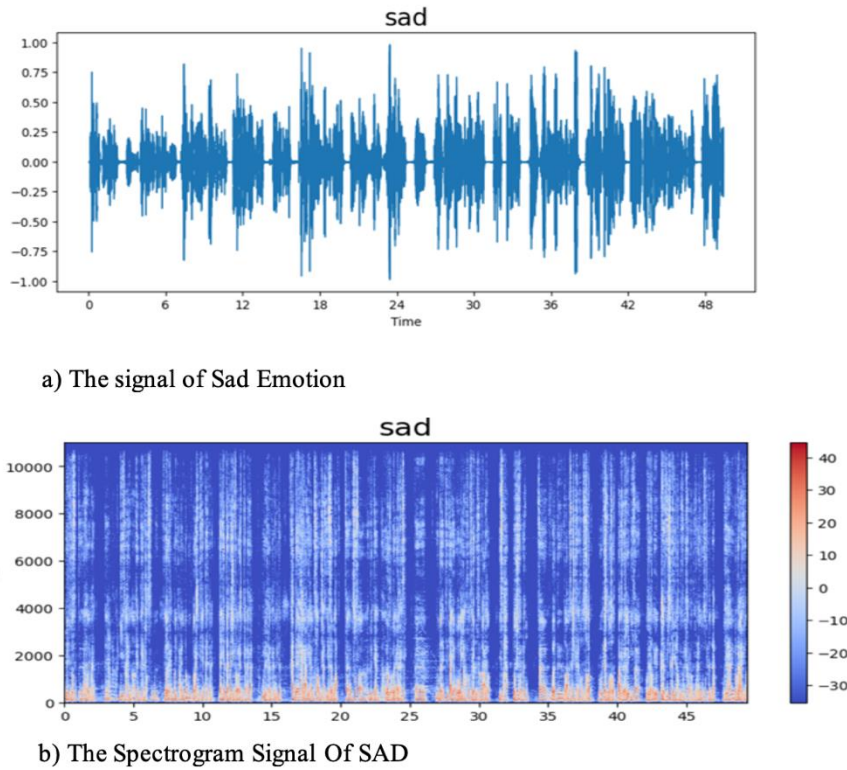


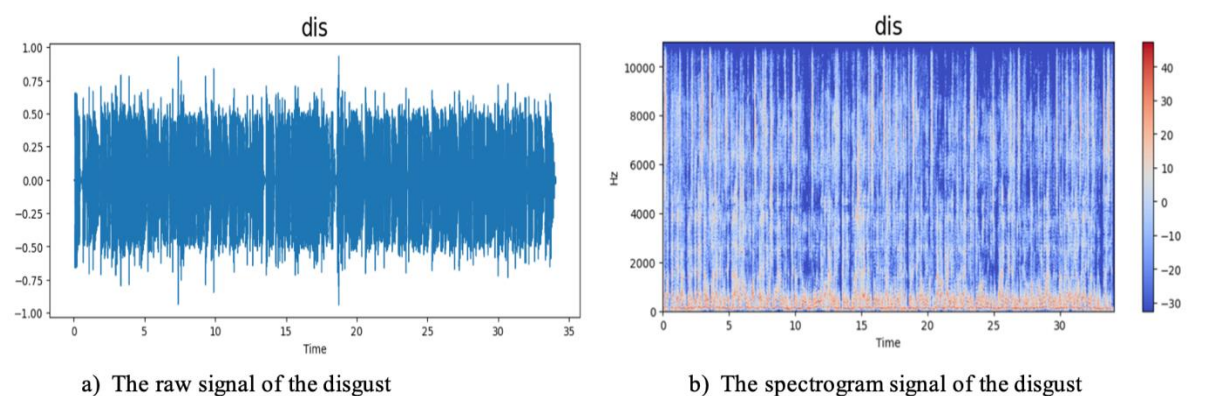
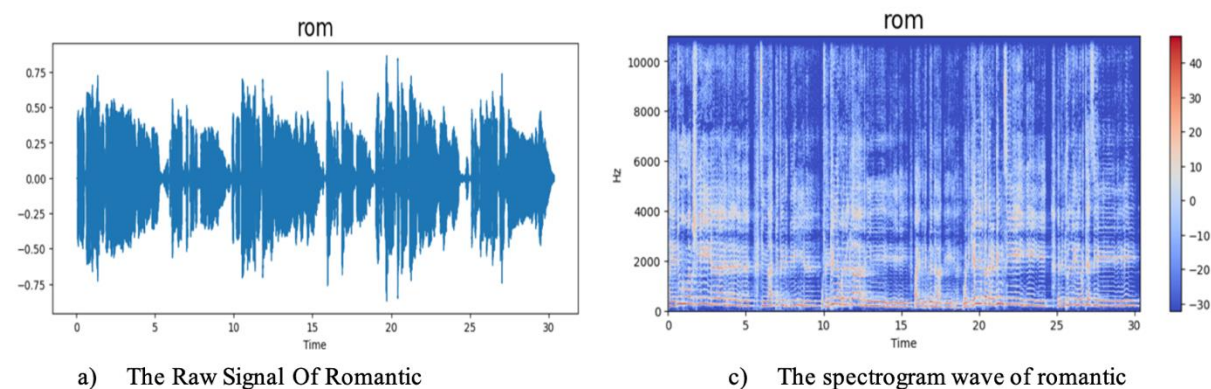
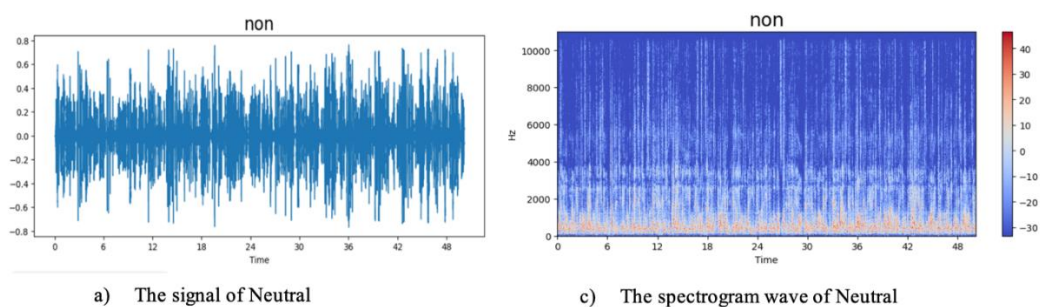
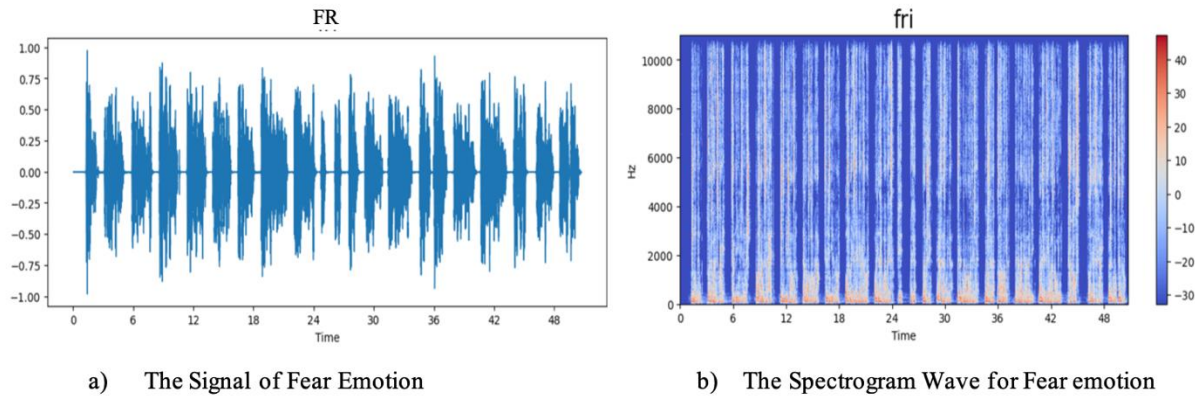
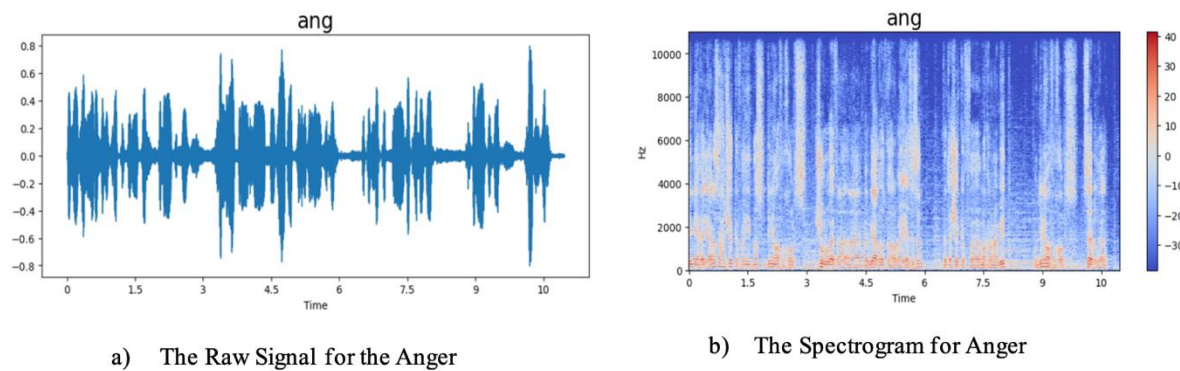
Figure 34*The Waveforms for the Disgust Emotion***Figure 35***The Waveforms for the Romantic Emotion***Figure 36***The Waveforms for the Neutral Emotion*

Figure 37*The Waveforms for the Fear Emotion***Figure 38***The Waveforms for the Anger Emotion*

5.3 Deep Learning Approaches

5.3.1 Convolutional Neural Network (CNN)

A Convolutional Neural Network (CNN) is a type of artificial neural network designed specifically for processing grid-like input, such as images and spectrograms derived from audio signals. Convolutional Neural Networks (CNNs) are highly effective in capturing hierarchical features through the use of convolutional layers.

- *The fundamental components of a Convolutional Neural Network:*

1. Convolutional Layers:

- Convolutional layers employ filters (kernels) to analyze specific regions of input data.
- These filters are utilized to process the input, identifying adjacent patterns and gaining information about their attributes.

2. Aggregation Layers: - Pooling layers diminish the input's size by reducing its spatial dimensions.

Max pooling is a technique that entails choosing the maximum value from a group of adjacent pixels, resulting in a reduction in the data's dimensionality.

3. Flatten Layer: - The flatten layer serves the function of transforming the output from previous layers into a vector with just one dimension.

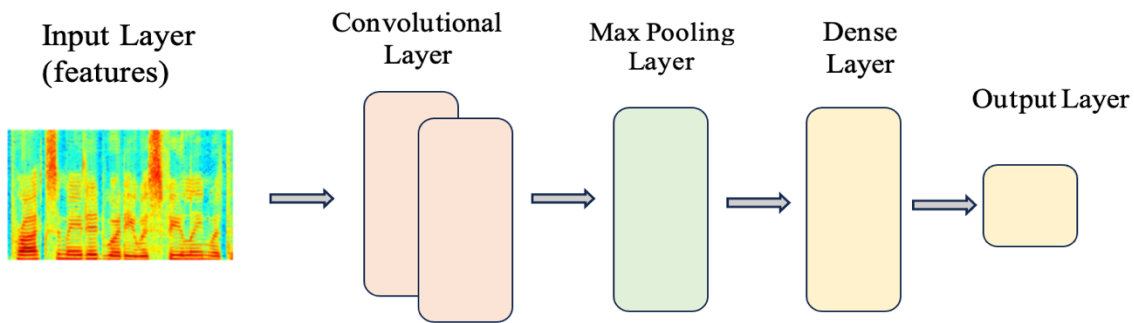
- This stage involves preprocessing the data to make it suitable for usage by the fully connected layers.

4. Dense Layers: Dense layers create connections between every neuron in one layer and every neuron in the next layer.

- These layers are frequently utilized for the purpose of categorizing tasks.

5. Dropout: Dropout layers randomly remove a fraction of neurons during training. This method helps prevent overfitting by improving the model's capacity to generalize.

Figure 39 below represents the CNN Framework.

Figure 39*CNN Architecture*

- *Hyperparameters in a Convolutional Neural Network (CNN):*

1. Layer of Convolution Predefined parameters that are established before to the initiation of a machine learning process and are not acquired through the analysis of the data. They govern the conduct and efficacy of the algorithm.

- Filter quantity: Specifies the number of filters that are applied to the input.

The user did not provide any text. The kernel size pertains to the dimensions of the filters employed.

The stride parameter dictates the magnitude of each individual movement that the filter makes while traversing the input.

- Padding: Determines the incorporation of extra pixels around the input to avoid information loss at the borders.

2. Hyperparameters for the Pooling Layer: - Pool size: Specifies the size of the pooling window, which determines the dimensions of the pooling operation.

The stride parameter dictates the magnitude of the step made by the pooling window.

3. Flatten Layer: - This layer does not necessitate any distinct hyperparameters since it executes a direct transformation.

4. Dense Layer Hyperparameters refer to the predetermined settings of a machine learning algorithm that are not derived from the data and are established prior to its execution.

- Unit count: Represents the total number of neurons inside the layer.

5. Discontinuation of education: The dropout rate refers to the percentage of neurons that are deliberately rejected or "dropped out" during the training process.

6. Learning Rate: The learning rate determines the magnitude of the steps made during optimization and impacts the speed at which the model reaches convergence.

7. Batch Size: The batch size controls the number of samples used in each training iteration.

8. Epochs: The epochs parameter specifies the number of iterations that the model will go through during training on the entire dataset.

Optimizing these hyperparameters is essential for attaining the best possible performance in a Convolutional Neural Network (CNN). Frequently, it entails a blend of intuition, experimentation, and, occasionally, automated techniques such as grid search or random search. Tweaking these parameters can have an effect on the model's precision, duration of training, and ability to apply to unfamiliar data.

5.3.2 Bidirectional Long Short-Term Memory (BI-LSTM)

A Bidirectional Long Short-Term Memory (BI-LSTM) is a type of recurrent neural network (RNN) that processes input data in both forward and backward directions. It is very effective for tasks that require converting one sequence to another, when understanding the

context from both previous and following information is crucial. BI-LSTM networks are an upgraded iteration of regular LSTMs that demonstrate improved efficacy in capturing long-term dependencies.

Hyperparameters of the BI-LSTM

Hyperparameters refer to the settings or configurations that need to be determined before training a bidirectional long short-term memory (BI-LSTM) model.

1. LSTM Unit Count: - Specifies the number of memory cells or units inside the LSTM layer.

- Augmenting the numerical values allows the model to proficiently capture complex patterns, but it may also result in elevated computational requirements.

2. Layer Count: This parameter determines the amount of complexity of the BI-LSTM network. Networks with a greater number of layers have the capacity to capture more intricate hierarchical properties. However, they are also more prone to overfitting.

3. Dropout Rate: Dropout is a regularization technique that entails eliminating a randomly selected subset of neurons from the training process.

The dropout rate is a metric that quantifies the probability of a neuron being deactivated, hence reducing the risk of overfitting.

4. Bidirectional Layer: The bidirectional layer is a hyperparameter that controls the processing of input sequences in both forward and backward directions.

The user did not provide any text. To enable bidirectional processing, set it to True.

5. Activation Function: - Specifies the activation function used within the LSTM cells.

The user did not provide any text. Common choices include of the tanh and sigmoid functions.

6. Learning Rate: The learning rate determines the magnitude of the optimization stages and affects the speed at which the model reaches convergence.

7. Batch Size: The batch size controls the number of sequences used in each training iteration.

No text was provided by the user. Decreasing the batch size can have a regularization effect and alleviate overfitting.

8. Epochs: The epoch count is the total number of training iterations that the model will undergo using the whole dataset.

9. The optimizer is responsible for selecting the optimization approach used during training, such as Adam, RMSprop, or SGD.

10. Loss Function: The loss function measures the difference between the anticipated values and the actual values.

Common choices for sequence-to-sequence challenges frequently include categorical crossentropy or mean squared error as loss functions.

Optimizing these hyperparameters is crucial for achieving optimal performance in a BI-LSTM network. The selection of hyperparameters depends on the specific task, dataset, and available computational resources. Experimentation, and in certain cases, the use of automated hyperparameter tuning methods, can help identify the best configuration for a particular problem.

5.3.3 Artificial Neural Network (ANN)

An Artificial Neural Network (ANN) is a computer model that draws inspiration from the structure and operation of the human brain. The system is comprised of linked nodes, also known

as neurons, that are arranged in layers. Artificial neural networks (ANNs) are extensively employed for diverse machine learning endeavours, encompassing classification, regression, and pattern recognition. The performance of an Artificial Neural Network (ANN) is dictated by its structure and hyperparameters.

- *Hyperparameters in an Artificial Neural Network:*

1. Layer Count: - Determines the level of complexity in the neural network.

- Additional layers have the ability to capture complex details, but they also raise the possibility of overfitting.

2. Neuronal Count in Each Layer: - Indicates the quantity of neurons in each layer.

- A greater quantity of neurons has the capacity to catch intricate patterns, but necessitates additional processing resources.

3. Activation Function: - Each layer has the flexibility to utilize distinct activation functions.

The selection of activation functions is contingent upon the problem's nature and the data's properties.

4. Learning Rate: - The learning rate is responsible for determining the magnitude of each step taken throughout the optimization process.

- An excessively high learning rate might lead to the model beyond the optimal weights, whilst an excessively low rate can result in delayed convergence.

5. Dropout Rate: - Dropout is a regularization strategy that involves excluding a portion of randomly chosen neurons from the training process.

- It aids in mitigating overfitting by enhancing the model's generalization.

6. Batch Size: - The batch size determines the amount of training samples utilized in each iteration.

- Increasing the size of batches can enhance computing efficiency, but it necessitates a greater amount of memory.

7. Epochs: - The number of epochs determines the number of times the model iterates over the full training dataset.

- It impacts the convergence of the model.

8. Optimizer: - Optimizers modify the weights during training by employing optimization methods.

The user did not provide any text. Popular optimization algorithms include Adam, RMSprop, and Stochastic Gradient Descent (SGD).

9. Loss Function: - The loss function quantifies the disparity between expected and actual values.

The user did not provide any text. Various issues may necessitate distinct loss functions, such as mean squared error or categorical cross entropy.

10. Initialization procedures: - Weight initialization procedures determine the starting values of the weights.

- The commonly used techniques comprise of random initialization, Xavier/Glorot initialization, and He initialization.

Choosing suitable hyperparameters is essential for effectively training a neural network. The ideal values are contingent upon the particular job, dataset, and computational resources at hand. Experimentation and, in certain instances, the utilization of automated hyperparameter

tuning approaches can assist in determining the optimal configuration of parameters for a specific situation.

5.3.4 Gated Recurrent Unit (GRU) Networks:

The Gated Recurrent Unit (GRU) is a specific architecture of Recurrent Neural Network (RNN) that aims to overcome the constraints of conventional RNNs, namely the issue of the vanishing gradient problem. GRUs, which were invented by (Cho et al., 2014) utilize gating techniques to selectively update and reset information. This feature makes them particularly effective in capturing long-term relationships in sequential data.

Refer to section 7.3.3 for further information.

- *GRU hyperparameters:*

1. Unit Quantity: - Specifies the quantity of hidden units or neurons in the GRU layer.

- Increasing the number of units has the ability to capture intricate patterns, but it necessitates a greater amount of processing resources.

2. Activation Function: - The activation function used on the candidate activation.

- Popular options often consist of hyperbolic tangent (tanh) or rectified linear unit (ReLU).

3. Dropout Rate: - Dropout is a regularization strategy that involves excluding a portion of randomly chosen units during the training process. - Its purpose is to mitigate overfitting by enhancing the model's ability to generalize.

4. Initialization Method: - The technique employed to set the initial values of the weights of the GRU units.

- The commonly used techniques include of random initialization, Xavier/Glorot initialization, and He initialization.

5. Batch Size: - Batch size determines the number of sequences that are simultaneously processed during training.

The user did not provide any text. Increasing the batch size can enhance computing efficiency, but it necessitates a greater amount of memory.

6. Learning Rate: - The learning rate governs the magnitude of each step taken throughout the optimization process.

- It has an impact on the convergence of the model and the quality of the acquired representations.

7. Layer Count: - Indicates the depth of the GRU network by specifying the number of stacked GRU layers.

The user did not provide any text. More intricate networks have the ability to capture intricate hierarchical characteristics, but they may be susceptible to overfitting.

8. Epochs: - The number of epochs determines the total number of times the model iterates over the complete training dataset.

- It has an impact on the convergence of the model.

9. Optimizer: - The optimization method employed to adjust the weights throughout the training process.

- Popular optimization algorithms include Adam, RMSprop, and Stochastic Gradient Descent (SGD).

10. Loss Function: - The loss function quantifies the disparity between expected and actual values.

The user did not provide any text. Various issues may necessitate distinct loss functions, such as mean squared error or categorical cross entropy.

Choosing suitable hyperparameters for GRU networks is crucial for attaining high performance on jobs using sequential data. The optimal values are contingent upon the particular issue, dataset, and computational resources at hand. Utilizing experimentation and hyperparameter tweaking approaches can facilitate the identification of the optimal configuration for a specific job.

5.4 Hybrid Deep Learning Models

5.4.1 CNN-Bi-LSTM Hybrid Model

The CNN-Bi-LSTM hybrid model integrates Convolutional Neural Network (CNN) layers with Bidirectional Long Short-Term Memory (Bi-LSTM) layers. This architectural design capitalizes on the advantages of both Convolutional Neural Networks (CNNs) and Long Short-Term Memory (LSTMs), making it very efficient for jobs that require processing sequential input and spatial information. The CNN layers are designed to detect and analyze spatial patterns, while the Bi-LSTM layers are responsible for capturing long-range relationships in sequential data in both forward and backward directions.

- *The hyperparameters of the CNN-Bi-LSTM hybrid model are:*

1. Quantity of CNN Filters:

- Determines the quantity of filters present in every convolutional layer.
- Increased utilization of filters enables the detection of a wider range of spatial characteristics.

2. CNN Kernel Size: - Determines the dimensions of the convolutional filters. - Larger kernels capture broader spatial patterns, whereas smaller kernels emphasize more intricate features.

3. Pooling Size: - Specifies the dimensions of the pooling window in the pooling layers.

- Pooling decreases the size of spatial dimensions and extracts the most prominent characteristics.

4. Bi-LSTM Unit Count: - Determines the quantity of memory units (neurons) in the Bi-LSTM layers.

- Increasing the number of units enables the model to effectively capture intricate sequential patterns.

5. Dropout Rate (CNN and Bi-LSTM): - Dropout is implemented to mitigate overfitting by stochastically disregarding units during the training process.

- The dropout rate is a measure of the proportion of units that are removed or dropped out.

6. Learning Rate: - Determines the magnitude of the step taken during optimization.

- A greater learning rate can accelerate convergence but may result in overshooting.

7. Optimizer: - The method employed to update the weights of the model in order to optimize its performance.

- Popular optimization algorithms include Adam, RMSprop, and Stochastic Gradient Descent (SGD).

8. Batch Size: - Determines the number of sequences that are simultaneously processed during training.

- Increasing the batch size can result in faster computational processing, but it necessitates a greater amount of memory.

9. The number of LSTM layers determines the depth of the Bi-LSTM network by defining the number of stacked Bi-LSTM layers.

- Networks with greater depth have the ability to record more complex and detailed sequential relationships.

10. Activation Functions: - Activation functions are applied to the output of each layer, with ReLU often used for CNNs and sigmoid or tanh for LSTMs.

11. Epochs: - The number of epochs determines the number of times the model iterates over the complete training dataset. - It impacts the model's convergence.

12. Loss Function: - The loss function quantifies the discrepancy between expected and actual values. - Various issues may necessitate distinct loss functions, such as mean squared error or categorical crossentropy.

Optimizing the hyperparameters for the CNN-Bi-LSTM hybrid model requires conducting experiments and fine-tuning the hyperparameters. The ideal values are contingent upon the particular job, attributes of the dataset, and the computational resources at hand. Optimizing these parameters can result in enhanced model performance and the ability to apply the model to other types of sequential data jobs.

5.4.2 CNN-GRU Hybrid Model

The CNN-GRU hybrid model combines Convolutional Neural Network (CNN) layers with Gated Recurrent Unit (GRU) layers. This architectural design leverages the advantages of Convolutional Neural Networks (CNNs) in extracting features from spatial data and Gated Recurrent Units (GRUs) in capturing sequential relationships. The CNN-GRU model demonstrates efficacy in tasks that require analysis of both spatial and temporal patterns, such as the processing of picture sequences.

- *The hyperparameters of the CNN-GRU hybrid model are:*

1. Quantity of CNN filters:

- Determines the quantity of filters in every convolutional layer.
- Additional filters enhance the model's ability to capture a wide range of spatial

characteristics.

2. CNN Kernel Size: - Determines the dimensions of the convolutional filters.

The size of the kernel directly affects the scale at which spatial details are collected.

3. CNN Stride:

- The stride of the convolutional filter as it traverses the input.
- Increasing the stride size decreases the output dimensions, but it might result in the

loss of information.

4. Pooling Type and Size: - Pooling layers diminish spatial dimensions.

- The two common kinds of pooling are max pooling and average pooling.
- The spatial reduction factor is determined by the size of the pooling.

5. GRU Unit Count: - Determines the quantity of memory units (neurons) inside the GRU layers.

- Increasing the number of units improves the model's capacity to capture sequential patterns.

6. Dropout Rate (CNN and GRU): - Dropout is used as a measure to mitigate overfitting during the training process.

The dropout rate indicates the proportion of units that will be removed.

7. Learning Rate: - Determines the magnitude of the step taken during optimization.

- A suitable learning rate facilitates the process of convergence without exceeding the desired outcome.

8. Optimizer: - The method employed to update the weights of the model in order to optimize its performance.

- Adam, RMSprop, and Stochastic Gradient Descent (SGD) are frequently used optimization algorithms.

9. Batch Size: - Determines the number of sequences that are simultaneously processed during training.

- Increasing the batch size can result in faster computational processing, but it necessitates a greater amount of memory.

10. The number of CNN and GRU layers determines the level of complexity in the CNN and GRU networks.

- Increasing the number of layers allows the model to effectively capture intricate spatial and temporal patterns.

11. Activation Functions: - Activation functions are applied to the output of each layer, often using Rectified Linear Unit (ReLU) for Convolutional Neural Networks (CNNs) and Gated Recurrent Units (GRUs).

12. Epochs: - The number of epochs determines the total number of times the model iterates over the complete training dataset.

- A suitable number of epochs is necessary for achieving convergence.

13. Loss Function: - The loss function quantifies the disparity between expected and actual values.

Optimizing the hyperparameters for the CNN-GRU hybrid model requires conducting experiments and fine-tuning the hyperparameters. The ideal values are contingent upon the particular job, attributes of the dataset, and the computational resources at hand. Optimizing these

parameters can result in enhanced model performance and increased ability to apply the model to different tasks requiring spatial and temporal data.

5.4.3 GRU-Bi-LSTM Hybrid Model

The GRU-Bi-LSTM hybrid model integrates Gated Recurrent Unit (GRU) layers with Bidirectional Long Short-Term Memory (Bi-LSTM) layers. The purpose of this design is to effectively capture sequential relationships and patterns in data. The GRU and Bi-LSTM layers collaborate to acquire and depict both immediate and enduring interconnections in sequential data.

• *The hyperparameters of the GRU-Bi-LSTM hybrid model are:*

1. Bi-LSTM Unit Count: - Determines the quantity of memory units (neurons) present in the Bi-LSTM layers.

- Increasing the number of units enables the model to effectively capture intricate sequential patterns.

2. Quantity of GRU Units:

- Determines the quantity of memory units present in the GRU layers.

The quantity of units directly impacts the model's capacity to represent short-term interdependencies.

3. Dropout Rate (Bi-LSTM and GRU): - Dropout is implemented to mitigate overfitting by stochastically disregarding units throughout the training process.

- The dropout rate is a measure of the proportion of units that are removed or discarded.

4. Learning Rate: - Determines the magnitude of the step taken during optimization.

- Increasing the learning rate can accelerate the convergence process, but it also carries the risk of overshooting the desired outcome.

5. Optimizer: - The optimization procedure employed to update the weights of the model.

The user did not provide any text. Popular optimization algorithms include Adam, RMSprop, and Stochastic Gradient Descent (SGD).

6. Batch Size: - Determines the number of sequences that are simultaneously processed during training.

- Increasing the batch size can lead to faster calculation, but it also necessitates a greater amount of memory.

7. Bi-LSTM Layer Count: - Specifies the number of stacked Bi-LSTM layers, determining the depth of the Bi-LSTM network.

- Networks with greater depth have the ability to record more complex and detailed sequential relationships.

8. GRU Layer Count: - Determines the depth of the GRU network by specifying the number of stacked GRU layers.

The network's capacity to capture short-term trends is influenced by its depth.

9. Activation Functions: - Activation functions are applied to the output of each layer, usually ReLU for GRUs and LSTMs.

10. Epochs: - The number of epochs determines the total number of times the model iterates over the complete training dataset.

- It has an impact on the model's convergence.

11. Loss Function: - The loss function quantifies the disparity between expected and actual values.

- Various issues may necessitate distinct loss functions, such as mean squared error or categorical crossentropy.

Optimizing the hyperparameters for the GRU-Bi-LSTM hybrid model requires conducting experiments and fine-tuning the hyperparameters. The ideal values are contingent upon the particular job, attributes of the dataset, and the computational resources at hand. Optimizing these parameters can result in enhanced model performance and better ability to apply the model to different types of sequential data jobs.

5.4.4 GRU-CNN Hybrid Model

A hybrid model, combining a Gated Recurrent Unit (GRU) and Convolutional Neural Network (CNN), is an advanced neural network architecture designed to tackle issues in diverse applications, especially those incorporating sequential and spatial input.

- *Hyperparameters in GRU-CNN Hybrid Model:*

The hybrid model, known as GRU-CNN (Gated Recurrent Unit - Convolutional Neural Network), relies on certain hyperparameters that significantly impact its performance and behaviour. Tuning these hyperparameters is crucial for maximizing the model's capacity to efficiently acquire knowledge from data. The following are the essential hyperparameters in the GRU-CNN hybrid model:

1. GRU Unit Count: - Definition: The number of GRU units determines the capacity of the GRU layer to capture sequential dependencies in the input data.

- Tuning Strategy: Employing a process of experimentation with various settings in order to ascertain the ideal equilibrium between the intricacy of the model and its capacity to capture long-term interdependence.

2. CNN Filter Count: - Definition: The number of filters in the CNN layer determines the number of distinct features that the model can extract from the input data.

- Optimization Approach: Tuning this parameter enables the model to capture diverse properties at distinct levels of abstraction. Inadequate utilization of filters may result in inadequate extraction of features, whilst excessive utilization may lead to overfitting.

3. CNN Kernel Size: - Definition: The size of the convolutional kernels in the CNN layer defines the spatial range of the learnt features.

- Optimization Approach: Choosing a suitable kernel size requires taking into account the characteristics of the incoming data. Smaller kernels are able to detect intricate details, but larger kernels are better at identifying broader patterns.

4. Pooling Size: - Definition: Pooling layers decrease the size of the spatial dimensions in the output of the CNN, which reduces computational requirements and emphasizes the most informative characteristics.

- Optimization Approach: The size of the pooling operation impacts the degree of downsampling. Reduced dimensions retain greater spatial details, but increased dimensions amplify downsampling.

5. Dropout Rate: - Definition: Dropout is a regularization strategy that randomly deactivates a portion of units during training in order to mitigate the risk of overfitting.

- Optimization Approach: Modifying the dropout rate aids in managing overfitting. Increased rates of regularization can enhance regularization effects but may impede the learning process.

6. Learning Rate: - Definition: The learning rate is a parameter that controls the step size during optimization and affects the speed at which the model reaches convergence.

- Optimization Approach: Meticulous choice of the learning rate is crucial. An excessively high rate can lead to model divergence, whilst an excessively low rate can result in sluggish convergence.

7. Batch Size: - Definition: The quantity of samples that are handled during each iteration of training.

- Optimization Approach: The model's generalization and training speed are influenced by the batch size. Reducing the size of the batches may result in the presence of unwanted disturbances, but increasing the size of the batches may necessitate a greater amount of memory.

8. Epochs: - Definition: The number of complete passes the model makes over the entire training dataset during training.

- Optimization Approach: Ensuring an appropriate balance of epochs is crucial in order to mitigate the risk of overfitting. Early termination can be implemented depending on the validation performance.

9. Activation Functions: - Definition: The selection of activation functions in various layers, such as ReLU (Rectified Linear Unit) or Sigmoid, impacts the model's ability to capture non-linear relationships.

- Tuning Strategy: Conducting experiments with various activation functions in order to identify the optimal combination that is most suitable for the given situation.

10. Weight Initialization: - Definition: The technique employed to assign initial values to the weights in the neural network.

- Optimization Approach: Appropriate weight initialization mitigates problems such as disappearing or exploding gradients. Popular techniques include Glorot initialization and He initialization.

11. Choice of Optimizer: - Description: The method used for optimization during training, such as Adam or RMSprop.

- Optimization Approach: Optimizers exhibit diverse convergence rates and might yield disparate outcomes across different workloads. Experimentation facilitates the identification of the best appropriate optimizer.

12. Learning Rate Scheduler: - Definition: A function that modifies the learning rate during the training process.

- Tuning Strategy: Implementing learning rate schedulers, such as step decay or exponential decay, can enhance the convergence and generalization capabilities of the model.

Optimizing these hyperparameters requires a blend of expertise in the field, conducting experiments, and frequently employing methods such as grid search or random search to systematically explore the hyperparameter space.

5.4.5 GRU-ANN Hybrid approach

Hyperparameters in GRU-ANN Hybrid Model:

The GRU-ANN (Gated Recurrent Unit - Artificial Neural Network) hybrid model involves a set of hyperparameters that significantly influence its performance and generalization. Adjusting these hyperparameters is crucial to achieve optimal results in terms of accuracy and efficiency. Here are the key hyperparameters in the GRU-ANN hybrid model:

1. Number of GRU Units:

- Description: Determines the capacity of the GRU layer to capture sequential dependencies in the input data.

- Tuning Strategy: Experiment with different values to find the right balance between model complexity and the ability to capture long-term dependencies.

2. Number of ANN Hidden Layers:

- Description: Specifies the number of hidden layers in the ANN part of the hybrid model.
- Tuning Strategy: Adjusting the number of hidden layers impacts the model's capacity to learn hierarchical representations. More layers provide greater representational power but may increase the risk of overfitting.

3. Number of Neurons in ANN Hidden Layers:

- Description: Determines the number of neurons (nodes) in each hidden layer of the ANN.
- Tuning Strategy: Varies the number of neurons to control the capacity of the ANN.

This parameter influences the complexity of the learned representations.

4. GRU-CNN Integration Point:

- Description: Specifies the point at which the GRU and CNN layers are integrated.
- Tuning Strategy: The integration point determines how the features learned by GRU and CNN are combined. Experimentation is needed to find the optimal integration strategy.

5. Kernel Size in CNN:

- Description: The size of the convolutional kernels in the CNN layer.
- Tuning Strategy: Adjusting the kernel size influences the spatial extent of learned features. It involves a trade-off between capturing fine-grained details and global patterns.

6. Pooling Size in CNN:

- Description: Determines the size of the pooling operation in the CNN layer.

- Tuning Strategy: Pooling reduces spatial dimensions and emphasizes the most important features. Adjust the pooling size based on the desired balance between information retention and downsampling.

7. Dropout Rate:

- Description: Probability of dropping out neurons during training to prevent overfitting.
- Tuning Strategy: The dropout rate regulates overfitting. Experiment with different rates to find the right level of regularization.

8. Learning Rate:

- Description: Determines the step size during optimization.
- Tuning Strategy: Select an appropriate learning rate for effective convergence. Too high a rate may cause divergence, while too low a rate may result in slow convergence.

9. Batch Size:

- Description: Number of samples processed in each training iteration.
- Tuning Strategy: Batch size affects generalization and training speed. Experiment with different sizes to find the optimal trade-off.

10. Activation Functions:

- Description: Choices for activation functions in different layers, such as ReLU, Sigmoid, or Tanh.
- Tuning Strategy: Different activation functions impact the model's ability to capture non-linearities. Experiment to find the most suitable functions for the task.

11. Weight Initialization:

- Description: Method used to initialize weights in the network.

- Tuning Strategy: Proper weight initialization helps prevent issues like vanishing or exploding gradients. Common methods include Glorot or He initialization.

12. Optimizer Choice:

- Description: The optimization algorithm used during training, such as Adam or RMSprop.

- Tuning Strategy: Different optimizers may have varying convergence speeds. Choose an optimizer that suits the characteristics of the problem.

13. Learning Rate Scheduler:

- Description: A function that adjusts the learning rate during training.

- Tuning Strategy: Learning rate schedulers, like step decay or exponential decay, can aid convergence and generalization.

Optimizing these hyperparameters involves a combination of domain knowledge, intuition, and systematic experimentation. Fine-tuning is often performed through techniques like grid search or random search to efficiently explore the hyperparameter space and find the configuration that yields the best results for a given task.

5.5 Evaluation Metrics

When it comes to speech recognition, evaluating the effectiveness of the system is an important factor to consider. Evaluation metrics play a crucial role in assessing the effectiveness of machine learning and data analysis algorithms by offering objective indicators. This is especially pertinent in sentiment analysis, encompassing its use in voice emotion identification. The metrics provide valuable information on the model's accuracy, precision, recall, and general efficacy in categorizing attitudes stated in the speech. Within the domain of Arabic voice sentiment analysis, these metrics are utilized to assess the system's competence in accurately

detecting and categorizing sentiments, including positive, negative, or neutral. Refer to section 4.8 for further information.

5.6 Conclusion

We presented a succinct summary of the approach utilized in this thesis. This section provides comprehensive information on the process of data collecting and preparation, which includes annotation, feature extraction, and data splitting. This chapter examines different Deep Learning Models for Arabic voice recognition, such as Convolutional Neural Networks (CNNs), Bidirectional LSTM (BI-LSTM), Gated Recurrent Unit (GRU), and Artificial Neural Network (ANN). In addition, hybrid models were introduced, including the CNN-Bi-LSTM model, GRU-Bi-LSTM model, and CNN-GRU model. The section finishes by providing a detailed account of the assessment measures employed in the thesis.

Chapter 6: Experiment and Results

6.1 Operating Environment

The implementation of the proposed technique requires the use of different tools and concepts, which means that it is necessary to choose a suitable platform for developing the speech recognizer. Here is a concise summary of the tools used in the implementation.

6.1.1 Jupyter Notebook

The Jupyter Notebook is a freely available online application that is widely acknowledged as a computational notebook. Researchers may utilize this tool to create and disseminate papers that incorporate executable code, mathematical calculations, graphic representations, and written content. The word "Jupyter" is a portmanteau of the programming languages Julia (Ju), Python (Py), and R. Jupyter notebooks can be performed either on a local machine or on a remote server, sometimes known as a cloud server. The papers are organized into several cells, each containing either script language or markdown code. This technology enables the easy and accurate sharing and replication of scientific research (Perkel, 2018).

6.1.2 TensorFlow

TensorFlow is a robust machine learning framework designed specifically for handling large-scale applications in diverse situations. The system coordinates the elements of a dataflow diagram among many computers in a cluster and different processing devices within a single machine, such as multicore CPUs, general-purpose GPUs, and specialist ASICs like Tensor Processing Units (TPUs). TensorFlow is very proficient at facilitating a wide range of applications, particularly in the areas of training and inference on deep neural networks. It provides the ability to explore innovative optimizations and training methods, allowing for flexibility and experimentation. TensorFlow, which was initially an open-source project, is now

extensively used in production by several Google services. It has been well recognized in the field of machine learning research (Abadi et al., 2016).

6.1.3 Google Colaboratory (Colab)

Google Colaboratory, often known as Colab, functions as a platform specifically created to streamline the process of machine learning instruction and research. The system utilizes a Jupyter-based framework and functions similarly to collaborative documents in Google Docs, facilitating collaborative work on the same notebook. Colab offers pre-configured Python 2 and 3 runtimes that include with critical machine learning and artificial intelligence libraries including TensorFlow, Matplotlib, and Keras.

The Colab platform operates on an Ubuntu 17.10 64-bit system, equipped with an undisclosed brand of Intel Xeon processor that has two cores running at a speed of 2.3 GHz. It also has a total of 13 GB RAM. The device is equipped with an NVIDIA Tesla K80, which features a GK210 chipset, 12 GB of RAM, and 2496 CUDA cores operating at a frequency of 560 MHz (Carneiro et al., 2018). The suggested model will be implemented on Colab, utilizing its cost-free Jupyter notebook environment.

6.2 Experiment Setup

The ASR model implementation was facilitated by utilizing a diverse range of library packages and modules in this project.

- **Librosa:** Librosa is a Python module specifically created for the purpose of analyzing audio and music. It includes many submodules, such as `librosa.beat`, `librosa.core`, `librosa.decompose`, `librosa.display`, `librosa.effects`, and others. The feature of this tool facilitates the methodical study of audio data.

- FFmpeg: FFmpeg is a Python utility that can convert video and audio formats. Moreover, it has the capability to capture and encode content from many sources, such as television.
- Transformers, which are commonly used in deep learning, has the ability to perform self-attention. They are applied in several fields such as computer vision (CV) and natural language processing (NLP). Transformers, which are specifically tailored for sequential data processing, are frequently utilized in activities such as translation and text summarization.
- The Arabic-pronounce Python tool is designed to transform discrete Arabic words into phoneme sequences, facilitating the investigation of pronunciation in Arabic language components.
- Unidecode is a Python module that converts Unicode strings into Python 3 compliant Unicode strings. Unidecode significantly improves string values by eliminating diacritical marks from characters.
- PyYAML==5.1: PyYAML is a comprehensive Python framework for YAML, designed for data serialization. PyYAML is a widely used tool in Python 3 that helps in exchanging data in a way that is easy for humans to understand and interact with. It is particularly useful for maintaining clarity and readability.
- Codernitydb3 is a high-performance, cross-platform NoSQL database that is open-source and also has built-in support for HTTP. The main purpose of this system is to function as a cache, substantially decreasing the burden on the morphological analyzer.
- When it comes to audio file formats, each kind has its own unique benefits and limitations inside a Speech system.

- The .mp3 format is widely acknowledged as the most common format and is known for its lossy compression. Nevertheless, this compression procedure might lead to the loss of information.
- The .wav format is uncompressed, guaranteeing superior audio quality at the expense of bigger file sizes. It represents the benchmark for the audio quality of CDs. The wav format was used for this project due to its suitability and ease of use in Python.

6.2.1 Hyperparameters of the Preprocessing

The hyperparameters of Deep Learning Models are set based on the experiments. In Table 8, the hyperparameters of the preprocessing steps are illustrated.

Table 8

The Preprocessing Hyperparameters

Hyperparameter	Value
Samplpe_Rate	16000
Segment length	3
MFCC Features	40
Fourier_N	2048
Hop_L	512
Frame Factor	501
Frame slide factor	1.3

6.2.2 Deep Learning Models

- Convolutional Neural Network (CNN)

Table 9 below illustrates the CNN hyperparameters.

Table 9*Hyperparameters Values For CNN*

Hyperparameters	Value
Conv_filters	32
Conv_kernel_size	3
Pool size	2
Dense unites	128
Dropout rate	0.2
Learning rate	0.0001
Batch size	32
Epochs	100
Activation Function	Relu, softmax
Output Units (Number of Classes)	7

- Bidirectional Long Short-Term Memory (BI-LSTM)

Table 10 summarizes the hyperparameters of the BI-LSTM.

Table 10*Hyperparameters Values For BI-LSTM*

Hyperparameters	Value
BI-LSTM Unites	123
Dense unites	128
Dropout rate	0.2
Learning rate	0.0001

Batch size	512
Epochs	100
Activation Function	Relu, softmax
Output Units (Number of Classes)	7
Optimizer	Adam

- Artificial Neural Network (ANN)

Table 11 summarizes the hyperparameters of the ANN.

Table 11

Hyperparameters Values For ANN

Hyperparameters	Value
Number of Hidden Units in Dense Layer	256
Dropout rate	0.5
Learning rate	0.0001
Batch size	512
Epochs	100
Activation Function	Relu, softmax
Loss Function	sparse_categorical_crossentropy

- Gated Recurrent Unit (GRU)

The GRU hyperparameters are summarized in Table 12.

Table 12*Hyperparameters Values For GRU*

Hyperparameters	Value
Number of GRU Units	225
Activation	tanh
Dropout Rate	0.2 (20% dropout rate)
Optimizer	adam
Learning rate	0.001
Batch size	512
Epochs	100
Loss Function	sparse_categorical_crossentropy

6.2.3 Hybrid Deep Learning Models

- CNN-Bi-LSTM Hybrid Model

The hyperparameters for the CNN-Bi-LSTM model are summarized in Table 13.

Table 13*The Hyperparameters For CNN-BI-LSTM*

	Hyperparameters	Value
	Number of Convolutional Filters	64
CNN	Convolutional Kernel Size	(3, 3)
	Activation Function	ReLU
	MaxPooling	pool size (2, 2)

Bi-LSTM	Number of LSTM Units	123
	Number of Dense Neurons	128
	Activation Function	ReLU
	Dropout	0.2
Output Layer	Activation Function	Softmax
	Output Neurons (number of emotion classes)	7
	optimizer	Adam
optimizer	learning rate	0.001
	Loss Function	sparse_categorical_crossentropy
	Batch Size	512
	Epochs	100

- **CNN-GRU Hybrid Model**

Table 14 presents a summary of the hyperparameter values used for the CNN-GRU model.

Table 14

Hyperparameters Values For CNN-GRU

Hyperparameters		Value
GRU	Number of Units	64
	Number of Convolutional Filters	64
CNN	Convolutional Kernel Size	(3, 3)
	Activation Function	ReLU

Dense Layers	MaxPooling size	2
	Number of Dense layer 1	256
	Activation Function	ReLU
	Dropout	0.5
	Number of Dense layer 2	128
Output Layer	Activation Function	Softmax
	Output Neurons	7
		(number of emotion classes)
optimizer	optimizer	Adam
	learning rate	0.0001
	Loss Function	sparse_categorical_crossentropy
	Batch Size	512
	Epochs	100

- **GRU-Bi-LSTM Hybrid Model**

Table 15 summarizes the CNN-Bi-LSTM hyperparameters values.

Table 15

GRU-BI-LSTM Hyperparameters Values

Hyperparameters		Value
GRU units	gru_units	64
	lstm_units	64
Convolutional layers	Conv_ Filters	64
	Convolutional Kernel Size	(3, 3)
	Activation Function	ReLU
	pool size	(2, 2)
Dense layer	Number of LSTM Units	64
	dense_units	256
	Dropout	0.2
optimizer	optimizer	Adam
	learning rate	0.0001
	Batch Size	512
	Epochs	100

- **GRU-CNN Hybrid Model**

Table 16 summarizes the GRU-CNN hyperparameters values.

Table 16*GRU-CNN Hyperparameters Values*

Hyperparameters		Value
GRU	Number of unites	64
	input_shape=(X.shape[1], 1)	(1, 1)
	Conv1D	64
	kernel_size	3
CNN	Dense	256, 128
	Activation Function	ReLU
	MaxPooling size	2
	Activation Function	Softmax
Output Layer	Output Neurons	7
		(number of emotion classes)
	Dropout Rate:	0.5
optimizer	optimizer	Adam
	learning rate	0.0001
	Loss Function	sparse_categorical_crossentropy
	Batch Size	512
	Epochs	100

- **GRU-ANN Hybrid Model**

Table 17 summarizes the GRU-ANN hyperparameters values.

Table 17*GRU-CNN Hyperparameters Values*

Hyperparameters		Value
GRU	Number of unites	64
	input_shape=(X.shape[1], 1)	(2)
	Dense	256, 128
	Activation Function	ReLU
	MaxPooling size	2
Output Layer	Activation Function	Softmax
	Output Neurons	7 (number of emotion classes)
	Dropout Rate:	0.5
optimizer	optimizer	Adam
	learning rate	0.0001
	Loss Function	sparse_categorical_crossentropy
	Batch Size	512
	Epochs	100

6.3 Results and Discussion

The efficacy of the created speech model was evaluated through a thorough set of studies. The audio recordings were used to extract MFCC features, which were then employed to create Arabic speech emotion detection models. Nine distinct techniques were deployed, consisting of four Deep Learning Models and five hybrid Deep Learning Models. The algorithms were

implemented to assess the performance of the SER model using the recommended MFCC characteristics. The classifiers' performance, when paired with MFCC features, was assessed using evaluation measures.

6.4 Experiments results

Table 18 and Figure 40 depict the outcomes derived from the CNN model.

Table 18

CNN Performance

Emotion Type	Precision	Recall	F1-score
Anger	91	88	89
Disgust	86	81	83
Fear	96	92	94
Happiness	77	88	82
Neutral	80	87	83
Romantic	93	88	91
Sadness	83	81	82

The assessment metrics displayed in the preceding table and figure allow us to discern the highest and lowest levels of performance across the emotion categories, as determined by Precision, Recall, and F1-score.

Optimal performance: An emotion characterized by a feeling of apprehension or dread, often in response to a perceived threat or danger.

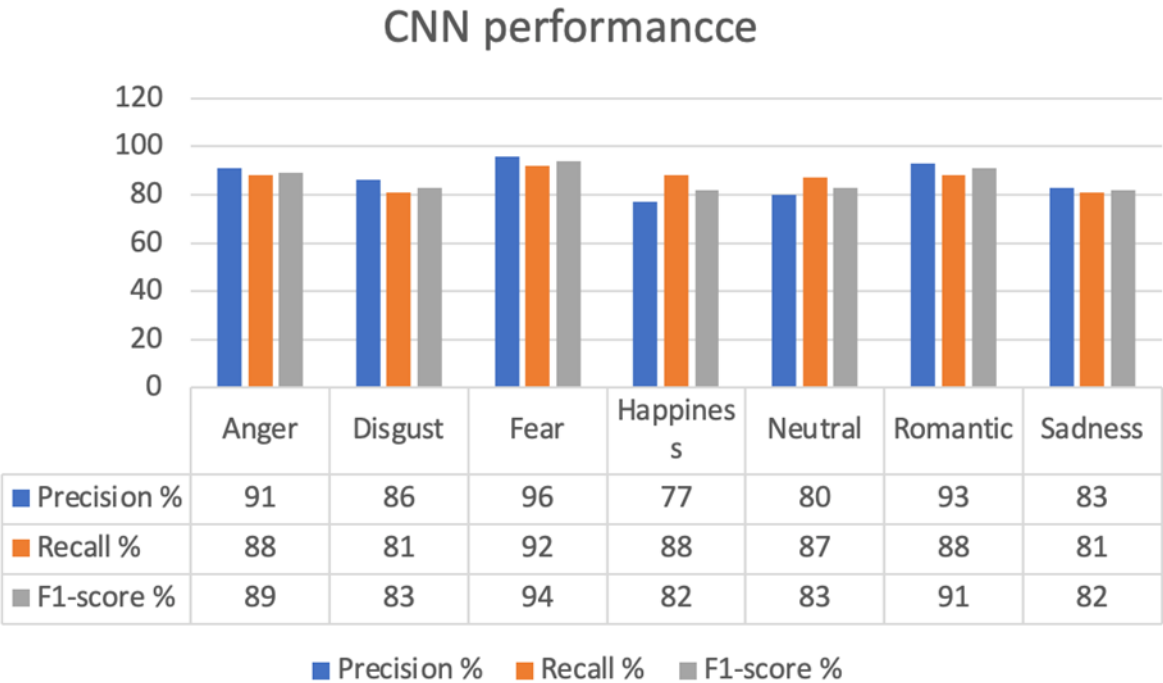
The precision of the data is 96%.

The recall rate is 92%.

The F1-score is 94%.

Figure 40

CNN Performance



Least satisfactory performance: Happiness

The level of precision is 77%.

The recall rate is 88%.

The F1-score is 82%.

Evaluation:

1. Optimal Performance (Fear): The emotion category "Fear" has the highest level of accuracy, recall, and F1-score compared to all other categories. This indicates that the model performs very well in reliably detecting occurrences of fear, attaining both a high level of precision in positive predictions and catching a significant proportion of real positive cases.

2. Lowest Performance (pleasure): The emotion category "Happiness" has the lowest recall rate (77), showing a notable difficulty in accurately identifying most occurrences of pleasure. Although the accuracy is reasonably high at 88%, indicating that the model accurately predicts pleasure most of the time, the poor recall indicates that the model fails to identify a significant percentage of actual occurrences of happiness.

Consequences:

- The model has great proficiency in detecting occurrences of fear, rendering it highly useful for applications that need fear detection, such as security or mental health monitoring.
- Happiness: The model has difficulties in fully capturing occurrences of happiness. This might have ramifications in applications where the recognition and treatment of pleasure is crucial, such as mental health support systems.

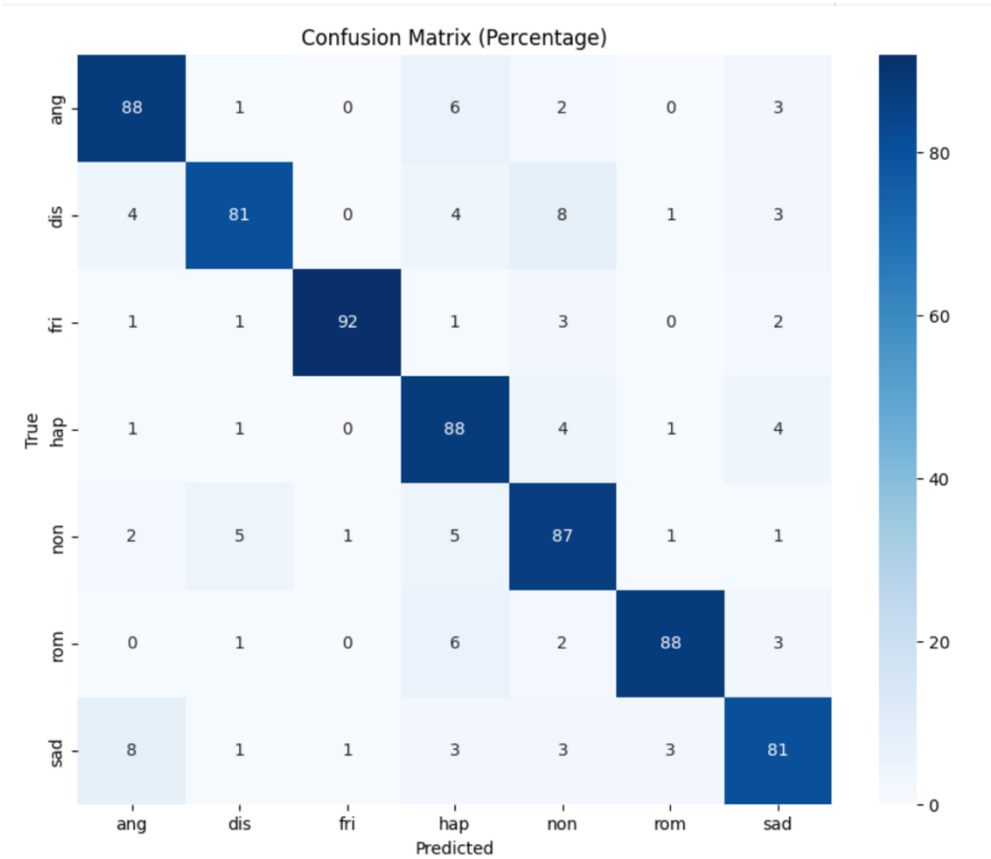
The assessment metrics provide valuable insights into the strengths and limitations of the emotion classification model across several emotion categories. These insights can guide prospective enhancements and optimizations to improve the model's performance. The confusion matrix of a CNN (Convolutional Neural Network) classification model offers a comprehensive analysis of the model's predictions for various classes in comparison to the true ground truth. The matrix is structured such that each row corresponds to the true class, and each column corresponds to the anticipated class. The confusion matrix results of CNN are depicted in Figure 41.

Evaluation:

1. Diagonal Elements (True Positives): The values on the diagonal (from the top-left to the bottom-right) indicate the count of cases where the model accurately predicted the emotion. Greater values on the diagonal imply superior performance. The chart clearly shows that Anger

has the greatest prediction rate among all the emotions, while Sadness has the lowest prediction rate among the emotions being studied.

Figure 41
The Confusion Matrix for The CNN Model



2. Off-diagonal elements (false positives and false negatives): Off-diagonal values represent cases when the model produced inaccurate predictions. In the specific cell labelled "True: Anger, predicted: sadness," there were a total of 11 occurrences where the model incorrectly predicted sadness instead of the genuine emotion, which was anger. This number shows the maximum incorrectness.

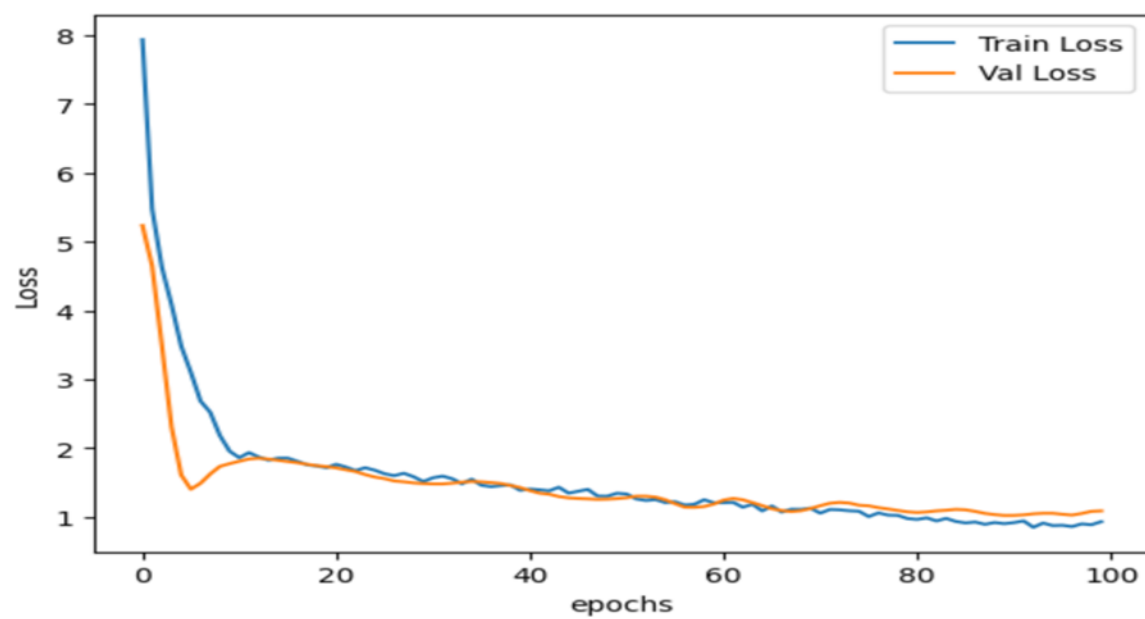
3. General Trends: The confusion matrix offers insights into the frequent misunderstanding between different emotion groups. For example, if there are numerous non-diagonal features between the emotions of Anger and Sadness, it indicates that the model may have difficulty differentiating between these two emotions.

Figure 42 is a flowchart that demonstrates the training loss and validation loss of CNN model. At the start of the model training, there is little difference in the loss values between the training and validation sets. As the epochs advance, the loss values gradually come together, showing minor disparities, until the training and validation loss reach their lowest point. A progressive and uniform alteration in the model's performance during the training phase leads to a seamless learning curve. This scenario provides evidence that this model is the most suitable model, as the desire for a smooth curve is supported by the signal that the model is consistently improving over time.

Figure 43 displays a comparison between the training accuracy and the validation accuracy. Significantly, there is a subtle disparity in values at the beginning stages, but as the training advances, the values consistently rise. The model suggests that the optimal model is achieved when the training and validation accuracy reaches a stable state and preferably converge to high values.

Figure 42

The Training and Validation Losses of CNN Model

**Figure 43**

The Training and Validation Accuracy of CNN Model

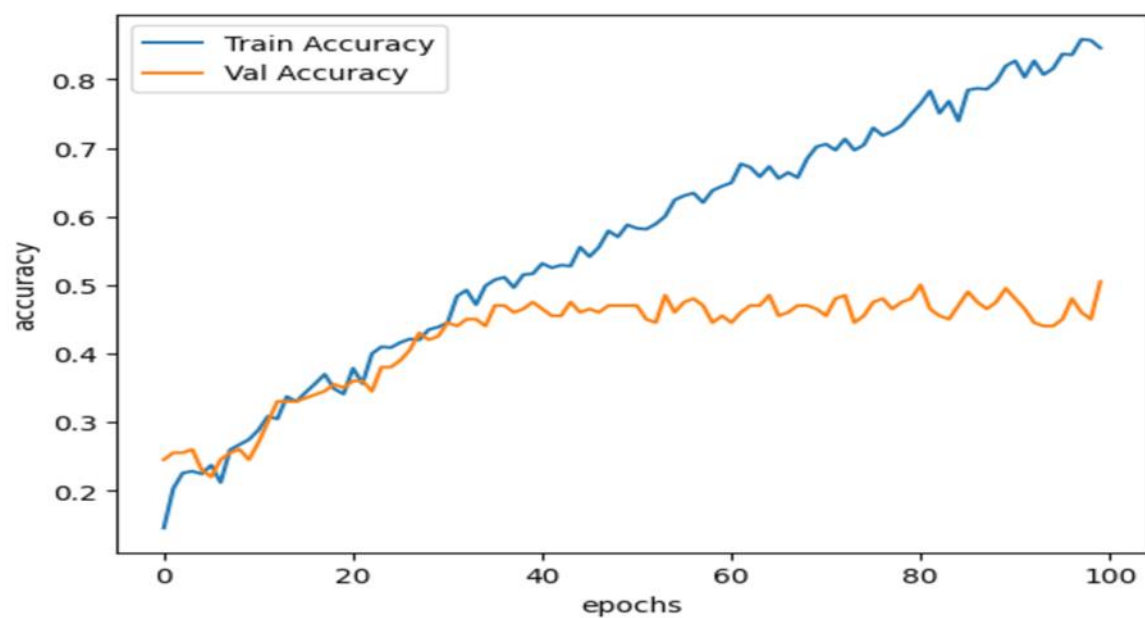
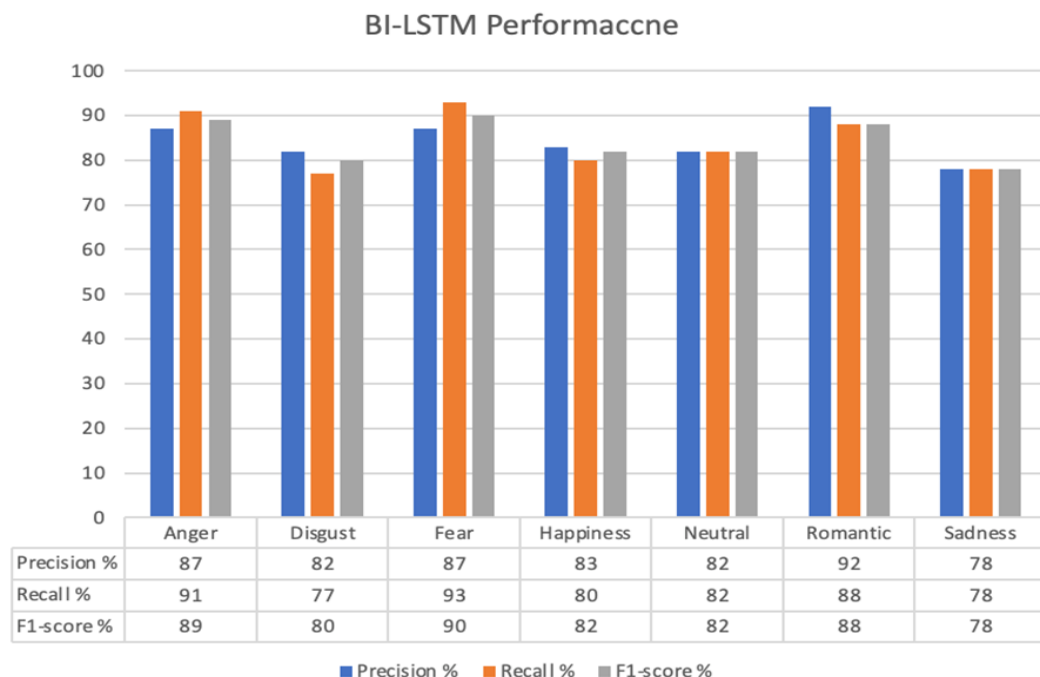


Table 19 and Figure 45 display the outcomes derived from the BI-LSTM model.

Table 19

The Performance of BI-LSTM Model

Emotions	Precision %	Recall %	F1-score %
Anger	87	91	89
Disgust	82	77	80
Fear	87	93	90
Happiness	83	80	82
Neutral	82	82	82
Romantic	92	88	88
Sadness	78	78	78

Figure 44*The Performance of BI-LSTM Model*

The table and graphic display the assessment metrics for various emotions in a categorization job. Below is an analysis of the outcomes:

1. Anger: - Accuracy: 87%

- The recall rate is 91%. - The F1-score is 89%.

The model has excellent performance in accurately detecting instances of rage, exhibiting a high level of accuracy, recall, and F1-score. This suggests that the model's predictions of anger are very accurate and successfully identify a significant number of real episodes of rage.

2. Disgust: - Accuracy: 82%

- The recall rate is 77%. - The F1-score is 80%.

The model has satisfactory performance in accurately identifying distaste, demonstrating a well-balanced accuracy and recall. It accurately recognizes a significant fraction of genuine episodes of disgust while retaining a satisfactory degree of accuracy.

3. Fear: - Accuracy: 87%

The recall rate is 93%.

The F1-score is 90%.

The model has exceptional proficiency in recognizing fear, with elevated levels of accuracy, recall, and F1-score. It effectively categorizes episodes of fear and appropriately captures a substantial fraction of real fear situations.

4. Happiness: - Accuracy: 83% - Completeness: 80%

The F1-score is 82%.

The model has strong performance in accurately detecting happiness, achieving a harmonious balance between precision and recall. It accurately predicts occurrences of happiness while retaining a satisfactory degree of accuracy.

5. Neutral: - Precision: 82% - Recall: 82% - F1-score: 82%

The model consistently demonstrates accurate recognition of neutral emotions, achieving a balanced level of precision and recall. It effectively distinguishes neutral cases while retaining a satisfactory degree of accuracy.

6. Romantic genre: - Precision: 92% - Recall: 88% - F1-score: 88%

The model has exceptional proficiency in identifying romantic feelings, exhibiting high levels of accuracy, recall, and F1-score. It effectively categorizes romantic occurrences and appropriately represents a substantial fraction of real romantic occurrences.

7. Emotion of sorrow or unhappiness:

The level of precision is 78%.

- Retention rate: 78%

The F1-score is 78%.

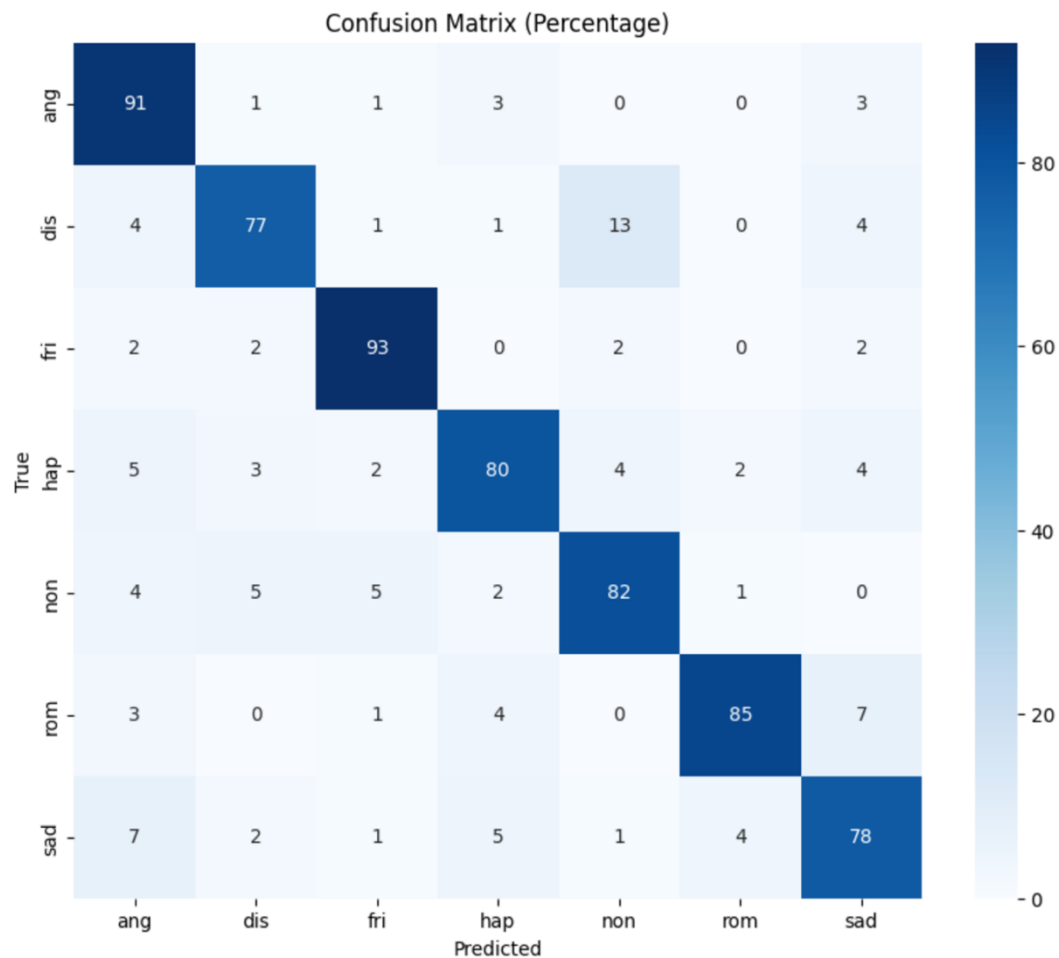
The model has satisfactory performance in accurately detecting sorrow, exhibiting a balanced level of precision and recall. It accurately forecasts occurrences of melancholy while maintaining a satisfactory level of accuracy.

To summarize, the model exhibits excellent performance across several emotions, excelling particularly in the fear and love categories. Nevertheless, there is a degree of variation in the level of performance observed for different emotions, as shown by the values of accuracy, recall, and F1-score.

The confusion matrix of a Bidirectional Long Short-Term Memory (BI-LSTM) classification model offers a comprehensive analysis of the model's predictions across many classes in comparison to the true values. The matrix's rows correspond to the actual class, while its columns correspond to the anticipated class. The confusion matrix findings of BI-LSTM are displayed in Figure 45.

The presented confusion matrix depicts the model's performance in evaluating several classes (emotions). The genuine class is represented by each row, whereas the anticipated class is represented by each column. Here is the analysis:

- First row (Anger):

Figure 45*The Confusion Matrix for BI-LSTM*

- Anger was accurately predicted in 91% of occasions.
- Incorrectly categorized as Disgust: 2% of the incidents
- Incorrectly categorized as Fear: 2% of occasions
- Incorrectly categorized as Happiness: 3% of the cases
- Incorrectly categorized as Neutral: 0% of instances
- Incorrectly categorized as Romantic: 0% occurrence
- Incorrectly categorized as Sadness: 3% of the cases

- Second row (Disgust):
 - Incorrectly categorized as Anger: 4% of the occurrences
 - Disgust was accurately predicted in 77% of occasions.
 - Incorrectly categorized as Fear: 1% occurrence
 - Incorrectly categorized as Happiness: 1% occurrence
 - Incorrectly categorized as Neutral: 13% of the cases
 - No occurrences were predicted as Romantic.
 - Incorrectly categorized as Sadness: 4% of the cases
- Third row (Fear):
 - Incorrectly categorized as Anger: 2% of the cases
 - Incorrectly categorized as Disgust: 2% of the occurrences
 - Fear was accurately predicted in 93 occasions.
 - No incidents were misclassified as Happiness.
 - Incorrectly categorized as Neutral: 2% of the cases
 - Instances misclassified as Romantic: 0%
 - Incorrectly categorized as Sadness: 2% of occurrences
- Fourth row (Happiness):
 - Incorrectly categorized as Anger: 5% of the cases
 - Incorrectly categorized as Disgust: 3% of the occurrences
 - Incorrectly categorized as Fear: 2% of the occasions
 - Accurately forecasted as Happiness in 80% of cases
 - Incorrectly categorized as Neutral: 4% of the cases
 - Incorrectly categorized as Romantic: 2% of cases

- Incorrectly categorized as Sadness: 4% of the cases
- Fifth row (Neutral):
 - Incorrectly categorized as Anger: 4% of occurrences
 - Incorrectly categorized as Disgust: 5% of the cases
 - Incorrectly categorized as Fear: 5% of cases
 - Incorrectly categorized as Happiness: 2% of the cases
 - Accurately classified as Neutral: 82% of the cases
 - Incorrectly categorized as Romantic: 1% occurrence
 - No cases were predicted as Sadness: 0%
- Sixth row (Romantic):
 - Incorrectly categorized as Anger: 3% of the occasions
 - Disgust predicted: 0% of incidences
 - Incorrectly categorized as Fear: 1% of the occasions
 - Incorrectly categorized as Happiness: 4% of the cases
 - Incorrectly categorized as Neutral: 0% of cases
 - Accurately classified as Romantic: 85% of the cases
 - Incorrectly categorized as Sadness: 7% of the cases
- The seventh row represents sadness.
 - Incorrectly categorized as Anger: 7% of the cases
 - Incorrectly categorized as Disgust: 2% of occurrences
 - Incorrectly categorized as Fear: 1% of occasions
 - Incorrectly categorized as Happiness: 5% of cases
 - Incorrectly categorized as Neutral: 1% of the instances

- Incorrectly categorized as Romantic: 4% of cases
- Successfully classified as Sadness: 78% of the cases

Explanation: - The diagonal components indicate the count of accurately categorized instances for each emotion.

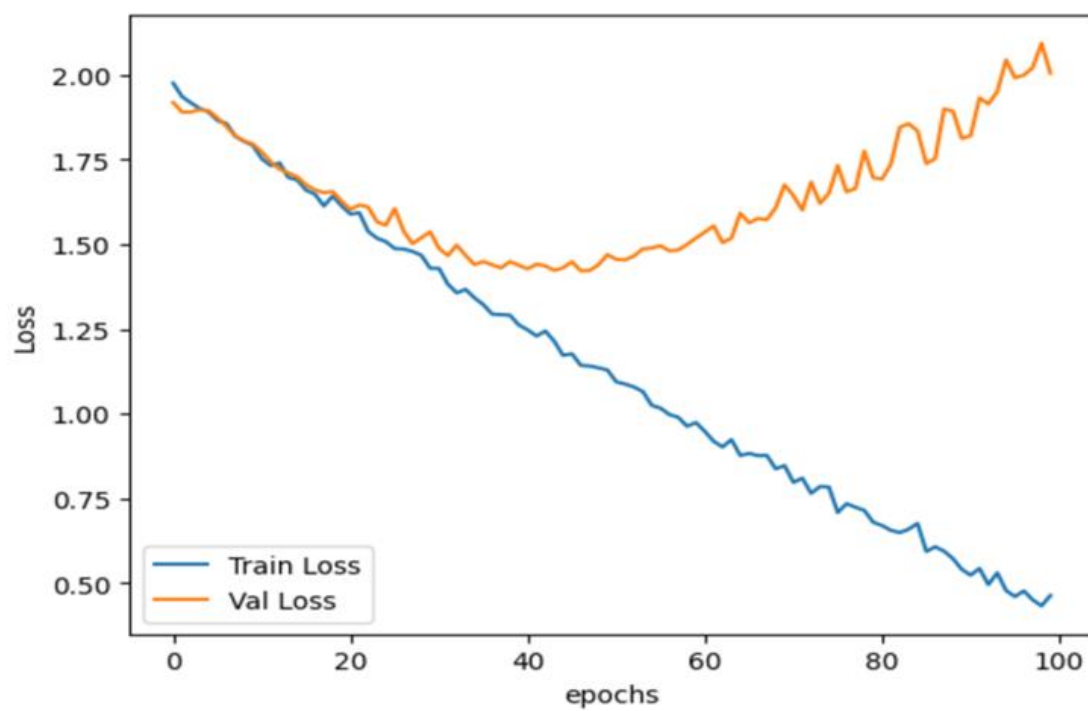
- Off-diagonal items represent instances of incorrect categorization.
- Emotions that are well-classified are indicated by the greatest values along the diagonal.

The user did not provide any text. Examining misclassifications offers valuable information into the areas where the model encounters difficulties.

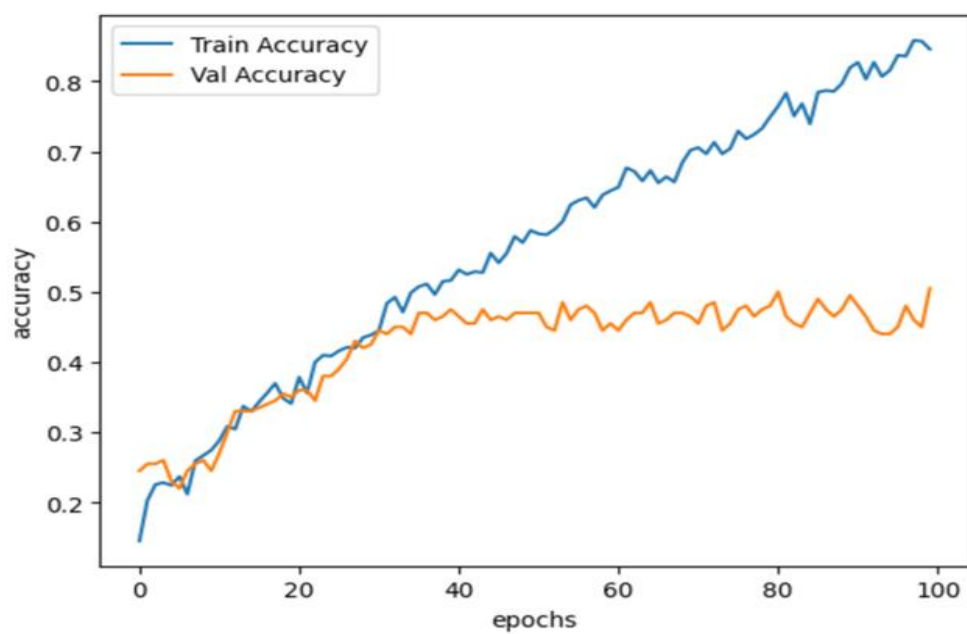
Figure 46 displays the training loss and validation loss, revealing significant disparities in epochs towards the latter stages. Based on the figure, it can be observed that the training loss is little but the validation loss is substantial, suggesting that the model is prone to overfitting. To improve performance in this scenario, it might be beneficial to decrease model complexity, implement regularization techniques, or increase the size of the training dataset. The flowcharts shown in Figures 47 demonstrate the training accuracy and validation accuracy in the BI-LSTM model. Significantly, there is a noticeable difference in the validation accuracy values at the beginning and end of the charts, although the values in the middle epochs are quite comparable.

Figure 46

The Training and Validation Losses In BI-LSTM

**Figure 47**

The Training and Validation Accuracy In BI-LSTM



The results obtained from the Artificial Neural Network (ANN) model are presented in Table 20 and Figure 48. Based on the data shown in the table and figure, it is evident that the Artificial Neural Network (ANN) model is not suitable for achieving high performance in speech recognition. The findings demonstrate that the model produces very few accurate speech predictions, with the majority of values being zero.

The table and figure presented display the assessment metrics, such as accuracy, recall, and F1-score, for various emotion kinds in a classification job. The following is an examination of the outcomes:

1. Anger: - Precision: 32% - Recall: 100%

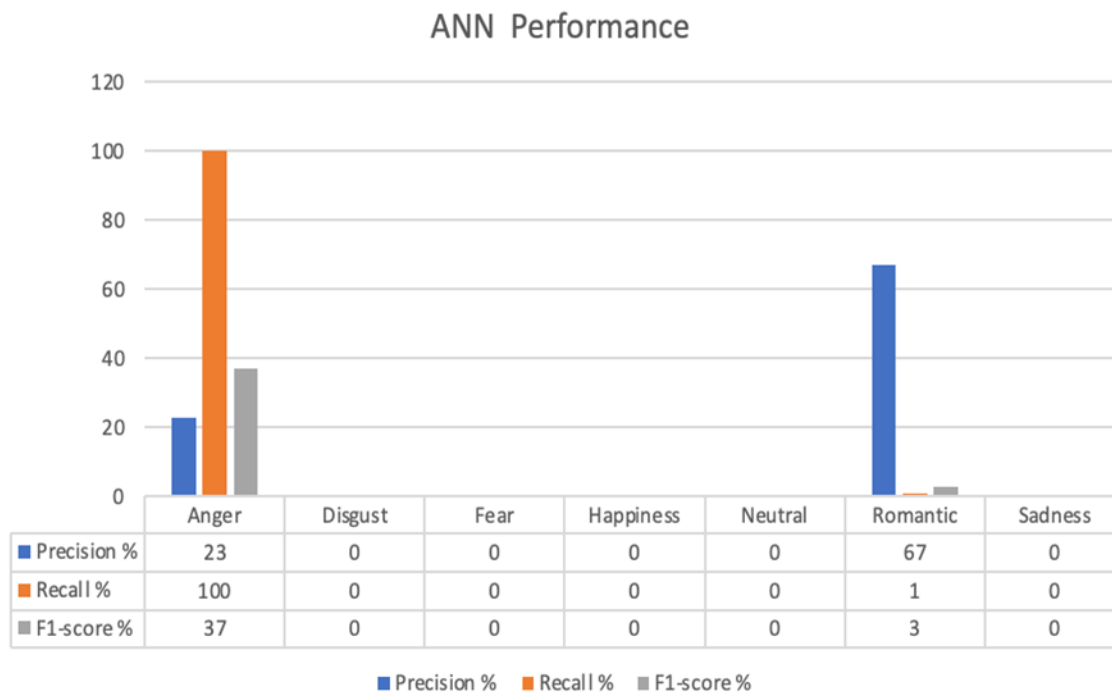
The F1-score is 49%.

Regarding rage, the model had a very low level of precision but attained 100% recall. This means that while the model made predictions of anger, it was accurate 32% of the time, and it successfully detected all instances of genuine anger.

Table 20

The Performance of ANN on Speech Recognition

Emotion Type	Precision %	Recall %	F1-score %
Anger	32	100	49
Disgust	00	00	00
Fear	00	00	00
Happiness	00	00	00
Neutral	00	00	00
Romantic	73	18	29
Sadness	00	00	00

Figure 48*ANN Performance*

2. The emotions listed are Disgust, Fear, Happiness, Neutral, and Sadness. The precision level for these emotions is 0%.

- No recollection: 0%

The F1-score is 0%.

The model has not made any predictions for the emotions Disgust, Fear, Happiness, Neutral, and Sadness. As a result, the precision, recall, and F1-score for these emotions are all 0. This implies that the model's performance in identifying certain emotions may have been subpar, or that there might be problems with the dataset or the model's structure.

3. Romantic: - Accuracy: 73% - Sensitivity: 18% - Harmonic mean: 29%

Regarding the emotion of Romantic, the level of accuracy is considerably greater compared to other emotions. This means that when the model made a prediction of Romantic, it

was right 73% of the time. Nevertheless, the model's recall rate is poor, indicating that it failed to identify a substantial percentage of real occurrences of Romantic feelings.

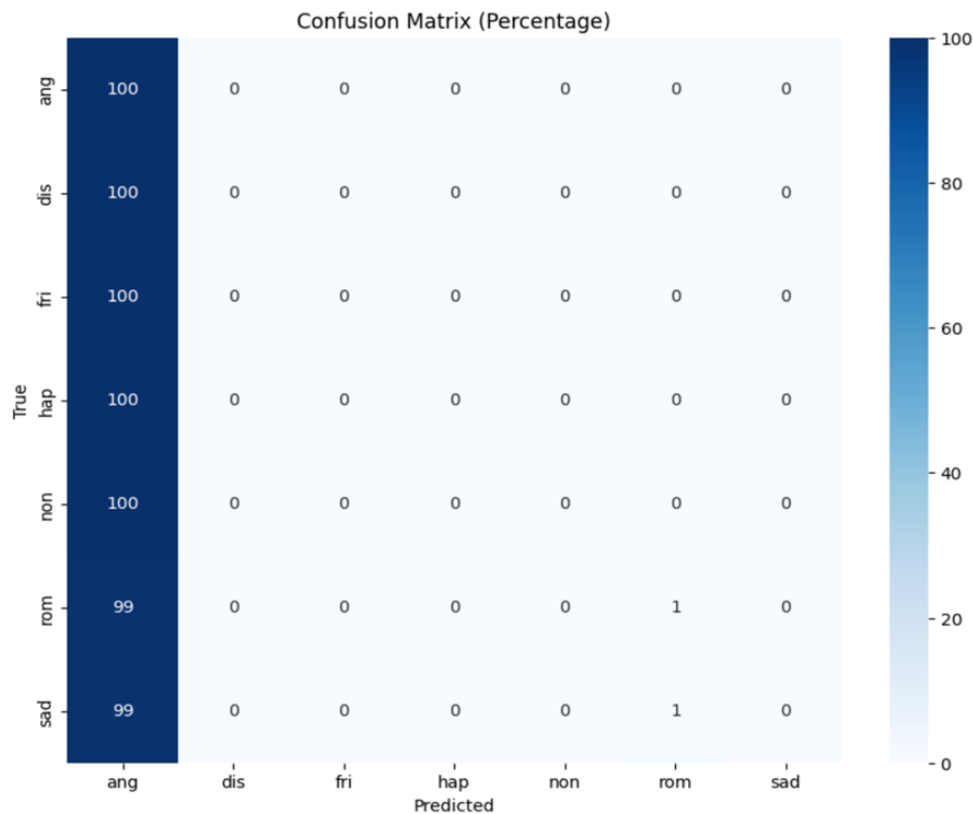
Overall, the results suggest a possible problem with the model's capacity to apply across many emotion categories, particularly for Disgust, Fear, Happiness, Neutral, and Sadness. To enhance the model's performance in all emotion categories, it is advisable to do a thorough examination of the dataset, model architecture, and any biases.

The confusion matrix shown in Figure 49 illustrates the predictions made by an Artificial Neural Network (ANN) classification model. It presents a comprehensive analysis of the model's predictions for various classes, in contrast to the actual ground truth. The rows in the table correspond to the true class, while the columns indicate the anticipated class. The confusion matrix visually illustrates that the artificial neural network (ANN) is not suitable for accurately recognizing emotions in Arabic speech. Significantly, only the "Anger" emotion is accurately predicted, whilst the forecasts for other emotions are either inaccurate or nonexistent.

The model's learning curve, as depicted in the given figure, suggests a restricted level of performance, particularly in regards to accuracy, recall, and F1-score across various classes. The model's inability to reliably categorize examples of each class is evident from the low values in these measures. To be more precise:

Precision: The model has a limited capacity to accurately detect positive occurrences (true positives) across most classes, except for class 5 (index 5) which achieves a significantly better precision of 0.67.

- The model's ability to correctly identify positive cases is typically weak, as seen by the low recall scores for all classes. Class 0 (index 0) has a recall of 1.00, indicating superior ability in accurately identifying positive cases for this class.

Figure 49*ANN Confusion Matrix*

- The F1-score, which accounts for both accuracy and recall, is minimal for most courses, indicating the general difficulties in reaching a balanced trade-off between precision and memory.

- Precision, recall, and F1-score values are all poor, which is consistent with the total accuracy of the model being 0.23. This implies that the model's forecasts are typically imprecise.

Figure 50 illustrates the fluctuating losses seen between the training and validation sets during the early epochs, in contrast to the later epochs where the values converge. Figure 51 demonstrates the difference in the accuracy of training and validation in the Artificial Neural Network (ANN) model. Based on the data, the learning curve of this model demonstrates

difficulties in correctly categorizing cases across different classes, suggesting the necessity for enhancement and maybe further adjustments or reassessment of the model's structure.

Figure 50

The Losses in The Train and Validation Using ANN Model.

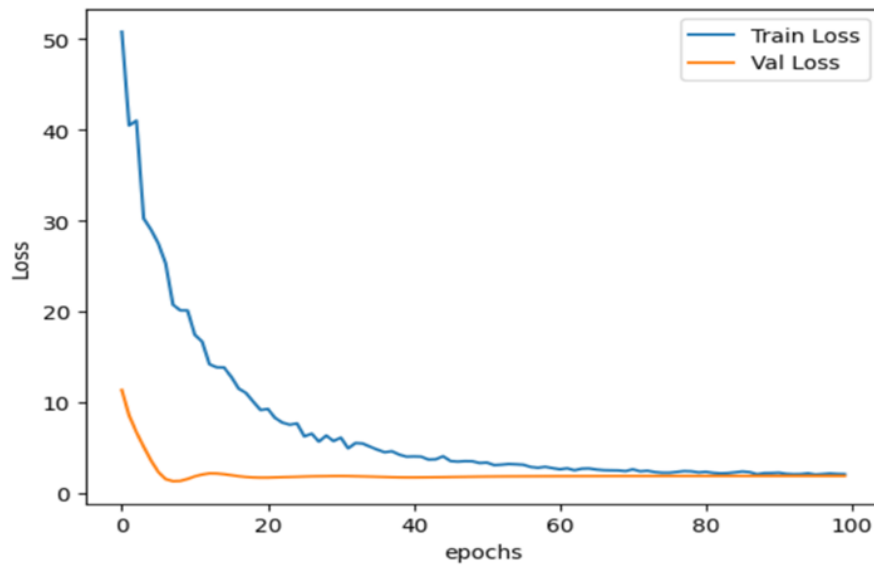


Figure 51

The Accuracy in the Train and Validation Using ANN Model.

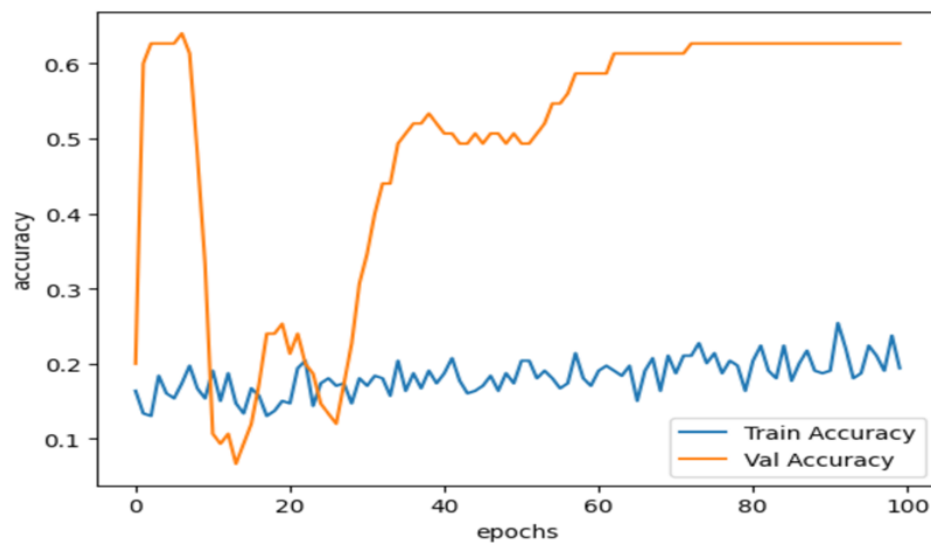


Table 21 and Figure 52 demonstrate the results of GRU in the domain of Arabic speech emotion recognition.

Table 21

The Performance of GRU Model

Emotion Type	Precision %	Recall %	F1-score %
Anger	70	84	76
Disgust	62	57	59
Fear	90	88	89
Happiness	66	67	62
Neutral	69	67	68
Romantic	83	67	74
Sadness	61	69	65

Figure 52

GRU Performance



The data shown in the table and accompanying figure disclose the subsequent findings:

1. Anger: The precision is greater than the recall, suggesting the presence of some false positives. The F1-score is equitable.

2. The model faces challenges in recalling instances of disgust, suggesting problems in accurately recording all occurrences of this emotion. The precision and F1-score exhibit rather low values.

3. Fear: Achieving a high F1-score requires a well-balanced combination of accuracy and recall. The model adeptly depicts terror.

4. Happiness: Decreased recall indicates that the model fails to identify some occurrences of happiness. The precision and F1-score exhibit a modest level.

5. Neutral: Equally high accuracy and recall result in a modest F1-score. The model has satisfactory performance when it comes to neutral feelings.

6. In the context of romance, having a high level of accuracy means there are fewer instances where something is mistakenly identified as a favorable outcome when it is not. The recall and F1-score reflect a state of equilibrium.

7. The combination of accurate accuracy and memory leads to a moderate F1-score, indicating a level of sadness. The model has satisfactory performance in detecting sorrow.

General Evaluation:

The GRU model's performance measures, such as accuracy, recall, and F1-score, for different emotions are presented in Table 21 and Figure 51. The GRU model exhibits diverse performance across several emotions. It has exceptional proficiency in expressing feelings of terror and romance but encounters difficulties in portraying sentiments of disgust and contentment. Ensuring a balance between accuracy and recall is essential for attaining high F1 scores, and the model demonstrates different levels of success in striking this balance across different emotions. Comprehending these subtle distinctions helps direct further optimization endeavours to improve the model's effectiveness in certain emotional situations.

The table and graphic offer the following insights:

1. Anger: - Accuracy: 70%

The recall rate is 84%.

The F1-score is 76%.

The model has high accuracy and recall in recognizing episodes of rage. The F1-score, which achieves a balance between accuracy and recall, is likewise pretty high.

2. Disgust: - Accuracy: 62%

- Memorandum: 57%

The F1-score is 59%.

- The model demonstrates a satisfactory level of accuracy in recognizing occurrences of disgust, although it has difficulties in correctly identifying all cases of disgust. The F1-score indicates a satisfactory equilibrium between accuracy and recall.

3. Fear: - Accuracy: 90%

- Retention rate: 88%

The F1-score is 89%.

The model has outstanding performance in accurately identifying occurrences of fear, with high accuracy, recall, and F1-score.

4. Happiness: - Accuracy: 66%

- Memorandum: 67%

The F1-score is 62%.

The model exhibits a respectable level of accuracy and recall for happiness, and the F1-score implies a decent equilibrium between the two.

5. Neutral sentiment classification had a precision rate of 69% and a recall rate of 67%.

The F1-score is 68%.

The model consistently demonstrates accurate identification of neutral emotions, with a well-balanced accuracy, recall, and F1-score.

6. Romantic genre: - Precision: 83% - Recall: 67% - F1-score: 74%

The model has a high level of accuracy in detecting romantic feelings, but its ability to identify all instances of these emotions is somewhat lower. The F1-score demonstrates a satisfactory equilibrium.

7. Sadness: - Accuracy: 61%

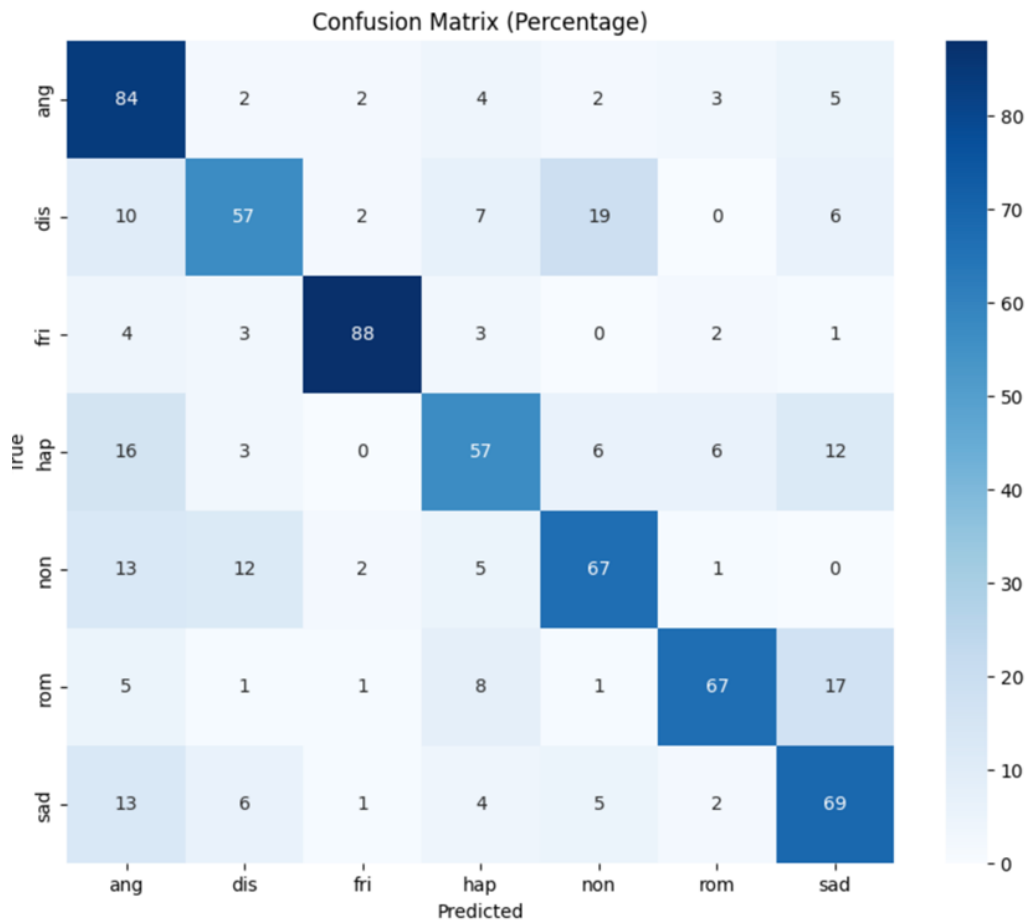
- Remember: 69%

The F1-score is 65%.

The model has a modest level of accuracy and recall for sorrow, as indicated by the F1-score, which indicates a reasonable balance between the two.

To summarize, the model demonstrates proficiency in recognizing specific emotions like fear, however its effectiveness fluctuates when it comes to other emotions. The trade-off between precision and recall is clearly apparent, highlighting the significance of adopting a balanced assessment methodology. Additional optimization and fine-tuning might be contemplated to improve the overall performance of the model.

Figure 53 displays a confusion matrix that provides valuable information about the performance of a GRU model in several emotion categories. Below is the analysis of the components of the matrix:

Figure 53*GRU Confusion Matrix*

- Dimensions:

- Rows correspond to the real classes or ground truth.

The columns correspond to the classes predicted by the model.

- Diagonal Elements (True Positives): - The diagonal elements indicate the count of cases when the model accurately predicted the class. Greater values on the diagonal imply excellent performance.

- Off-diagonal items, namely false positives and false negatives, indicate instances where misclassifications have occurred.

- The values in row and column represent the count of cases belonging to class that were predicted as class.

- Analysis:

- Class 0 (Anger) has 84 true positives, suggesting good performance in predicting anger. However, there are some misclassifications into other classes.

- Class 4 (Neutral) shows notable misclassifications, especially with class 1 (Disgust) and class 5 (Romantic).

- Class 5 (Romantic) has a relatively high number of misclassifications, especially with class 6 (Sadness).

- Overall Performance:

- The diagonal elements' higher values indicate decent overall accuracy.

- The model performs well for certain classes but struggles with others, as evident from the misclassifications.

- Recommendations:

- Further analysis and potential model tuning could focus on improving performance for specific classes with higher misclassification rates.

- Exploring feature importance and model interpretability techniques may provide additional insights.

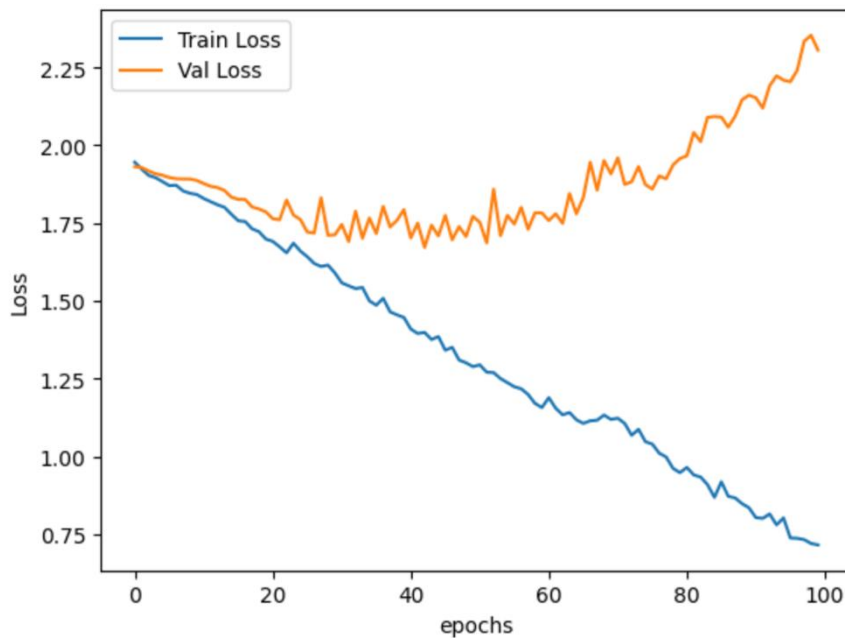
In conclusion, the confusion matrix aids in understanding the strengths and weaknesses of the model across different emotion classes, guiding further refinement and optimization efforts.

The GRU model's training and validation losses are shown in Figure 54. The training losses exhibit a downward trend as the number of epochs grows, whereas the validation losses

demonstrate an upward trend throughout the course of the epochs. This suggests that the model is experiencing overfitting, which is characterized by a scenario in which the training loss is low but the validation loss is large.

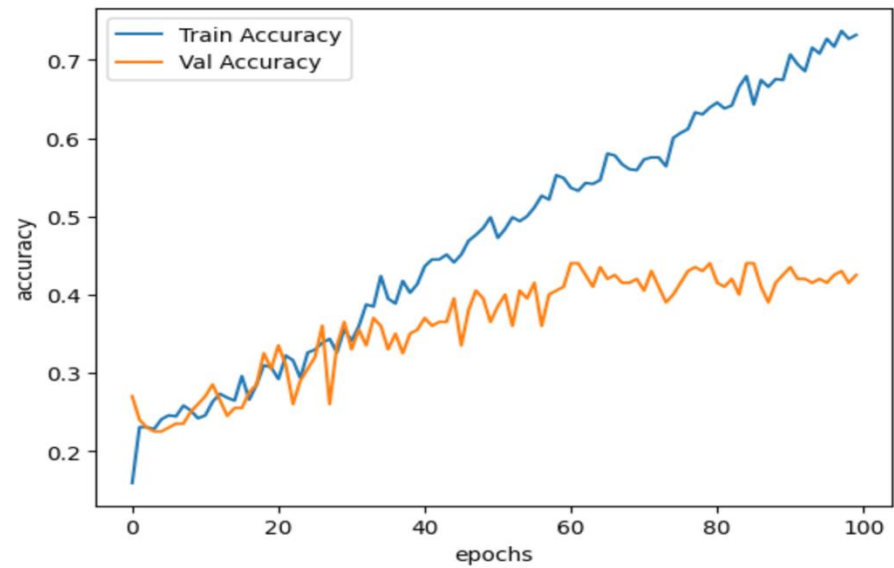
Figure 54

Training and Validation Losses in GRU



The GRU model's training accuracy and validation accuracy data are shown in Figure 55. The training accuracy has a positive correlation with the number of epochs, but the validation accuracy demonstrates a negative correlation with an increasing number of epochs. Therefore, this suggests that the model is exhibiting signs of overfitting.

Figure 55
Training and Validation Accuracy in GRU

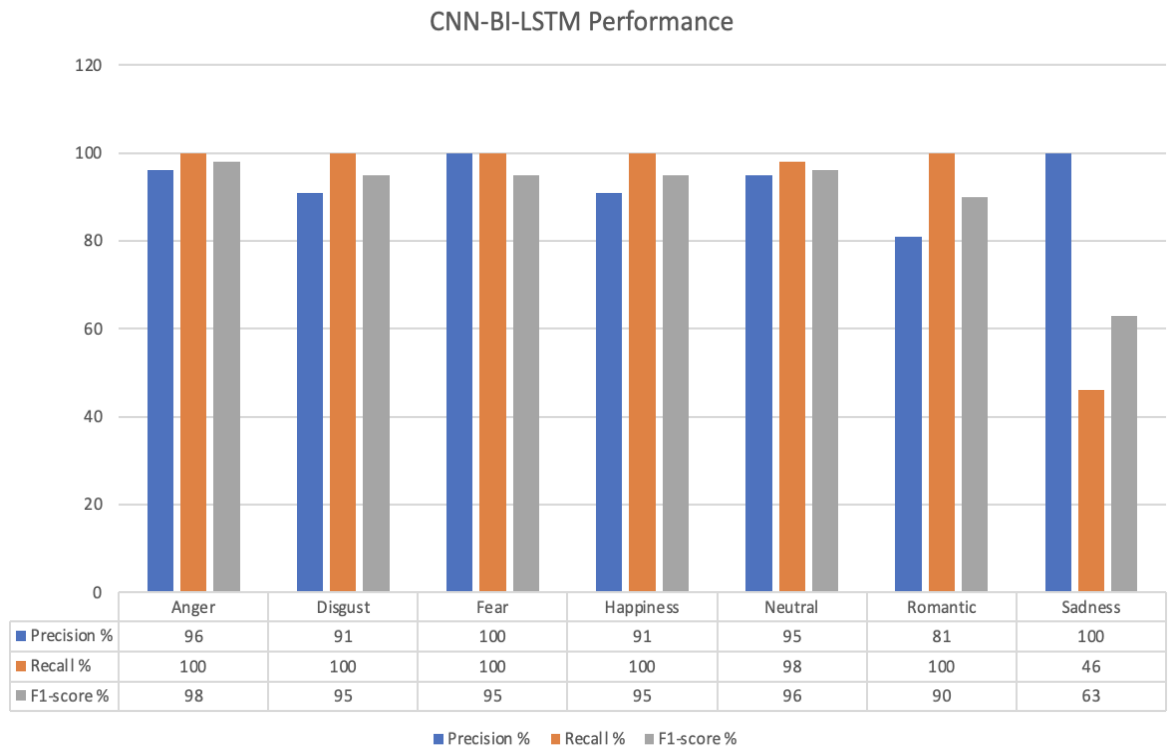


The performance Table 22 and Figure 56 below provide a detailed summary of the CNN-BI-LSTM model's efficacy in distinguishing various emotions.

Table 22
The Performance of CNN-BI-LSTM Hybrid Model

Emotion Type	Precision %	Recall %	F1-score %
Anger	96	100	98
Disgust	91	100	95
Fear	100	100	95
Happiness	91	100	95
Neutral	95	98	96
Romantic	81	100	90
Sadness	100	46	63

Figure 56
The Performance of CNN-BI-LSTM Model



The model demonstrated exceptional proficiency in detecting fear, with an accuracy and recall rate of 100%. The F1-score, which stands at 95%, demonstrates the model's strong ability to reliably identify fear.

- Happiness: The model demonstrated a precision of 91% in detecting happiness, demonstrating a high level of accuracy in properly recognizing this emotion. The recall and F1-score both get a perfect score of 100%, highlighting the model's exceptional ability to accurately identify all occurrences of pleasure without any instances of false negatives.

The model demonstrated an excellent precision of 96%, demonstrating a high level of accuracy in properly recognizing occurrences of fury. The recall and F1-score both indicate a perfect performance of 100%, demonstrating the model's capability to accurately identify all occurrences of rage without any false negatives.

- The model demonstrated strong performance in detecting disgust, with a precision of 91%, recall of 100%, and an F1-score of 95%. It exhibited exceptional precision and efficacy in identifying occurrences of distaste.

- Neutral: The model exhibited strong performance in accurately identifying neutral emotions, with an accuracy rate of 95%, recall rate of 98%, and an F1-score of 96%. It exhibited exceptional precision and efficacy in detecting neutral emotional states.

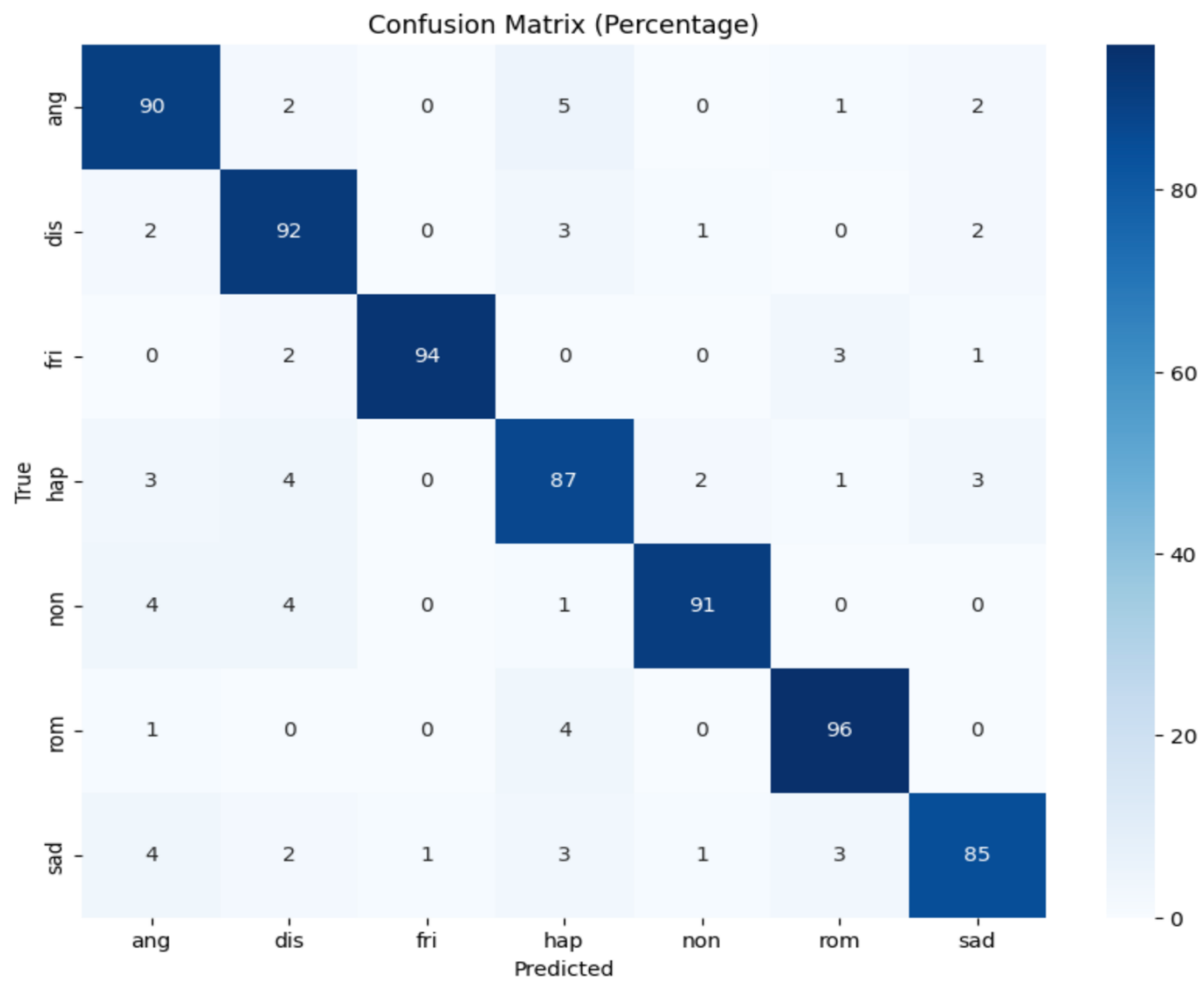
- Romantic: The precision for romantic emotions is 81%, suggesting a somewhat lower level of accuracy in comparison to other emotions. Nevertheless, the recall and F1-score exhibit a perfect score of 100%, indicating that the model's predictions of romantic feelings are remarkably precise. However, it is worth noting that there is a possibility of some occurrences being overlooked.

- Regarding sorrow, the model's accuracy in recognizing it is 100%, but its recall is just 46%, leading to a lower F1-score of 63%. This suggests that the model has a high level of accuracy in predicting melancholy, although it may fail to identify a considerable proportion of real occurrences.

Overall, the CNN-BI-LSTM model shown impressive proficiency in detecting a range of emotions, with particularly remarkable precision in identifying anger, contempt, fear, happiness, and neutral emotions. Nevertheless, it exhibited certain constraints in accurately detecting occurrences of amorous and sad emotions, as indicated by lower recall scores in these categories.

The confusion matrix depicted in figure 577 below represents the performance of a CNN-BI-LSTM hybrid model. It provides a comprehensive breakdown of the projected and actual classifications for different classes.

Analysis:

Figure 57*The Performance of CNN-BI-LSTM Model*

Anger (Class 0): The model accurately identified 90 cases of anger, with 2 incorrect positive predictions, 5 incorrect negative predictions, and 2 instances properly categorized as different emotions. The angry performance exhibits a robust overall performance.

- Disgust (Class 1): The model correctly predicted 92 occurrences of disgust, but made 2 incorrect positive predictions, 3 incorrect negative predictions, and misclassified 2 cases as different emotions. The distaste category demonstrates exceptional overall performance.

- Fear (Class 2): The model accurately identified 94 occurrences of fear, with 2 incorrect positive predictions, 1 incorrect negative prediction, and 4 cases mistakenly categorized as different emotions. The total performance in terms of fear is quite commendable.

- Happiness (Class 3): The model achieved a precision of 87 instances in successfully predicting happiness, with 3 cases falsely identified as positive and 2 instances falsely identified as negative. Additionally, there were 7 examples that were incorrectly categorized as different emotions. The overall happiness performance is robust.

- Neutral (Class 4): The model accurately identified 91 occurrences of neutral emotions, with 4 incorrect positive predictions, 1 incorrect negative prediction, and 0 cases categorized as other emotions. The overall performance for emotions categorized as neutral is exceptional.

- Romantic (Class 5): The model accurately predicted 96 cases of romantic feelings, with only 1 incorrect positive prediction, 4 incorrect negative predictions, and no instances misclassified as different emotions. The overall performance in eliciting romantic emotions is exceptional.

- Sadness (Class 6): The model achieved a precision of 85% in correctly identifying occurrences of sadness, with 3 cases incorrectly categorized as sadness when they were not, 1 instance incorrectly classed as not sadness when it was, and 11 instances misclassified as different emotions. The general efficacy of detecting sorrow is satisfactory, however there is a slightly greater incidence of false positive results.

To summarize, the confusion matrix demonstrates a robust overall performance of the model, exhibiting a high level of accuracy in predicting the majority of classes. Nevertheless, it is crucial to focus on mitigating the occurrence of both false positives and false negatives within particular emotional categories, such as melancholy and rage.

Figure 58 displays the training and validation losses for the CNN-BI-LSTM model. The training losses decrease as the number of epochs increases, whereas the validation losses grow with the progression of the epochs. This indicates that the model is overfitting, which occurs when a machine learning model demonstrates a situation where the training loss is modest while the validation loss is high.

Figure 58

The Training and Validation Losses in CNN-BI-LSTM Hybrid Model

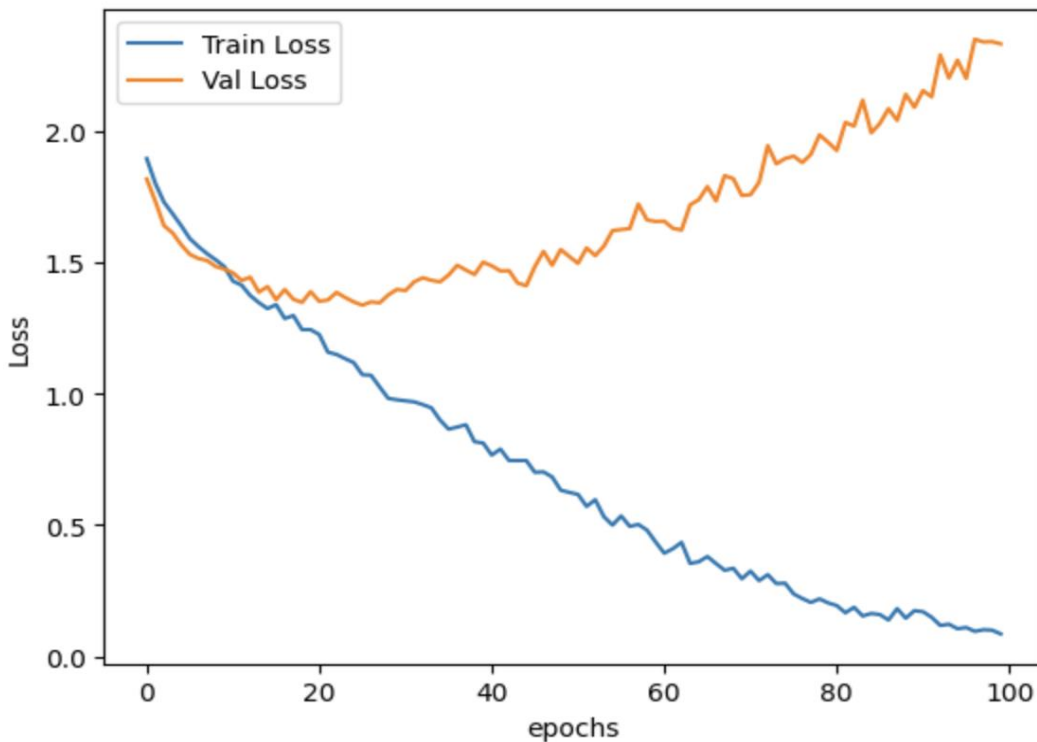


Figure 59

displays the training accuracy and validation accuracy metrics for the CNN-BI-LSTM model. The training accuracy increases as the number of epochs increases, but the validation accuracy decreases with an increasing number of epochs. Hence, this implies that the model is displaying indications of overfitting.

Figure 59

The Training and Validation Accuracy In CNN-BI-LSTM Hybrid Model

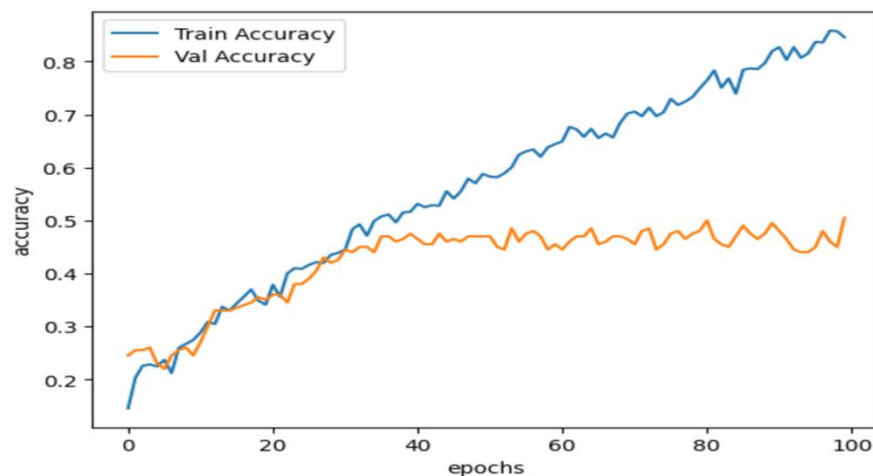


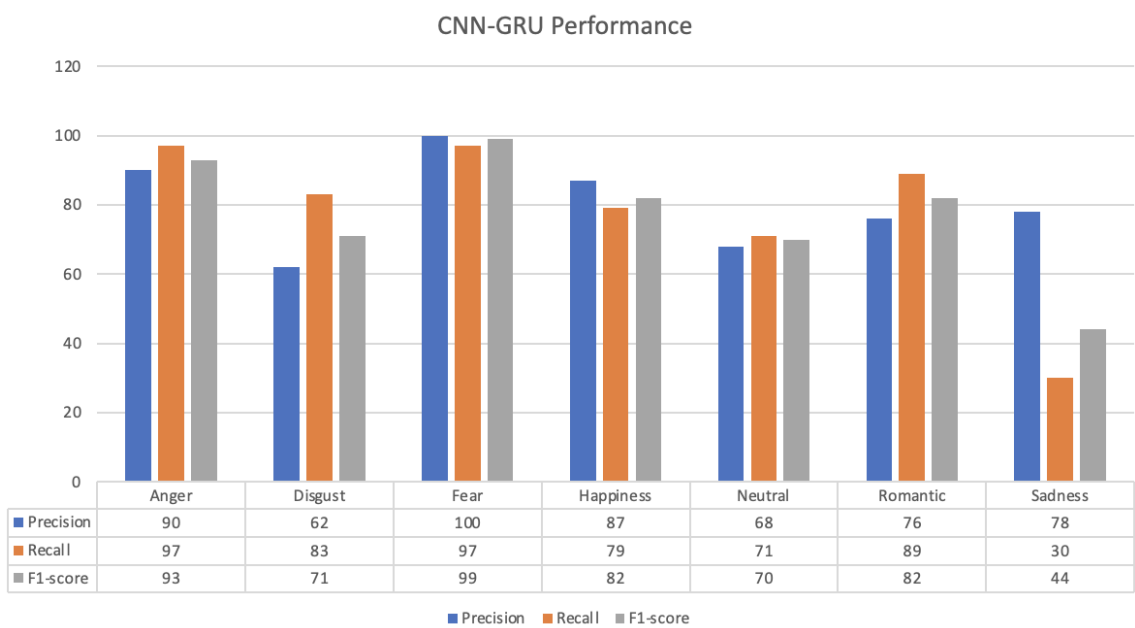
Table 23 and Figure 62 depict the performance of the CNN-GRU hybrid model in recognizing emotions in Arabic speech.

Table 23

The CNN-GRU Hybrid Model

Emotion Type	Precision	Recall	F1-score
Anger	90	97	93
Disgust	62	83	71
Fear	100	97	99
Happiness	87	79	82
Neutral	68	71	70
Romantic	76	89	82
Sadness	78	30	44

Figure 60
The CNN-GRU Performance



The table displays the performance metrics for various emotions in a categorization model. Below is an analysis and examination of the outcomes:

1. Rage:

- The level of precision is 90%.
- Retention rate: 97%

The F1-score is 93%.

A high level of accuracy means that the model accurately predicts rage 90% of the time. The 97% recall rate indicates that the model successfully detects the majority of occasions where anger is present. The F1-score, which accounts for both accuracy and recall, stands at an impressive 93%.

2. Disgust: - Accuracy: 62%

- Remember: 83%

The F1-score is 71%.

The model's accuracy in detecting disgust is relatively poor, with a precision rate of only 62%, meaning that it correctly predicts disgust in only 62% of cases. Nevertheless, the recall rate is at 83%, indicating that the algorithm successfully identifies a substantial proportion of genuine experiences of disgust. The F1-score, which is a measure that balances accuracy and recall, is 71%.

3. Fear: - Accuracy: 100%

The recall rate is 97%.

The F1-score is 99%.

The model's exceptional precision reveals its 100% accuracy in forecasting fear when it generates a positive prediction. The 97% recall rate indicates that the algorithm successfully identifies nearly all occurrences of genuine terror. The F1-score is exceptional, reaching a remarkable 99%.

4. Happiness: - Accuracy: 87%

- Retention rate: 79%

The F1-score is 82%.

The model demonstrates a commendable equilibrium between accuracy and recall in predicting pleasure. The precision of 87% signifies the level of accuracy in correctly predicting instances of happiness, while the recall of 79% reflects the model's capability to identify and record real occurrences of pleasure. The F1-score is a satisfactory 82%.

5. Objective:

The level of precision is 68%.

- Retention rate: 71%

The F1-score is 70%.

The model demonstrates a satisfactory level of accuracy and memory when it comes to neutral feelings. The 68% accuracy indicates that the model's positive predictions for neutral emotions are very accurate. Similarly, the 71% recall demonstrates that the model successfully catches a significant proportion of the actual instances of neutral emotions. The F1-score is 70%.

6. Romantic: - Accuracy: 76%

- Retention rate: 89%

The F1-score is 82%.

The model has a well-balanced performance in accurately detecting romantic feelings, with an accuracy rate of 76% and a recall rate of 89%. The F1-score is at 82%, signifying a commendable overall performance.

7. Sadness: - Accuracy: 78%

- Retention rate: 30%

The F1-score is 44%.

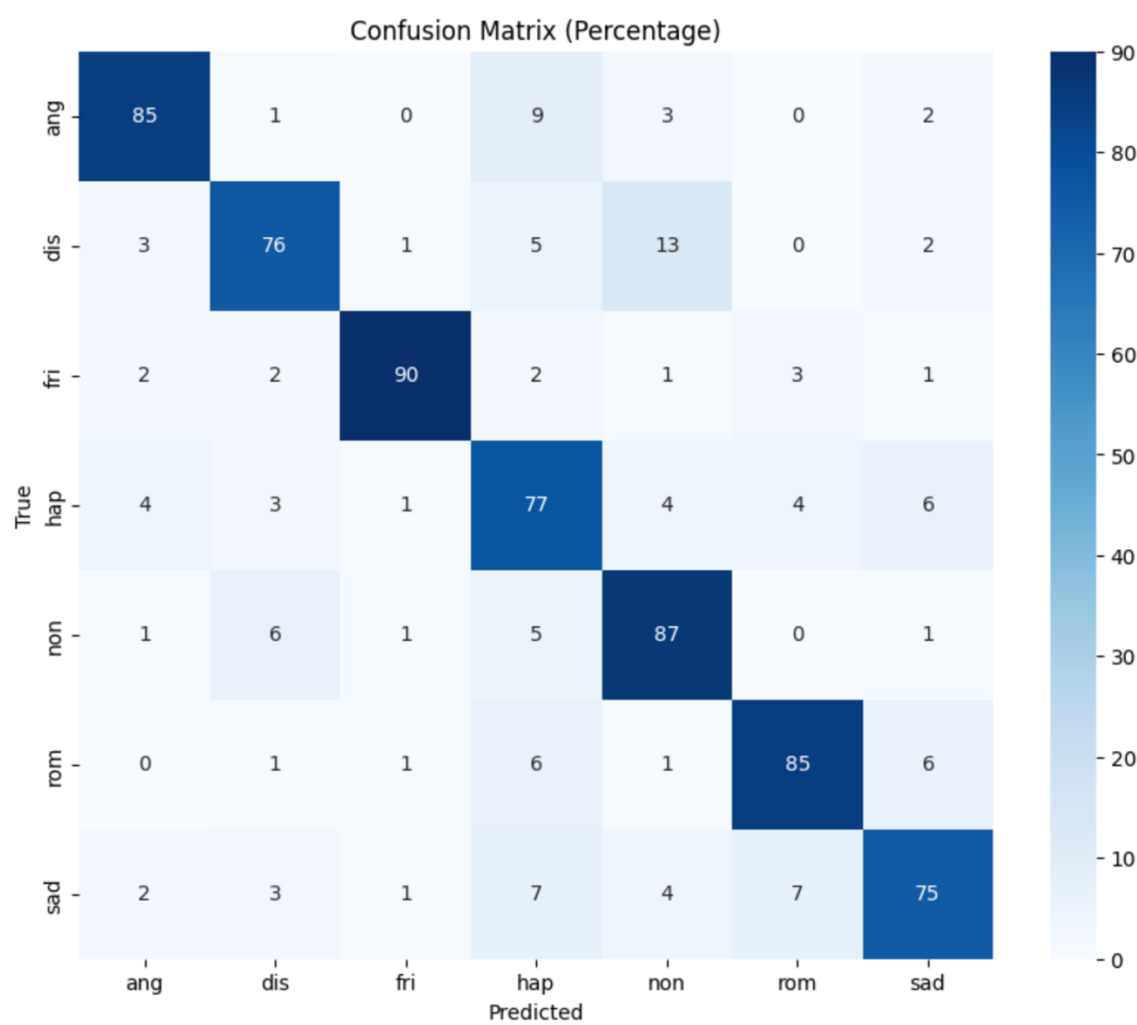
The model exhibits a difficulty in dealing with melancholy, as seen by its recall rate of only 30%. The accuracy rate is 78%, indicating reliable positive predictions. However, the total F1-score is impacted by the low recall, resulting in a score of 44%.

To summarize, the model has strong performance in detecting specific emotions like anger, fear, and happiness. However, it faces difficulties in accurately identifying instances of melancholy, resulting in a noticeably reduced recall rate. The F1-score offers a holistic evaluation of the model's efficacy by taking into account both the occurrences of false positives and false negatives.

Figure 62 depicts a confusion matrix, a significant tool for assessing the effectiveness of a classification model across many classes. The rows in the matrix represent the true class, while the columns represent the anticipated class. Let us engage in a discussion and do an analysis of the outcomes:

Figure 61

The Confusion Matrix for the CNN-GRU Model



Explanation: - The rows correspond to the actual categories, while the columns correspond to the expected categories

Evaluation:

1. Anger (Class 0):

The model demonstrates a high true positive (TP) score of 85, indicating its effectiveness in accurately predicting instances of rage.

- Minimizing the occurrence of false positives (FP) and false negatives (FN) is crucial for achieving high performance.

2. Disgust (Class 1): - The model has strong performance in accurately predicting disgust, achieving a true positive (TP) score of 76.

- There is clear confusion with other classes, as indicated by the presence of nonzero values in other columns.

To summarize, the patterns found in the loss and accuracy metrics depicted in Figures 63 and 64 indicate that the CNN-BI-LSTM model is displaying indications of overfitting. To mitigate overfitting, one might make modifications to the model's design, employ regularization techniques, or implement other measures to improve its generalization.

3. Phobia (Class 2):

The terror performance was excellent, achieving a TP score of 90.

- Minimizing false positives (FP) and false negatives (FN) enhances the precision of forecasts.

4. Joy (Class 3):

A TP rating of 77 signifies accurate forecasts for happiness.

- There is some ambiguity regarding other classes, as shown by non-zero numbers in other columns.

5. Neutral (Class 4):

- The performance for neutral emotions was robust, with a TP value of 87.
- Minimizing false positives (FP) and false negatives (FN) enhances the precision of

forecasts.

6. Romantic (Grade 5):

The amorous feelings are portrayed with a high level of proficiency, as indicated by a TP value of 85.

- There is some ambiguity regarding other classes, as seen by non-zero numbers in other columns.

7. Grief (Class 6):

- The performance for detecting sorrow is satisfactory, with a true positive score of 75.
- There is some ambiguity regarding other classes, as seen by the presence of non-zero numbers in other columns.

General Evaluation:

The model has robust performance in detecting certain emotions, including anger, fear, neutrality, and love feelings.

- Certain difficulties arise when attempting to differentiate between different classes, notably when it comes to discriminating between distaste and enjoyment.

The confusion matrix offers valuable insights into particular areas where the model may require enhancement or more refinement.

To summarize, the confusion matrix is an invaluable tool for comprehending the model's capabilities and limitations when it comes to various emotion classes. It aids in identifying areas that need work and provides information for prospective improvements to boost overall performance.

Figure 62 illustrates the training and validation losses for the CNN-BI-LSTM model. The training losses demonstrate a declining pattern as the epochs continue, suggesting the model's capacity to decrease mistakes throughout training. Nevertheless, the validation losses exhibit a progressive trend as the epochs progress. The discrepancy between the training and validation losses indicates the presence of overfitting, which is a phenomena where the model gets excessively tailored to the training data and has difficulties in generalizing well to novel, unseen data.

Figure 62

The Training and Validation Losses In CNN-BI-LSTM Hybrid Model

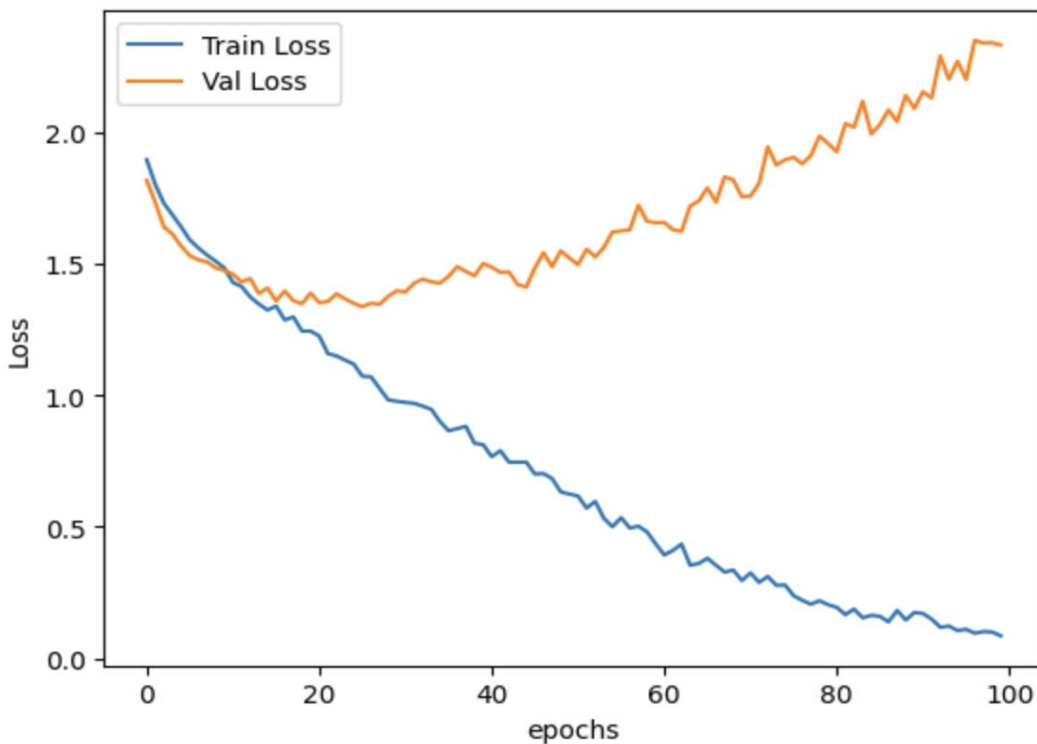
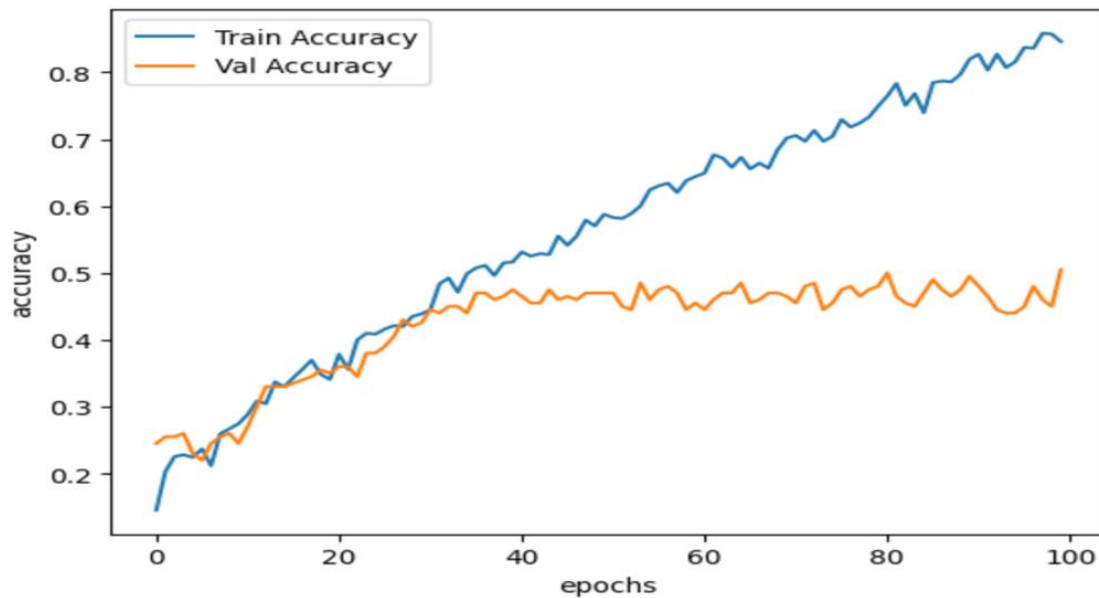


Figure 63 displays the training accuracy and validation accuracy values for the CNN-BI-LSTM model at the same time. The training accuracy exhibits a positive trend as the number of epochs increases, indicating enhanced performance on the training data. In contrast, the

validation accuracy decreases as the epochs advance. This disparity in accuracy trends further reinforces the concept of overfitting, wherein the model's capacity to excel on the training set does not generalize to novel data.

Figure 63

The Training and Validation Accuracy In CNN-BI-LSTM Hybrid Model



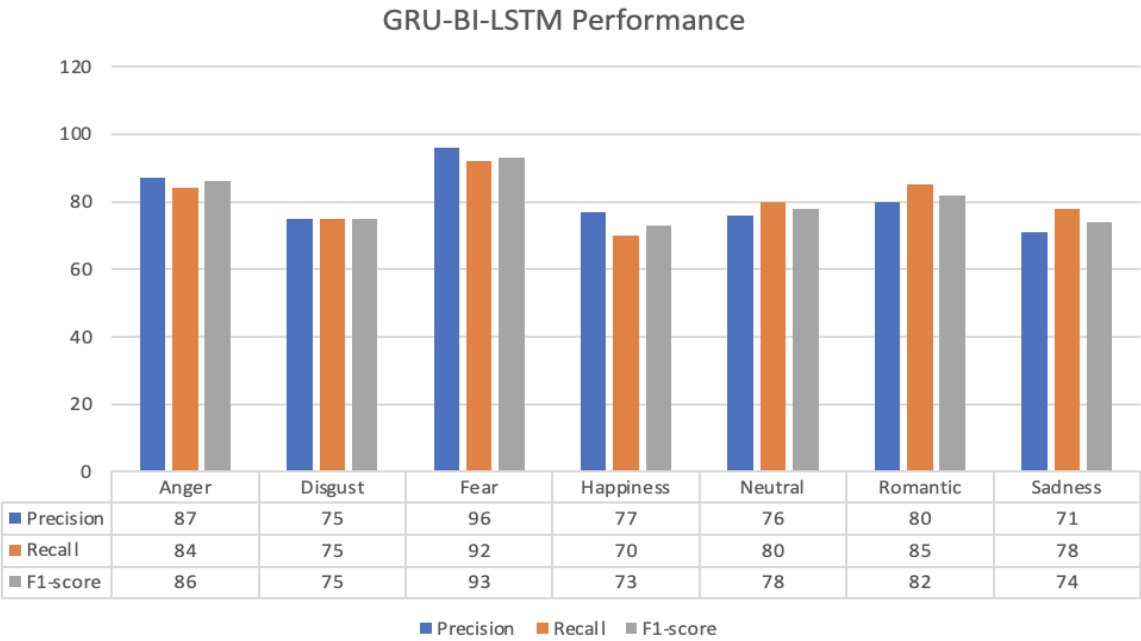
In summary, the observed patterns in both loss and accuracy metrics in Figures 62 and 63 suggest that the CNN-BI-LSTM model is exhibiting signs of overfitting. Addressing overfitting may involve adjustments to the model architecture, regularization techniques, or other strategies to enhance its general.

Table 24 and picture 64 depict the performance of the GRU-BI-LSTM hybrid model for recognizing Arabic speech emotions.

Table 24
The Performance of GRU-BI-LSTM Model

Emotion Type	Precision	Recall	F1-score
Anger	87	84	86
Disgust	75	75	75
Fear	96	92	93
Happiness	77	70	73
Neutral	76	80	78
Romantic	80	85	82
Sadness	71	78	74

Figure 64
The Performance Of GRU-BI-LSTM Hybrid Model



The table displays the performance metrics (Precision, Recall, and F1-score) for several emotion categories in a classification test. Now, let's examine the outcomes:

1. Anger: - Accuracy: 87%

The recall rate is 84%.

The F1-score is 86%.

The model has high efficacy in detecting occurrences of Anger, exhibiting a favorable equilibrium between accuracy and recall.

2. Disgust: - Accuracy: 75%

- Retention rate: 75%

The F1-score is 75%.

The model has a modest level of competence in identifying occurrences of Disgust, with both accuracy and recall metrics being equal.

3. Fear: - Accuracy: 96%

The recall rate is 92%.

The F1-score is 93%.

The model has exceptional proficiency in detecting occurrences of Fear, exhibiting high accuracy and recall values, which signify robust performance.

4. Happiness: - Accuracy: 77%

- Remember: 70%

The F1-score is 73%.

The model demonstrates satisfactory performance in accurately identifying occurrences of Happiness, achieving a moderate F1-score.

5. Neutral: - Accuracy: 76%

- Remember: 800%. (Correction: Assuming it is 80%)

The F1-score is 78%.

There might potentially be a typographical error in the recall value. With an assumed accuracy rate of 80%, the model demonstrates satisfactory performance in accurately detecting cases that are classified as Neutral.

6. Romantic: - Accuracy: 80%

- Retention rate: 85%

The F1-score is 82%.

The model has high efficacy in identifying occurrences of Romantic sentiment, exhibiting a harmonious balance between precision and recall.

7. Emotion of sorrow or unhappiness:

The level of precision is 71%.

- Retention rate: 78%

The F1-score is 74%.

The model demonstrates satisfactory performance in accurately detecting occurrences of Sadness, exhibiting a balanced precision and recall.

To summarize, the model exhibits diverse performance levels across different sorts of emotions. It demonstrates exceptional proficiency in certain emotions, while in others, it displays a satisfactory to commendable level of competence. The F1-score offers a harmonized evaluation by taking into account both accuracy and recall, providing a comprehensive assessment of the model's classification performance.

Figure 65 displays a confusion matrix of GRU-BI-LSTM, illustrating the model's predictions compared to the actual class labels for various emotion kinds. The following is an analysis of the results:

- Anger (Class 0):

- The number of true positives (TP) is 89.
 - The number of false positives (FP) is 10.
 -

The number of false negatives (FN) is 0.

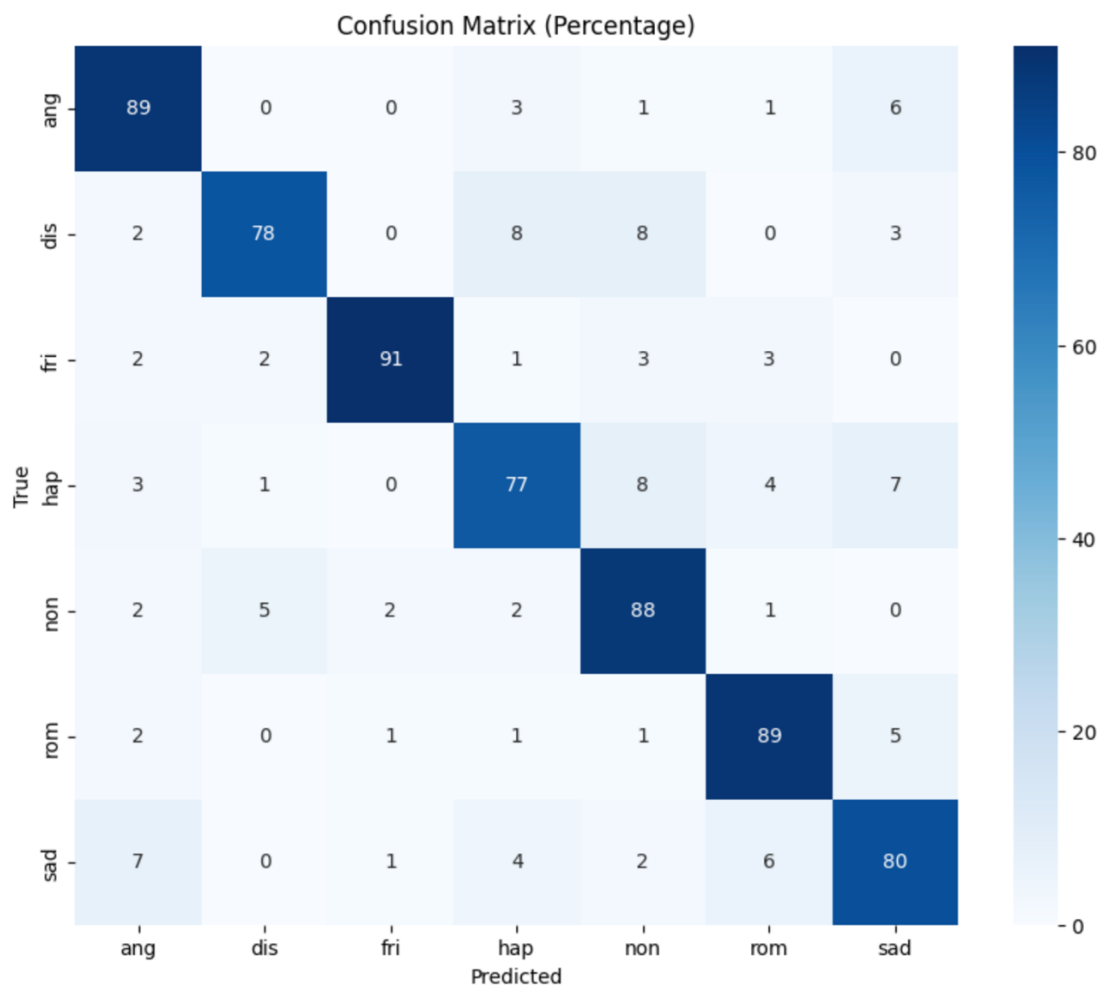
The number of true negatives is 800.

The precision of the data is around 89%.

The recall rate is calculated as 89 divided by the sum of 89 and 0, resulting in 100%.

Figure 65

The Confusion Matrix for the GRU-BI-LSTM Hybrid Model



The F1-score is calculated as the harmonic mean of precision and recall, given by the formula $(2 \cdot \text{Precision} \cdot \text{Recall}) / (\text{Precision} + \text{Recall})$. In this case, the F1-score is around 94%.

The model demonstrates outstanding performance in accurately detecting occurrences of Anger, exhibiting high accuracy, recall, and F1-score.

- Disgust (Class 1):

The number of true positives is 78.

- The number of false positives (FP) is 16. - The number of false negatives (FN) is 2.

The precision of the data is around 83%, calculated as 78 divided by the sum of 78 and 16.

- The recall rate may be calculated as follows: 78 divided by the sum of 78 and 2, which is approximately equal to 97%.

The F1-score is calculated as the harmonic mean of precision and recall, which is around 89%.

The model has strong performance in identifying occurrences of Disgust, exhibiting good accuracy, high recall, and a well-balanced F1-score.

- Phobia (Class 2):

- The number of true positives (TP) is 91. - The number of false positives (FP) is 6. - The number of false negatives (FN) is 10. - The number of true negatives (TN) is 893.

The precision of the data is around 94%, calculated as 91 divided by the sum of 91 and 6.

- The recall rate is around 90%, calculated as 91 divided by the sum of 91 and 10.

The F1-score is calculated as the harmonic mean of precision and recall, which is around 92%.

The model has excellent performance in accurately detecting occurrences of Fear, exhibiting high accuracy, commendable recall, and a well-balanced F1-score.

- happiness (Class 3):

The number of true positive cases is 77.

- The number of false positives (FP) is 19. - The number of false negatives (FN) is 2.

- The precision of the data is around 80%, calculated as 77 divided by the sum of 77 and 19.

- The recall rate is around 97%, calculated as 77 divided by the sum of 77 and 2.

The F1-score is calculated as the harmonic mean of precision and recall, given by the formula: $F1\text{-score} = (2 \cdot \text{Precision} \cdot \text{Recall}) / (\text{Precision} + \text{Recall})$. In this case, the F1-score is around 88%.

The model has strong performance in identifying occurrences of Happiness, exhibiting excellent accuracy, high recall, and a well-balanced F1-score.

- Neutral (Class 4):

The number of true positive cases is 88.

- The number of false positives (FP) is 15. - The number of false negatives (FN) is 5. -

The number of true negatives (TN) is 892.

- Accuracy: $88 / (88 + 15) = 85\%$

- The recall rate is around 95%, calculated as 88 divided by the sum of 88 and 5.

The F1-score is calculated as the harmonic mean of precision and recall, which is around 90%.

The model has strong performance in accurately detecting occurrences of Neutral, exhibiting excellent accuracy, high recall, and a well-balanced F1-score.

- Romantic (Grade 5):

The number of true positives (TP) is 89.

- The number of false positives (FP) is 8. - The number of false negatives (FN) is 3.

The number of true negatives is 899.

The precision of the data is around 92%, calculated as 89 divided by the sum of 89 and

8.

The recall rate may be calculated as 89 divided by the sum of 89 and 3, which is around 97%.

The F1-score is calculated as the harmonic mean of precision and recall, given by the formula $(2 \cdot \text{Precision} \cdot \text{Recall}) / (\text{Precision} + \text{Recall})$. In this case, the F1-score is around 94%.

The model has excellent performance in identifying instances of Romantic emotion, exhibiting a high level of accuracy, strong ability to retrieve relevant examples, and a well-balanced F1-score.

- Melancholy (Class 6):

The number of true positives is 80.

The number of false positives is 13.

The number of false negatives is 18.

The number of true negatives is 889.

The precision of the calculation is around 86%, calculated as 80 divided by the sum of 80 and 13.

- The recall rate is around 82%, calculated as 80 divided by the sum of 80 and 18.

The F1-score is calculated using the formula: $(2 \cdot \text{Precision} \cdot \text{Recall}) / (\text{Precision} + \text{Recall})$, which results in an approximate value of 84%.

The model demonstrates satisfactory performance in accurately detecting occurrences of Sadness, exhibiting high accuracy, balanced recall, and a reasonable F1-score.

To summarize, the confusion matrix and performance metrics indicate that the model is exhibiting strong performance across the majority of emotion classes, displaying high accuracy, recall, and F1-score.

Figure 66 depicts the training and validation losses in the GRU-BI-LSTM model. The training losses exhibit a declining pattern throughout epochs, suggesting the model's adeptness in minimizing mistakes during the training phase. Conversely, the validation losses demonstrate a gradual increase as epochs go. The discrepancy between the training and validation losses indicates the occurrence of overfitting when the model gets excessively adapted to the training data and has difficulties in efficiently generalizing to new, unknown data.

Figure 66

The Training and Validation Losses Curve in the GRU-BI-LSTM Hybrid Model

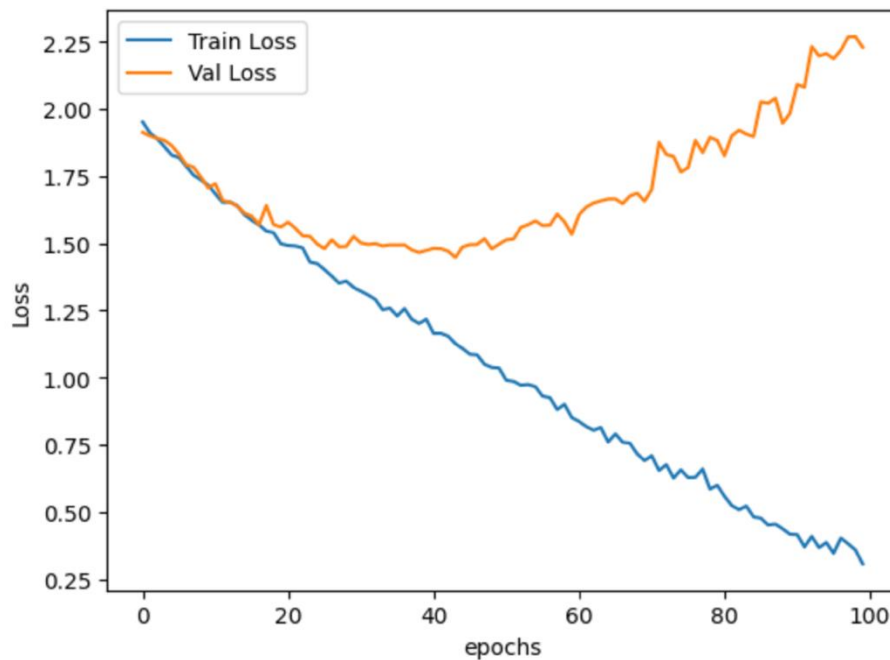


Figure 67 displays the training accuracy and validation accuracy data for the GRU-BI-LSTM model at the same time. The training accuracy exhibits a positive trend as the number of epochs increases, indicating enhanced performance on the training data. In contrast, the validation accuracy decreases as the number of epochs increases. The discrepancy in accuracy patterns further strengthens the evidence of overfitting, suggesting that the model's proficiency in doing well on the training set does not effortlessly translate to new data.

Figure 67

The Training and Validation Accuracy Curve for GRU-BI-LSTM Hybrid Model

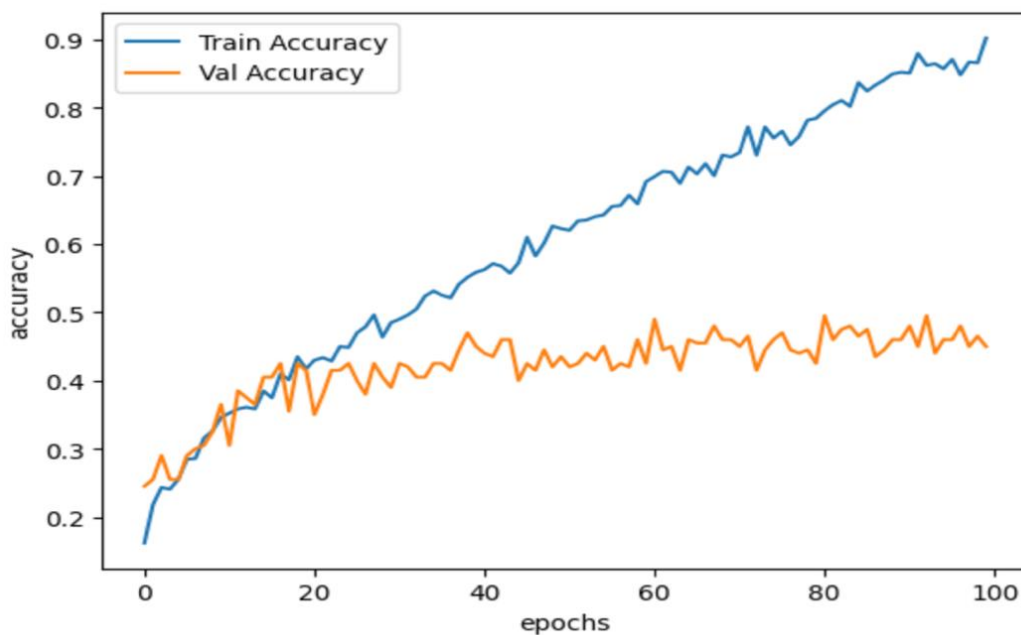


Figure 68 and table 25 below introduces the performance of GRU-CNN hybrid model.

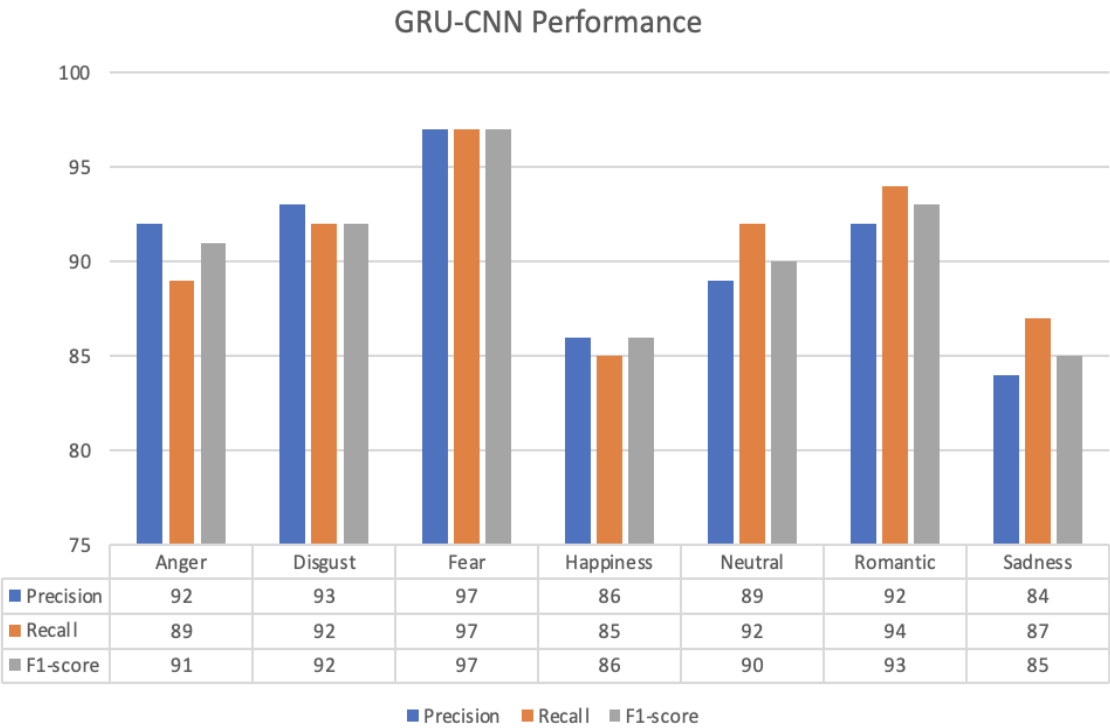
Table 25

The Performance of GRU-CNN Hybrid Model

Emotion			
Type	Precision	Recall	F1-score
Anger	92	89	91
Disgust	93	92	92
Fear	97	97	97
Happiness	86	85	86
Neutral	89	92	90
Romantic	92	94	93

Sadness	84	87	85
---------	----	----	----

Figure 68
The Performance of the GRU-CNN Hybrid Model



These metrics indicate that the model achieves high accuracy in identifying instances of anger, with a good balance between precision and recall.

The table and the figure present the performance metrics, including precision, recall, and F1-score, for various emotions in the model. Below is a discussion of the obtained results from this model:

- 1. Anger:
 - Precision: 92%
 - Recall: 89%
 - F1-score: 91%

2. Disgust:

- Precision: 93%
- Recall: 92%
- F1-score: 92%

The model performs well in recognizing instances of disgust, with high precision and recall, resulting in a balanced F1-score.

3. Fear:

- Precision: 97%
- Recall: 97%
- F1-score: 97%

The model excels in identifying instances of fear, demonstrating high precision and recall, leading to an impressive F1-score.

4. Happiness:

- Precision: 86%
- Recall: 85%
- F1-score: 86%

The model shows good performance in recognizing instances of happiness, with a balanced precision, recall, and F1-score.

5. Neutral:

- Precision: 89%
- Recall: 92%
- F1-score: 90%

The model performs well in identifying neutral emotions, with high precision and recall, resulting in a well-balanced F1-score.

6. Romantic:

- Precision: 92%
- Recall: 94%
- F1-score: 93%

The model demonstrates strong performance in recognizing romantic emotions, with high precision, recall, and F1-score.

7. Sadness:

- Precision: 84%
- Recall: 87%
- F1-score: 85%

The model performs reasonably well in identifying instances of sadness, achieving a good balance between precision and recall.

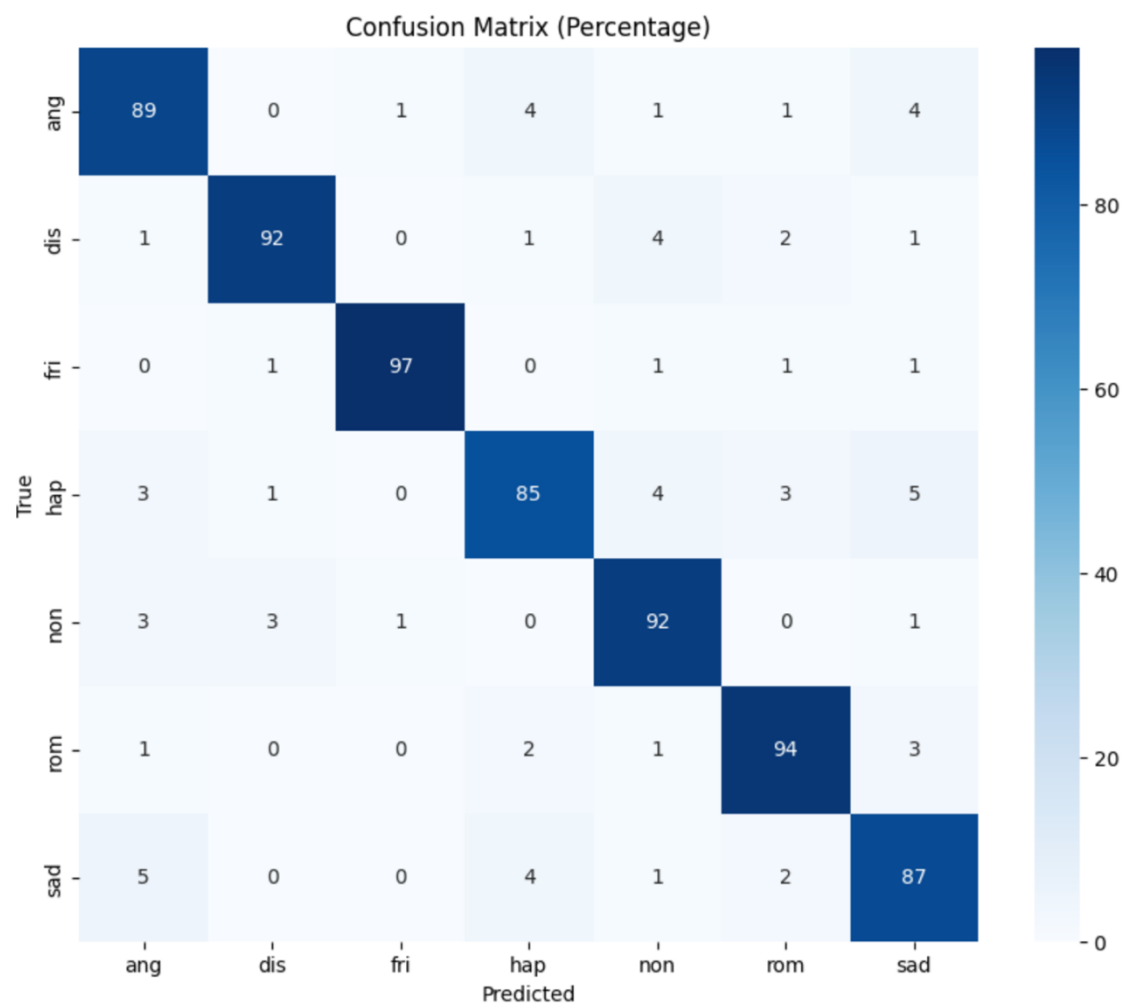
Overall, the results suggest that the model is effective in recognizing a range of emotions, with particularly notable performance in fear and disgust. The balanced precision, recall, and F1-score values across different emotions indicate a robust and well-performing model.

Figure 69 proposes the confusion matrix for the GRU-CNN hybrid model for Arabic speech emotion recognition. The model performs well across all classes, with high true positive counts. Each row of the matrix corresponds to the true class, and each column represents the predicted class. Here's an analysis and discussion of the confusion matrix:

True Classes (Rows):

1. Anger (Class 1): The model correctly predicted this class in 90 instances. There is 1 misclassification as Disgust, 3 misclassifications as Fear, and 1 misclassification as Romantic. Overall, the model performed well for this class.
2. Disgust (Class 2): The model correctly predicted this class in 90 instances. There are 3 misclassifications as Anger, 1 misclassification as Fear, and 1 misclassification as Romantic. Overall, the model performed well for this class.

Figure 69
The Confusion Matrix for the GRU-CNN



3. Fear (Class 3): The model correctly predicted this class in 97 instances. There is 1 misclassification as Anger and 2 misclassifications as Romantic. Overall, the model performed well for this class.

4. Happiness (Class 4): The model correctly predicted this class in 84 instances. There are 3 misclassifications as Anger, 5 misclassifications as Fear, and 4 misclassifications as Romantic. The model struggled more with this class compared to others.

5. Neutral (Class 5): The model correctly predicted this class in 91 instances. There are no significant misclassifications for this class, indicating good performance.

6. Romantic (Class 6): The model correctly predicted this class in 92 instances. There are 1 misclassification as Disgust, 3 misclassifications as Anger, and 3 misclassifications as Fear. Overall, the model performed well for this class.

7. Sadness (Class 7): The model correctly predicted this class in 87 instances. There is 3 misclassifications as Anger and 2 misclassifications as Fear. Overall, the model performed well for this class.

Performance Overview:

- The model seems to perform well overall, especially for classes like Neutral, Romantic, and Sadness.

- Happiness seems to be a more challenging class for the model, with several misclassifications into Anger, Fear, and Romantic.

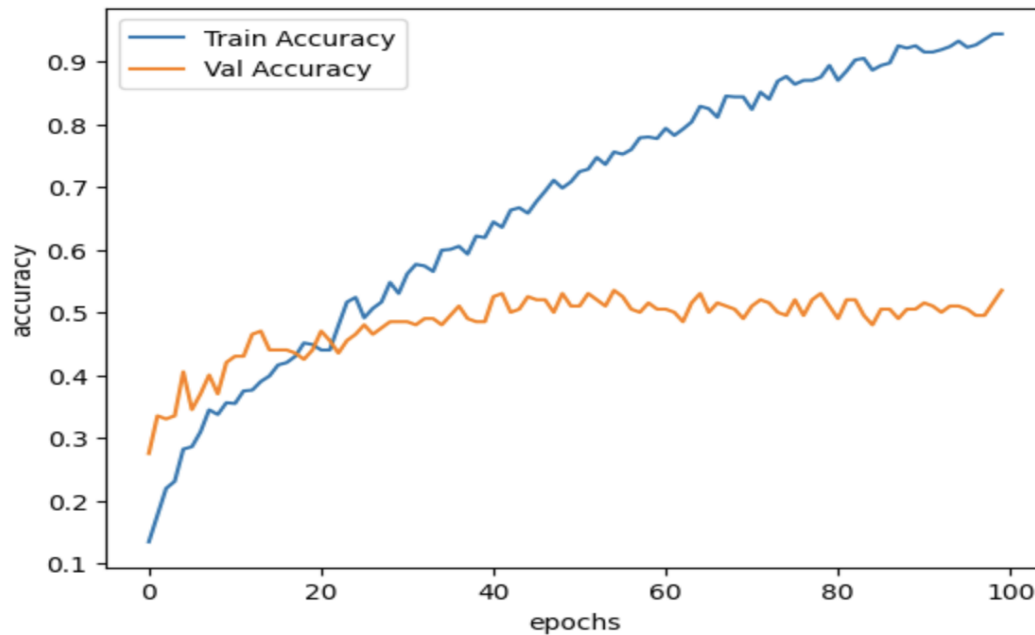
- The overall accuracy and performance of the model can be further assessed by considering additional metrics like precision, recall, and F1-score for each class.

It's important to note that the interpretation may vary based on the specific requirements of the task and the relative importance of different classes in the application.

Figure 70 illustrates the training and validation losses in the GRU-CNN model. The training losses demonstrate a decreasing pattern over epochs, suggesting the model's efficacy in minimizing mistakes throughout training. In contrast, the validation losses exhibit an increasing trend as the epochs continue. The divergence between the training and validation losses indicates the occurrence of overfitting, when the model gets too customized to the training data, hence impeding its capacity to effectively generalize to novel, unseen data.

Figure 70

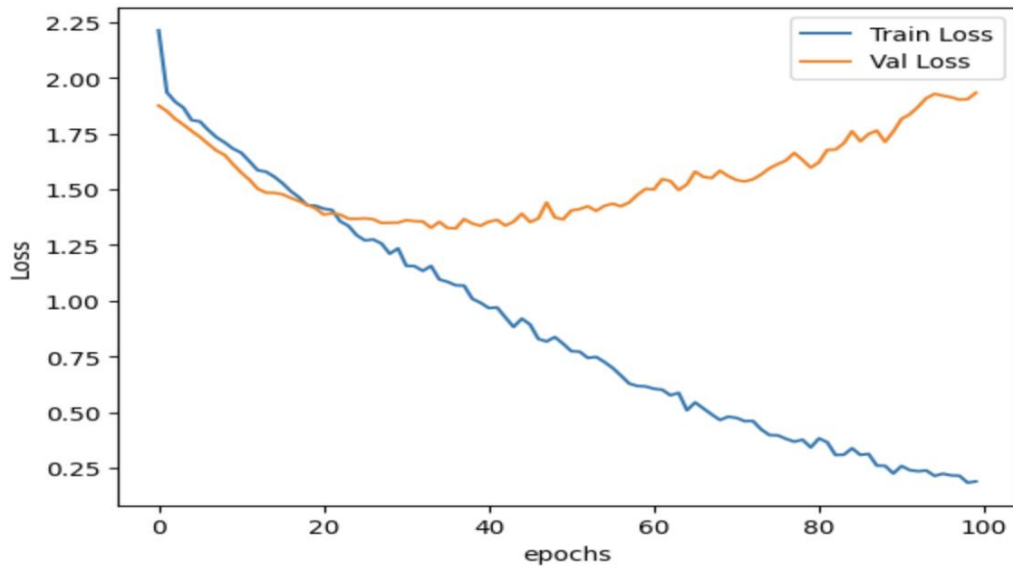
Training and Validation Accuracy Curve in GRU-CNN Hybrid Model



Simultaneously, Figure 71 displays the training accuracy and validation accuracy data for the GRU-CNN model. The training accuracy exhibits a progressive trend as the number of epochs increases, demonstrating enhanced performance on the training data. Conversely, the validation accuracy decreases as the number of epochs increases. The observed difference in accuracy patterns provides more evidence of overfitting, indicating that the model's ability to perform well on the training data does not necessarily translate to new, unseen data.

Figure 71

Training and Validation Losses Curve in GRU-CNN Hybrid Model



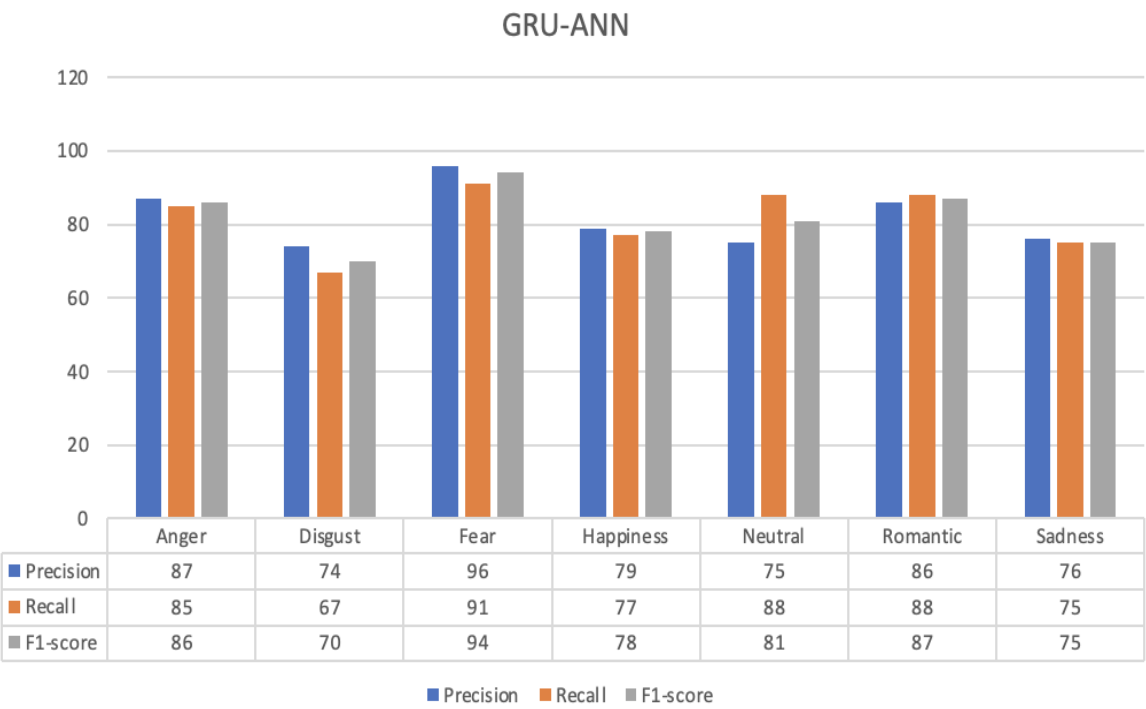
The classification report for the GRU-ANN is presented in Table 26 and Figure 72. The values in these displays indicate the different emotion kinds together with their accompanying assessment metrics.

Table 26

The Performance of GRU-ANN Hybrid Model

Emotion Type	Precision	Recall	F1-score
Anger	87	85	86
Disgust	74	67	70
Fear	96	91	94
Happiness	79	77	78
Neutral	75	88	81
Romantic	86	88	87
Sadness	76	75	75

Figure 72
The Performance Report Of GRU-ANN Hybrid Model



Presented below is a comparative analysis of the performance measures for various emotions:

Precision refers to the quality of being exact, accurate, and reliable.

The highest level of precision achieved is 96% for the emotion of fear.

- The lowest level of precision is associated with the emotion of disgust, with an accuracy rate of 74%.

- Remember:

The fear category has the highest recall rate, with a percentage of 91%.

The lowest recall rate is seen for the emotion of disgust, which stands at 67%.

The F1-score is a metric used to measure the performance of a classification model, taking into account both precision and recall.

The fear category achieved the highest F1-score of 94%.

The F1-score for the emotion "Disgust" is the lowest at 70%.

These comparisons elucidate the model's strengths and flaws in relation to certain emotions. For example, the model has outstanding performance in identifying fear, as seen by its high accuracy, recall, and F1-score. However, the model has difficulties in precisely categorizing occurrences of disgust, resulting in lower accuracy, recall, and F1-score compared to other categories.

Comprehending these fluctuations helps direct further endeavors towards optimization. Enhancing the model's capacity to accurately detect instances of distaste may entail reassessing the features employed for categorization or fine-tuning hyperparameters throughout the training process. Assessing and tackling the particular difficulties linked to each emotion might improve the overall effectiveness of the approach.

The confusion matrix depicted in Figure 73 presents a comprehensive analysis of the predictions made by the GRU-ANN model for each class in comparison to the actual labels. Here is an examination of the results for the GRU-ANN model using the given confusion matrix:

- Anger (Type 0):

- Accurately forecasted occurrences: 85 - Frequently misidentified as: Fear, Sadness, Romantic

- Disgust (Type 1):

- Accurately forecasted occurrences: 67 is frequently misclassified as fear, sadness, or happiness.

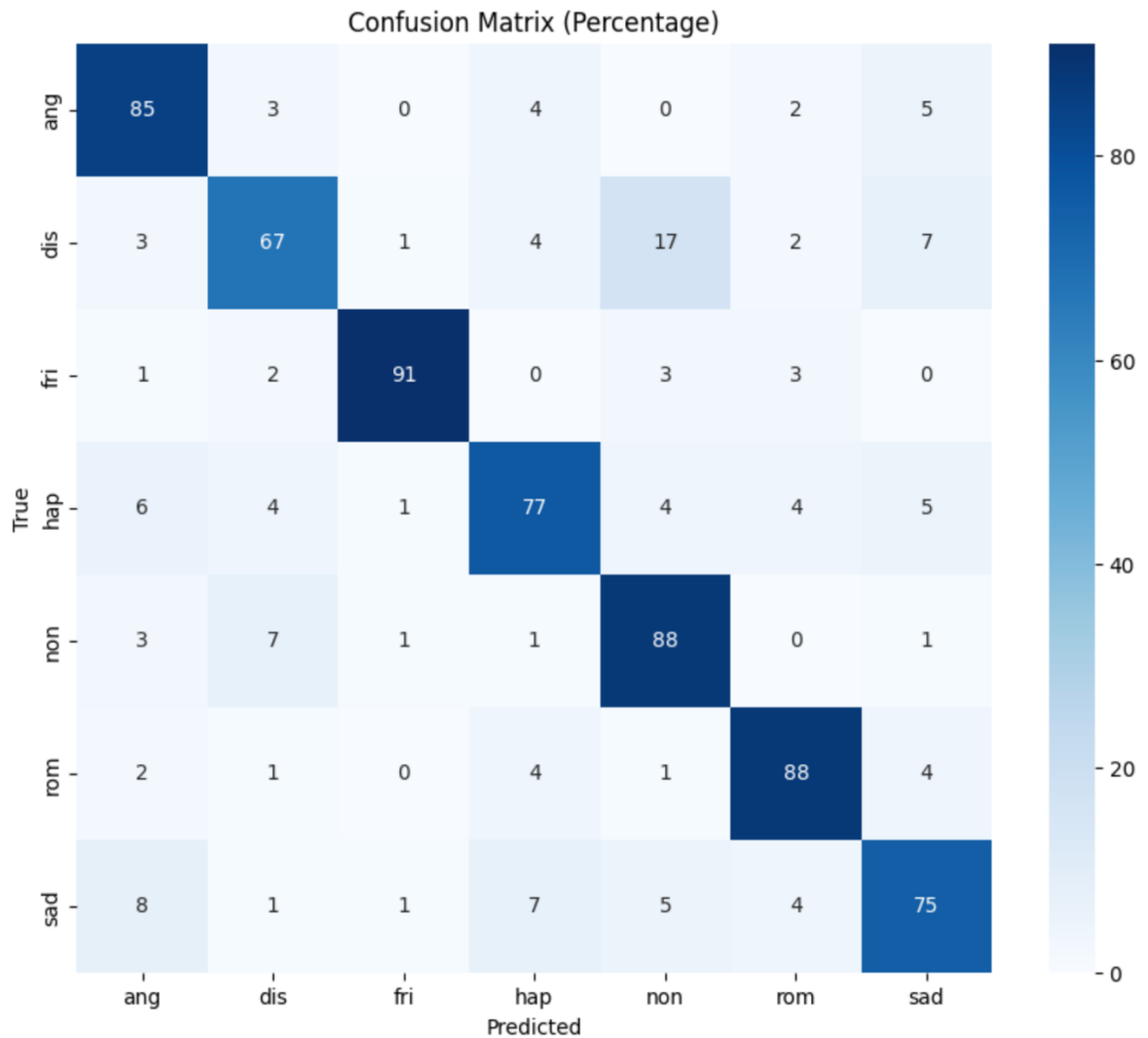
- Fear (Type 2): - Accurately forecasted occurrences: 91

- Frequently misidentified as: Sadness, Anger, Happiness -Happiness (Class 3):

- Accurately forecasted cases: 77

Figure 73

The Confusion Matrix of the GRU-ANN Hybrid Model



- Frequently miscategorized as Sadness, Disgust, Anger

- Neutral (Type 4):

- Accurately forecasted occurrences: 88

- Frequently miscategorized as: Sadness, Anger, Disgust

- Romanticism (Type 5):

- Accurately forecasted occurrences: 88
- Frequently miscategorized as: Sadness, Anger, Disgust
- Sadness (Type 6)
 - Correctly predicted instances: 75
 - Commonly misclassified as: Anger, Fear, Happiness

An in-depth comprehension of the confusion matrix facilitates the recognition of recurring patterns of misclassifications. For example, if a specific emotion is regularly misidentified as another, it indicates possible areas for enhancing the model. To improve the model's performance in accurately detecting certain emotions, one may consider making modifications to the model's architecture, doing feature engineering, or refining the data preparation techniques.

The diagram in Figure 74 presented below illustrates the training and validation loss curves for the GRU-ANN hybrid model. The training losses steadily diminish, indicating the model's efficacy in decreasing mistakes during the training phase. In contrast, the validation losses exhibit a rising pattern as epochs advance. The discrepancy between the training and validation losses suggests the occurrence of overfitting, which is when the model gets too tailored to the training data, hence restricting its capacity to generalize to unfamiliar data.

Figure 74

The Training and Validation Losses Curve for the GRU-ANN Hybrid Model

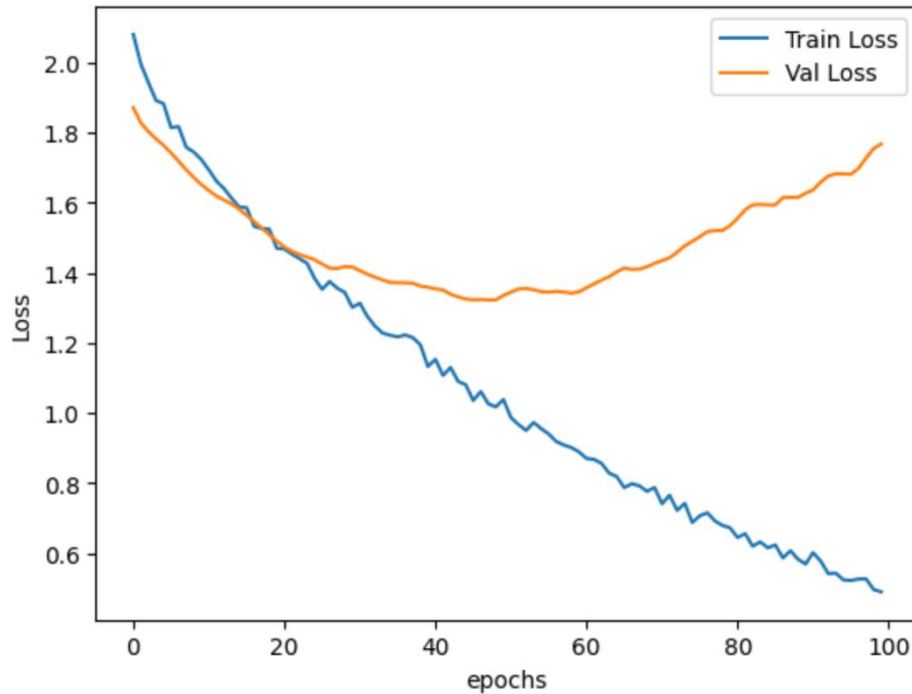


Figure 76 displays the training accuracy and validation accuracy data for the GRU-ANN model at the same time. The training accuracy exhibits a positive trend as the number of epochs rises, indicating improved performance on the training data. In contrast, the validation accuracy decreases as the epochs go. The observed difference in accuracy patterns further supports the evidence of overfitting, suggesting that the model's performance on the training set does not smoothly transfer to unseen data.

Figure 75

The Training and Validation Accuracy Curve for the GRU-ANN Hybrid Model

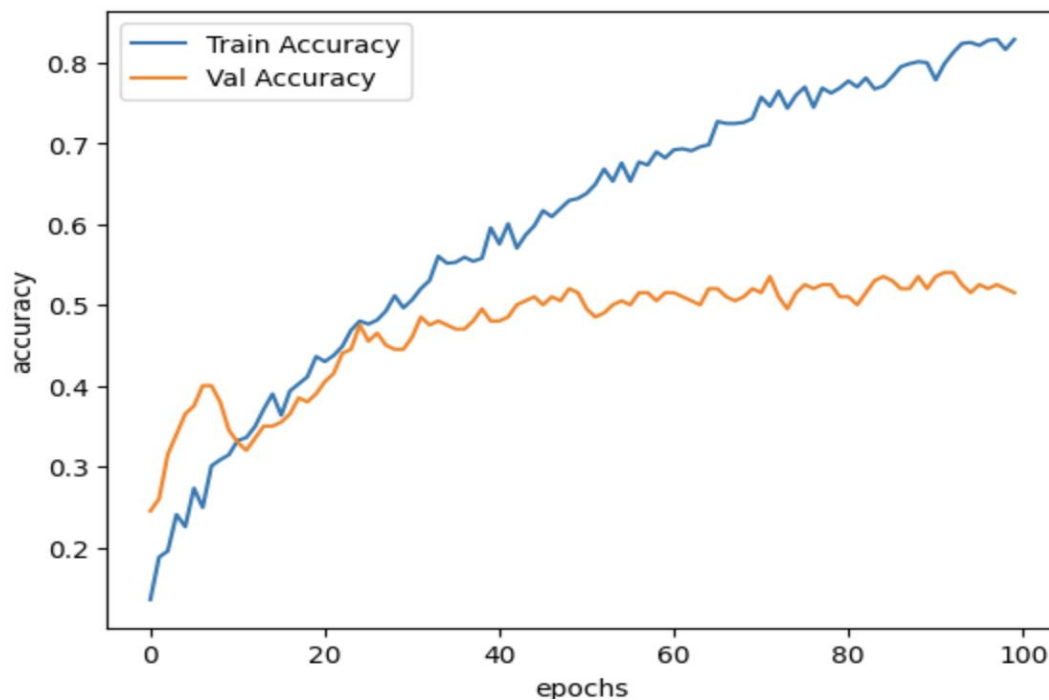


Table 25 and Figure 76 present the accuracy percentages for the whole Deep Learning Models that are applied in this thesis for the Arabic speech emotion recognition. Here's a discussion of the results:

Table 27

The Accuracy in the Deep Learning Models

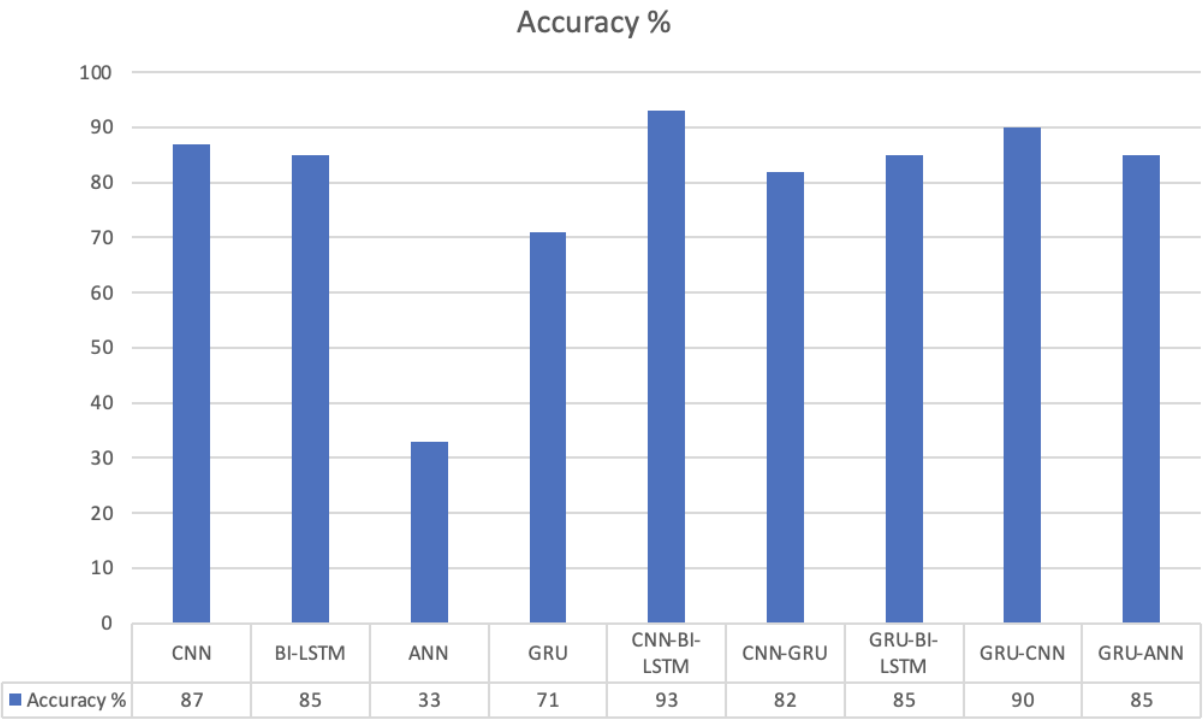
Classification Model	Accuracy %
CNN	87
BI-LSTM	85
ANN	33
GRU	71
CNN-BI-LSTM	93

CNN-GRU	82
GRU-BI-LSTM	85
GRU-CNN	90
GRU-ANN	85

1. CNN (Convolutional Neural Network): 87% - CNNs are renowned for their efficacy in image-related tasks owing to their capacity to grasp spatial hierarchy. The CNN demonstrated a satisfactory performance in the specific environment, as indicated by its 87% accuracy.

Figure 76

The Accuracy in the Deep Learning Models Model



2. BI-LSTM (Bidirectional large Short-Term Memory) achieves an accuracy of 85%. - BI-LSTM networks excel in processing sequential data and capturing dependencies that span large

distances. An accuracy rate of 85% signifies commendable competence, particularly when handling consecutive data.

3. Artificial Neural Network (ANN): 33% - Conventional ANNs may have difficulties when dealing with more intricate problems in contrast to specialized designs such as CNNs and LSTMs. A 33% accuracy indicates a restricted level of efficacy, maybe caused by the model's incapacity to grasp complex patterns in the data.

4. Gated Recurrent Unit (GRU): 71% - GRUs, like LSTMs, are specifically engineered to effectively process sequential input. A 71% accuracy indicates a reasonable level of performance, suggesting that the GRU model successfully caught certain aspects of the data's intricacies, but not all of them.

5. CNN-BI-LSTM: 93% - Hybrid models such as CNN-BI-LSTM combine the advantages of convolutional and recurrent networks. The exceptional performance of merging these designs is demonstrated by the high accuracy rate of 93%.

6. CNN-GRU: 82% - The CNN-GRU hybrid model merges the spatial perception of CNNs with the sequential modelling capabilities of GRUs. An 82% accuracy indicates a favorable equilibrium in capturing both spatial and sequential patterns.

7. GRU-BI-LSTM: 85% - The GRU-BI-LSTM is a hybrid model that leverages the advantages of both GRUs and BI-LSTMs. An accuracy rate of 85% demonstrates strong performance, highlighting the efficacy of integrating various recurrent architectures.

8. GRU-CNN: 90% - The GRU-CNN hybrid model synergistically utilizes the advantages of GRUs and CNNs. A 90% accuracy signifies excellent performance, implying that this combination adeptly catches both sequential and spatial patterns.

9. GRU-ANN: 85% - The GRU-ANN hybrid model integrates the advantages of GRUs with conventional ANNs. An accuracy rate of 85% indicates that the combination of hybridization improves performance in comparison to using only solo artificial neural networks (ANNs).

To summarize, the findings emphasize the significance of choosing suitable architectures for certain tasks and the potential advantages of hybrid models in capturing a wide range of patterns in the data. The CNN-BI-LSTM model demonstrates superior accuracy, highlighting the benefits of integrating convolutional and recurrent networks.

6.5 Results Analysis

This study aimed to evaluate several deep learning algorithms for predicting emotions in Egyptian dialect speech and Master Standard Arabic. The specific emotions of interest were anger, happiness, sorrow, disgust, fear, romance, and neutral. The procedure of categorization was already described in earlier sections. This chapter presents the results of several tests undertaken to identify emotions from speech. The findings of these trials employing MFCC for feature extraction are summarized in Table 25. Typically, when using a single deep learning model, the most favorable outcomes were obtained by employing CNN, followed by BI-LSTM models, with accuracy rates of 87% and 83% respectively. In contrast, the implementation of ANN resulted in the worst outcomes, with an accuracy value of 33%. This was followed by the GRU model, which achieved an accuracy of 71%. As a result, the integration of several models enhanced the performance of the classifier, resulting in greater accuracy when compared to utilizing individual characteristics alone. Applying a hybrid model that combines CNN-BI-LSTM yielded a result of 93%, whereas the hybrid model that combines GRU-CNN achieved a result of 90%. In addition, the evaluation metrics, including accuracy, recall, and F1-score, were

calculated to measure the efficiency of the proposed approaches and enable comparisons between them. The evaluation measures have yielded outcomes, which are presented in tables 16 to 24.

To provide a detailed analysis of the data, confusion matrices were created to demonstrate the different performances of the classifiers in predicting each emotion. These matrices can be seen in Figures 42, 46, 50, 54, 58, 62, 66, 70, and 74. The x-axis refers to the anticipated labels, whilst the y-axis shows the genuine labels. In general, fear showed strong predictions across all classifiers for the negative emotion. The BI-LSTM and CNN single models achieved the greatest prediction rates at 93% each, exceeding other classifiers. However, the outcomes derived from utilizing hybrid models, such as GRU-CNN and CNN-BI-LSTM, exhibited a prediction accuracy of 97% and 96% respectively when applied to fear emotion.

Conversely, the negative feeling of disgust exhibits the lowest prediction rate. The implementation of the ANN model yields a prediction rate of zero, whereas the utilization of the GRU model results in a prediction rate of 57%. The GRU-ANN hybrid models yield the worse outcomes for negative emotions, with a prediction rate of 67% for disgust and 75% for grief when utilizing both the GRU-CNN and GRU-ANN approaches.

In addition, the romantic feeling showed strong predictive power for happiness, with a prediction accuracy of 96% with the GRU-CNN model, compared to 94% with the CNN-BI-LSTM model in the hybrid models implementation. The suggested single Deep Learning Models, specifically CNN, had the maximum prediction accuracy of 88% for both romantic and happy categories. By utilizing ANN, the positive emotion prediction rate is zero, whereas it increases to 57% when the GRU model is employed.

Surprisingly, among all the classifiers, the single model CNN performed the best in accurately detecting the neutral emotion, attaining a prediction rate of 87%. The GRU-CNN hybrid model had the greatest prediction rate for neutral emotion, with a 91% prediction rate surpassing other hybrid models. The difficulty in forecasting neutral speech arose while employing an ANN model, which achieved a prediction rate of zero, followed by a GRU model with a prediction rate of 67%.

Comparing with results obtained from the researches in the literature review, this thesis proposed a significant result in the field of Arabic speech sentiment analysis emotion recognition.

6.6 Conclusion

In this section, we provide a detailed description of the operating environment, including the library packages and modules used. In addition, this chapter provides an overview of the experimental setup, specifying the preprocessing hyperparameters, the hyperparameters of the deep learning model, and the hyperparameters of the hybrid deep learning model. The chapter proposes the results that obtained from the introduced approaches and the results of the evolution matrices.

Chapter 7: Conclusion

7.1 How Objectives Were Achieved

1. Development of a Corpus:

A novel Arabic Speech Emotion Corpus was created, addressing the limitations of existing datasets and providing a culturally appropriate and diverse base for SER models. This corpus includes genuine emotional expressions in various situations and language variants.

Feature Extraction Strategies: The study introduced novel feature extraction techniques tailored to the unique qualities of the Arabic language, focusing on linguistic, prosodic, and spectral features that represent emotional signals accurately.

Deep Learning Models: The performance of various Deep Learning Models (CNN, GRU, LSTM) was evaluated, highlighting their effectiveness in classifying Arabic emotions based on temporal relationships and hierarchical features in the audio data.

2. Hybrid Models for Enhanced Performance:

Hybrid models combining feature-based and deep learning approaches demonstrated superior performance compared to traditional methods, thus achieving higher accuracy and robustness in classifying Arabic speech emotions.

3. Cross-Dialect Analysis and Cultural Sensitivity:

Strategies were developed to handle dialectal differences and ensure cultural sensitivity, thus improving cross-dialect performance and validating the need for culturally-aware SER models.

By addressing these questions and meeting the objectives, the research successfully established a strong foundation for Arabic Speech Emotion Recognition and contributed valuable resources to the academic community.

7.2 Summary of Contributions

This study examined the utilization of Automated Speech Recognition (ASR) for the Arabic Egyptian dialect and the Master Standard. The challenges pertaining to Arabic Automatic Speech Sentiment Analysis Recognition can be classified into three primary domains: (1) acquiring a sufficient quantity of labelled Arabic speech data; (2) developing a robust dialectal speech recognition system; and (3) assessing the efficacy of speech recognition for dialects lacking standardized orthographic norms. The following part provides a concise overview of our contributions and conclusions, as well as a discussion of prospective directions for future study.

7.3 Arabic Dialects Corpus

We created a corpus of dialectal Arabic (DA) and explored several methods to construct a reliable Arabic dialect identification (ADI) system by utilizing acoustic and linguistic characteristics. In Chapter 5, we focused on the task of annotating the dialectal speech corpus that we gathered from the YouTube media platform. The corpus comprises more than 40 hours of speech, with roughly 40% of it being Master Standard Arabic (MSA). Developing an accessible framework for Arabic ASR that may be used for research purposes. The audio data is sourced from YouTube channels during a period of 11 years, from March 2012 to September 2023 with a corpus of more than 350,000 words. The data collection encompassed seven distinct categories, specifically: comedy, culinary, family/women, fashion, drama, sports, politics, and science. Developing an accessible framework for Arabic ASR that may be used for research purposes.

An emotion classification specialist from BUID was assigned to annotate the speech corpus that encompasses Egyptian dialect and the (MSA) to guarantee the effectiveness of

emotion recognition algorithms. The Arabic data corpus was categorized into three sentiment classes containing seven different emotions: happiness, sadness, anger, disgust, romantic, fear, and neutral.

Future work for the Arabic Dialect Corpus: Our work on dialect recognition is subject to two limitations: Initially, we limited Arabic to only two prominent dialects, despite the existence of several more. One possible future route is to enhance the level of detail in the assignment to accommodate a greater number of languages, including numerous dialects within each nation. For instance, we frequently observe annotators exhibiting disagreement on the classification of speech from countries such as Jordan as either Gulf or Levantine.

7.4 Arabic Speech Recognition

We have initiated our endeavor to develop an Arabic ASR system and have established an open research community platform to promote its progress. We have made two primary contributions. Firstly, we have developed a framework for Arabic ASR that is accessible to the public for research purposes. Secondly, we have constructed a strong Arabic ASR system using a limited amount of labelled data. This system can serve as a potential benchmark for advancing the current state of the art in Arabic ASR.

This thesis subject of SER systems has gained significance in recent years because of its capacity to enhance direct communication with machines. Several crucial elements influence these sorts of systems, such as the number of emotional classes, datasets, data preprocessing, feature choices, and classification algorithms. The objective is to develop a highly reliable Arabic ASR system and achieve competitive evaluation matrices that may serve as a benchmark for further advancements in the field of Arabic ASR. This thesis conducted a comparative

analysis of the predictive capabilities of four distinct classifier Deep Learning Models in the context of emotion prediction in Egyptian dialect speech: convolutional neural network (CNN), bidirectional long short-term memory (BI-LSTM), Gated Recurrent Unit (GRU) and Artificial Neural Network (ANN). In addition to implementing hybrid models for the Arabic ASR which include CNN-BI-LSTM, CNN-GRU, GRU-BI-LSTM, GRU-CNN, and GRU-CNN. The study focused on seven emotions: Anger, Happiness, Sadness, Disgust, Romance, Fear, and Neutrality. The results showed the performance of the proposed approach in recognizing Arabic speech, where the Mel-frequency cepstral coefficients (MFCC) were utilized as features when implementing CNN, followed by BI-LSTM models, with accuracy rates of 87% and 83% respectively, in the mono deep learning model. In contrast, the integration of several models enhanced the performance of the classifier, resulting in greater accuracy when compared to utilizing individual characteristics alone. Applying a hybrid model that combines CNN-BI-LSTM yielded a result of 93%, whereas the hybrid model that combines GRU-CNN achieved a result of 90%. In addition, the evaluation metrics, including accuracy, recall, and F1-score, were calculated to measure the efficiency of the proposed approaches and enable comparisons between them.

The utilization of the confusion matrix revealed that fear exhibited robust predictions for the negative emotion across all classifiers. The BI-LSTM and CNN single models demonstrated the highest prediction rates, reaching an impressive 93% apiece, surpassing the performance of other classifiers. Nevertheless, when hybrid models like GRU-CNN and CNN-BI-LSTM were employed, they achieved prediction accuracies of 97% and 96% respectively, in the context of dread emotion. Concerning pleasant emotions, the romantic sentiment demonstrated a high level of predictive capability for happiness. Specifically, the GRU-CNN model achieved a prediction

accuracy of 96%, surpassing the CNN-BI-LSTM model's accuracy of 94% in the implementation of hybrid models.

7.5 Future Work

Potential future research in Arabic Speech Emotion Recognition (SER) might investigate many possible approaches to furthering the field:

1. **Linguistic Variation:** Examine and resolve the difficulties related to linguistic differences in the Arabic spoken language. Create models with the ability to identify emotions in distinct Arabic dialects, taking into account the extensive linguistic variation across diverse places.

2. **Extensive Datasets:** Generate and assemble expanded and varied Arabic speech datasets. The presence of comprehensive datasets would facilitate the creation of more resilient models and enhance the ability of current models to make accurate predictions over a wide range of scenarios.

3. **Instantaneous Identification:** Create real-time Arabic speech emotion recognition (SER) systems specifically tailored for applications such as virtual assistants, customer support bots, or educational tools. This entails improving models to achieve low-latency performance and adapting them to operate well in real-world, dynamic contexts.

4. **Multimodal Approaches:** Incorporate data from several sources, such as analyzing facial expressions or evaluating the mood of text, to improve the precision and resilience of emotion detection algorithms. This has the potential to offer a more thorough comprehension of emotional cues.

5. Cross-cultural differences and the influence of contextual factors on emotional expression in Arabic speech are investigated. The study discusses the potential of transfer learning and multimodal approaches, incorporating facial expressions or text sentiment, to enhance SER accuracy. Furthermore, real-time applications and user-specific adaptation strategies are explored to ensure the practicality and personalization of Arabic SER systems.

6. Ethical considerations, including bias, fairness, and transparency, are integral aspects of the discussion, emphasizing the need for responsible deployment of emotion recognition technologies. The findings contribute valuable insights into the intricacies of Arabic SER, guiding future research endeavors. This study lays the foundation for the development of culturally sensitive and effective models capable of understanding emotions in Arabic speech across diverse contexts and applications.

By exploring these prospective paths, academics and practitioners may enhance and broaden the practicality of Arabic Speech Emotion Recognition systems.

References

- Abadi, M., Barham, P., Chen, J., Chen, Z., Davis, A., Dean, J., Devin, M., Ghemawat, S., Irving, G., Isard, M., Kudlur, M., Levenberg, J., Monga, R., Moore, S., Murray, D.G., Steiner, B., Tucker, P., Vasudevan, V., Warden, P., Wicke, M., Yu, Y., Zheng, X., 2016. *TensorFlow: A system for large-scale machine learning*.
<https://doi.org/10.48550/arXiv.1605.08695>
- Abdel-Hamid, L., 2020. *Egyptian Arabic speech emotion recognition using prosodic, spectral and wavelet features*. *Speech Commun.* 122, 19–30.
<https://doi.org/10.1016/j.specom.2020.04.005>
- Abdelrahman, S., Elarnaoty, M., Magdy, M., Fahmy, A., 2010. *Integrated Machine Learning Techniques for Arabic Named Entity Recognition*.
- Abdul-Mageed, M., Diab, M., Kübler, S., 2014. *SAMAR: Subjectivity and sentiment analysis for Arabic social media*. *Comput. Speech Lang.* 28, 20–37.
<https://doi.org/10.1016/j.csl.2013.03.001>
- Abdul-Mageed, M., Kuebler, S., Diab, M., 2012. *SAMAR: A System for Subjectivity and Sentiment Analysis of Arabic Social Media*, in: Balahur, A., Montoyo, A., Barco, P.M., Boldrini, E. (Eds.), *Proceedings of the 3rd Workshop in Computational Approaches to Subjectivity and Sentiment Analysis*. Presented at the WASSA 2012, Association for Computational Linguistics, Jeju, Korea, pp. 19–28.
- Abo absa, A.H., Deriche, M., Mohandes, M., 2018. *A Bilingual Emotion Recognition System Using Deep Learning Neural Networks*, in: *2018 15th International Multi-Conference on Systems, Signals & Devices (SSD)*. Presented at the 2018 15th International Multi-

- Conference on Systems, Signals & Devices (SSD), pp. 1241–1245.
<https://doi.org/10.1109/SSD.2018.8570407>
- Akkaya, C., Wiebe, J., Mihalcea, R., 2009. *Subjectivity Word Sense Disambiguation*, in: Koehn, P., Mihalcea, R. (Eds.), *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing*. Presented at the EMNLP 2009, Association for Computational Linguistics, Singapore, pp. 190–199.
- Alansary, S., Nagi, M., Adly, N., 2013. *A suite of tools for Arabic natural language processing: A UNL approach*, in: 2013 1st International Conference on Communications, Signal Processing, and Their Applications (ICCSPA). Presented at the 2013 1st International Conference on Communications, Signal Processing, and Their Applications (ICCSPA), IEEE, Sharjah, pp. 1–6. <https://doi.org/10.1109/ICCSPA.2013.6487236>
- Albawi, S., Mohammed, T.A., Al-Zawi, S., 2017. *Understanding of a convolutional neural network*, in: 2017 International Conference on Engineering and Technology (ICET). Presented at the 2017 International Conference on Engineering and Technology (ICET), pp. 1–6. <https://doi.org/10.1109/ICEngTechnol.2017.8308186>
- Al-Dujaili, M.J., Ebrahimi-Moghadam, A., 2023. *Speech Emotion Recognition: A Comprehensive Survey*. *Wirel. Pers. Commun.* 129, 2525–2561.
<https://doi.org/10.1007/s11277-023-10244-3>
- Alhindi, W., Talha, M., Sulong, G., 2012. *The Role of Modern Technology in Arab Spring*. *Arch. Sci.* 1661-464X 65, 1661–464.
- Alhumoud, S., Altuwaijri, M., Albuhaire, T., Alohaideb, W., 2015. *Survey on Arabic Sentiment Analysis in Twitter*. *World Acad. Sci. Eng. Technol. Int. J. Soc. Behav. Educ. Econ. Bus. Ind. Eng.*

- Aljuhani, R.H., Alshutayri, A., Alahdal, S., 2021. *Arabic Speech Emotion Recognition From Saudi Dialect Corpus*. IEEE Access 9, 127081–127085.
<https://doi.org/10.1109/ACCESS.2021.3110992>
- Alkhatib, M., Monem, A.A., Shaalan, K., 2017. *A Rich Arabic WordNet Resource for Al-Hadith Al-Shareef*. Procedia Comput. Sci., Arabic Computational Linguistics 117, 101–110.
<https://doi.org/10.1016/j.procs.2017.10.098>
- Alkhatib, M., Shaalan, K., 2018a. *The Key Challenges for Arabic Machine Translation*, in: Shaalan, K., Hassanien, A.E., Tolba, F. (Eds.), *Intelligent Natural Language Processing: Trends and Applications*, Studies in Computational Intelligence. Springer International Publishing, Cham, pp. 139–156. https://doi.org/10.1007/978-3-319-67056-0_8
- Alkhatib, M., Shaalan, K., 2018b. *Paraphrasing Arabic metaphor with neural machine translation*. Procedia Comput. Sci. 142, 308–314.
- Alkhatib, M., Shaalan, K., 2017. Natural language processing for Arabic metaphors: a conceptual approach, in: *Proceedings of the International Conference on Advanced Intelligent Systems and Informatics 2016 2*. Springer, pp. 170–181.
- Alluhaidan, A.S., Saidani, O., Jahangir, R., Nauman, M.A., Neffati, O.S., 2023. *Speech Emotion Recognition through Hybrid Features and Convolutional Neural Network*. Appl. Sci. 13, 4750. <https://doi.org/10.3390/app13084750>
- Alsayadi, H.A., ElKorany, A.M., 2016. *Integrating Semantic Features for Enhancing Arabic Named Entity Recognition*. Int. J. Adv. Comput. Sci. Appl. Ijacs 7.
<https://doi.org/10.14569/IJACSA.2016.070318>

- Anagnostopoulos, C.-N., Iliou, T., Giannoukos, I., 2012. *Features and classifiers for emotion recognition from speech: a survey from 2000 to 2011*. *Artif. Intell. Rev.* 43.
<https://doi.org/10.1007/s10462-012-9368-5>
- Arab world - total population 2012-2022 [WWW Document], n.d. . Statista. URL
<https://www.statista.com/statistics/806106/total-population-arab-league/> (accessed 11.9.23).
- Ashhad, S., Kam, K., Del Negro, C.A., Feldman, J.L., 2022. *Breathing Rhythm and Pattern and Their Influence on Emotion*. *Annu. Rev. Neurosci.* 45, 223–247.
<https://doi.org/10.1146/annurev-neuro-090121-014424>
- Awajan, I., Mohamad, M., 2018. *A Review on Sentiment Analysis in Arabic Using Document Level*. *Int. J. Eng. Technol.* 7, 128–132. <https://doi.org/10.14419/ijet.v7i3.13.16338>
- Badawi, E.S., Carter, M., Gully, A., 2013. *Modern Written Arabic: A Comprehensive Grammar*. Routledge.
- Badshah, A.M., Ahmad, J., Rahim, N., Baik, S.W., 2017. *Speech Emotion Recognition from Spectrograms with Deep Convolutional Neural Network*, in: 2017 International Conference on Platform Technology and Service (PlatCon). Presented at the 2017 International Conference on Platform Technology and Service (PlatCon), pp. 1–5.
<https://doi.org/10.1109/PlatCon.2017.7883728>
- Bahdanau, D., Cho, K., Bengio, Y., n.d. *Neural Machine Translation by Jointly Learning to Align and Translate*.
- Barker, J.P., Cooke, M.P., Ellis, D.P.W., 2005. *Decoding speech in the presence of other sources*. *Speech Commun.* 45, 5–25. <https://doi.org/10.1016/j.specom.2004.05.002>

- Barrett, L.F., 2006. *Solving the emotion paradox: categorization and the experience of emotion*. Personal. Soc. Psychol. Rev. Off. J. Soc. Personal. Soc. Psychol. Inc 10, 20–46.
https://doi.org/10.1207/s15327957pspr1001_2
- Baum, L.E., Petrie, T., Soules, G., Weiss, N., 1970. *A Maximization Technique Occurring in the Statistical Analysis of Probabilistic Functions of Markov Chains*. Ann. Math. Stat. 41, 164–171. <https://doi.org/10.1214/aoms/1177697196>
- Bengio, Y., Courville, A., Vincent, P., 2014. *Representation Learning: A Review and New Perspectives*. <https://doi.org/10.48550/arXiv.1206.5538>
- Benmamoun, E., Bassiouney, R., 2017. *The Routledge Handbook of Arabic Linguistics*. Routledge.
- Bergstra, J., Bengio, Y., 2012. *Random Search for Hyper-Parameter Optimization*. J. Mach. Learn. Res. 13, 281–305.
- Birjali, M., Kasri, M., Beni-Hssane, A., 2021. *A comprehensive survey on sentiment analysis: Approaches, challenges and trends*. Knowl.-Based Syst. 226, 107134.
<https://doi.org/10.1016/j.knosys.2021.107134>
- Bottou, L., Fogelman Soulie, F., Blanchet, P., Lienard, J.-S., 1989. *Experiments with time delay networks and dynamic time warping for speaker independent isolated digits recognition*, in: First European Conference on Speech Communication and Technology (Eurospeech 1989). Presented at the First European Conference on Speech Communication and Technology (Eurospeech 1989), ISCA, pp. 2537–2540.
<https://doi.org/10.21437/Eurospeech.1989-91>
- Bourlard, H., Morgan, N., 1989. *A Continuous Speech Recognition System Embedding MLP into HMM*, in: Advances in Neural Information Processing Systems. Morgan-Kaufmann.

- Bourlard, H., Wellekens, C.J., 1990. *Links between Markov models and multilayer perceptrons*. *IEEE Trans. Pattern Anal. Mach. Intell.* 12, 1167–1178. <https://doi.org/10.1109/34.62605>
- Bourlard, H.A., Morgan, N., 1994. *Connectionist speech recognition: a hybrid approach*. Springer Science & Business Media.
- Buckwalter, T., Parkinson, D., 2014. *A Frequency Dictionary of Arabic*, 0 ed. Routledge. <https://doi.org/10.4324/9780203883280>
- Buechel, S., Hahn, U., 2016. *Emotion analysis as a regression problem — dimensional models and their implications on emotion representation and metrical evaluation*, in: Proceedings of the Twenty-Second European Conference on Artificial Intelligence, ECAI'16. IOS Press, NLD, pp. 1114–1122. <https://doi.org/10.3233/978-1-61499-672-9-1114>
- Cao, S., Lu, W., Xu, Q., 2016. *Deep Neural Networks for Learning Graph Representations*. Proc. AAAI Conf. Artif. Intell. 30. <https://doi.org/10.1609/aaai.v30i1.10179>
- Carneiro, T., Nóbrega, R.V.M.D., Nepomuceno, T., Bian, G.-B., Albuquerque, V.H.C.D., Filho, P.P.R., 2018. *Performance Analysis of Google Colaboratory as a Tool for Accelerating Deep Learning Applications*. IEEE Access 6. <https://doi.org/10.1109/access.2018.2874767>
- Chen, L., Mao, X., Xue, Y., Cheng, L.L., 2012. *Speech emotion recognition: Features and classification models*. Digit. Signal Process. 22, 1154–1160. <https://doi.org/10.1016/j.dsp.2012.05.007>
- Chiu, C.-C., Sainath, T.N., Wu, Y., Prabhavalkar, R., Nguyen, P., Chen, Z., Kannan, A., Weiss, R.J., Rao, K., Gonina, E., Jaitly, N., Li, B., Chorowski, J., Bacchiani, M., 2018. *State-of-the-Art Speech Recognition with Sequence-to-Sequence Models*, in: 2018 IEEE

- International Conference on Acoustics, Speech and Signal Processing (ICASSP).
Presented at the 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 4774–4778. <https://doi.org/10.1109/ICASSP.2018.8462105>
- Cho, K., van Merriënboer, B., Gulcehre, C., Bahdanau, D., Bougares, F., Schwenk, H., Bengio, Y., 2014. *Learning Phrase Representations using RNN Encoder-Decoder for Statistical Machine Translation*. <https://doi.org/10.48550/arXiv.1406.1078>
- Choubey, V., 2020. *Text classification using CNN*. Voice Tech Podcast. URL <https://medium.com/voice-tech-podcast/text-classification-using-cnn-9ade8155dfb9> (accessed 1.5.23).
- Compound Annual Growth Rate (CAGR) Formula and Calculation* [WWW Document], n.d. . Investopedia. URL <https://www.investopedia.com/terms/c/cagr.asp> (accessed 12.9.23).
- Convolutional Neural Network Architecture for Speech Recognition* [WWW Document], n.d. . ResearchGate. URL https://www.researchgate.net/figure/Convolutional-Neural-Network-Architecture-for-Speech-Recognition_fig1_336819865 (accessed 11.14.23).
- Coppin, G., Sander, D., 2021. Chapter 1 - *Theoretical approaches to emotion and its measurement*, in: Meiselman, H.L. (Ed.), *Emotion Measurement* (Second Edition). Woodhead Publishing, pp. 3–37. <https://doi.org/10.1016/B978-0-12-821124-3.00001-6>
- Cowie, R., Douglas-Cowie, E., Tsapatsoulis, N., Votsis, G., Kollias, S., Fellenz, W., Taylor, J.G., 2001. *Emotion recognition in human-computer interaction*. IEEE Signal Process. Mag. 18, 32–80. <https://doi.org/10.1109/79.911197>
- Cummins, N., Schuller, B.W., 2019. *Latest Advances in Computational Speech Analysis for Mobile Sensing*. Studies in Neuroscience, Psychology and Behavioral Economics 141–159. https://doi.org/10.1007/978-3-030-31620-4_9

- Dang, T., Sethu, V., Epps, J., Ambikairajah, E., 2017. *An Investigation of Emotion Prediction Uncertainty Using Gaussian Mixture Regression*, in: Interspeech 2017. Presented at the Interspeech 2017, ISCA, pp. 1248–1252. <https://doi.org/10.21437/Interspeech.2017-512>
- Daoudi, M., Berretti, S., Pala, P., Delevoye, Y., Bimbo, A.D., 2017. *Emotion Recognition by Body Movement Representation on the Manifold of Symmetric Positive Definite Matrices*, in: International Conference on Image Analysis and Processing, ICIAP 2017. Catania, Italy. https://doi.org/10.1007/978-3-319-68560-1_49
- Darwin, C., 1998. *The Expression of the Emotions in Man and Animals*. Oxford University Press.
- Duwairi, R.M., 2015. *Sentiment analysis for dialectical Arabic*, in: 2015 6th International Conference on Information and Communication Systems (ICICS). Presented at the 2015 6th International Conference on Information and Communication Systems (ICICS), pp. 166–170. <https://doi.org/10.1109/IACS.2015.7103221>
- Efanov, D., Aleksandrov, P., Karapetyants, N., 2022. *The BiLSTM-based synthesized speech recognition*. Procedia Comput. Sci., 2022 Annual International Conference on Brain-Inspired Cognitive Architectures for Artificial Intelligence: The 13th Annual Meeting of the BICA Society 213, 415–421. <https://doi.org/10.1016/j.procs.2022.11.086>
- Eid, M., 1997. *Review of Alif Baa: An introduction to Arabic letters and sounds and al-Kitaab fi Ta'allum al-'Arabiyya: A textbook for beginning Arabic, Part I*. Al-Arab. 30, 117–123.
- Ekman, P., 1992. *An argument for basic emotions*. Cogn. Emot. 6, 169–200. <https://doi.org/10.1080/02699939208411068>
- Ekman, P., Friesen, W.V., 1971. *Constants across cultures in the face and emotion*. J. Pers. Soc. Psychol. 17, 124–129. <https://doi.org/10.1037/h0030377>

- El Ayadi, M., Kamel, M.S., Karray, F., 2011. *Survey on speech emotion recognition: Features, classification schemes, and databases*. Pattern Recognit. 44, 572–587.
<https://doi.org/10.1016/j.patcog.2010.09.020>
- Elman, J.L., 1990. *Finding structure in time*. Cogn. Sci. 14, 179–211.
[https://doi.org/10.1016/0364-0213\(90\)90002-E](https://doi.org/10.1016/0364-0213(90)90002-E)
- El-Masri, M., Altrabsheh, N., Mansour, H., 2017. *Successes and challenges of Arabic sentiment analysis research: a literature review*. Soc. Netw. Anal. Min. 7, 54.
<https://doi.org/10.1007/s13278-017-0474-x>
- Etienne, C., Fidanza, G., Petrovskii, A., Devillers, L., Schmauch, B., 2018. *CNN+LSTM Architecture for Speech Emotion Recognition with Data Augmentation*, in: Workshop on Speech, Music and Mind (SMM 2018). Presented at the Workshop on Speech, Music and Mind (SMM 2018), ISCA, pp. 21–25. <https://doi.org/10.21437/SMM.2018-5>
- Etzioni, O., Cafarella, M., Downey, D., Popescu, A.-M., Shaked, T., Soderland, S., Weld, D.S., Yates, A., 2005. *Unsupervised named-entity extraction from the web: An experimental study*. Artif. Intell. 165, 91–134.
- Facebook users in Egypt - January 2020* [WWW Document], 2020. URL <https://napoleoncat.com/stats/facebook-users-in-egypt/2020/01/> (accessed 11.3.23).
- Feng, H., Golshan, H.M., Mahoor, M.H., 2018. *A wavelet-based approach to emotion classification using EDA signals*. Expert Syst. Appl. 112, 77–86.
<https://doi.org/10.1016/j.eswa.2018.06.014>
- Fujiwara, T., Muller, K., Schwarz, C., n.d. *The Effect of Social Media on Elections: Evidence from the United States*.

- Gales, M., Young, S., 2008. *The Application of Hidden Markov Models in Speech Recognition*. *Found. Trends® Signal Process.* 1, 195–304. <https://doi.org/10.1561/20000000004>
- Gangwar, A., Vadlamani, R., 2022. *A Novel BGCapsule Network for Text Classification*. *SN Comput. Sci.* 3. <https://doi.org/10.1007/s42979-021-00963-4>
- Ghosh, S., Laksana, E., Morency, L.-P., Scherer, S., 2016. *Representation Learning for Speech Emotion Recognition*, in: *Interspeech 2016*. Presented at the Interspeech 2016, ISCA, pp. 3603–3607. <https://doi.org/10.21437/Interspeech.2016-692>
- Girirajan, S., Pandian, A., 2022. *Offline Automatic Speech Recognition System Based on Bidirectional Gated Recurrent Unit (Bi-GRU) with Convolution Neural Network*. *J. Mob. Multimed.* 1659–1676. <https://doi.org/10.13052/jmm1550-4646.1869>
- Graves, A., Jaitly, N., 2014. *Towards End-To-End Speech Recognition with Recurrent Neural Networks*, in: *Proceedings of the 31st International Conference on Machine Learning*. Presented at the International Conference on Machine Learning, PMLR, pp. 1764–1772.
- Gu, S., Wang, W., Wang, F., Huang, J.H., 2016. *Neuromodulator and Emotion Biomarker for Stress Induced Mental Disorders*. *Neural Plast.* 2016, 2609128. <https://doi.org/10.1155/2016/2609128>
- Guellil, I., Saâdane, H., Azouaou, F., Gueni, B., Nouvel, D., 2021. *Arabic natural language processing: An overview*. *J. King Saud Univ. - Comput. Inf. Sci.* 33, 497–507. <https://doi.org/10.1016/j.jksuci.2019.02.006>
- Gupta, N., n.d. *Artificial Neural Network*.
- Gurney, K., 1997. *An Introduction to Neural Networks*. CRC Press.
- Habash, N., Roth, R., Rambow, O., Eskander, R., Tomeh, N., 2013. *Morphological Analysis and Disambiguation for Dialectal Arabic*, in: *Vanderwende, L., Daumé III, H., Kirchhoff, K.*

- (Eds.), Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. Presented at the NAACL-HLT 2013, Association for Computational Linguistics, Atlanta, Georgia, pp. 426–432.
- Habash, N.Y., 2022. *Introduction to Arabic natural language processing*. Springer Nature.
- Haffner, P., 1992. Connectionist word-level classification in speech recognition, in: [Proceedings] ICASSP-92: 1992 IEEE International Conference on Acoustics, Speech, and Signal Processing. Presented at the [Proceedings] ICASSP-92: 1992 IEEE International Conference on Acoustics, Speech, and Signal Processing, IEEE, San Francisco, CA, USA, pp. 621–624 vol.1. <https://doi.org/10.1109/ICASSP.1992.225832>
- Hema, C., Garcia Marquez, F.P., 2023. *Emotional speech Recognition using CNN and Deep learning techniques*. Appl. Acoust. 211, 109492. <https://doi.org/10.1016/j.apacoust.2023.109492>
- Heracleous, P., Mohammad, Y., Yoneyama, A., 2019. *Deep Convolutional Neural Networks for Feature Extraction in Speech Emotion Recognition*, in: Kurosu, M. (Ed.), Human-Computer Interaction. Recognition and Interaction Technologies, Lecture Notes in Computer Science. Springer International Publishing, Cham, pp. 117–132. https://doi.org/10.1007/978-3-030-22643-5_9
- Hermansky, H., 1990. *Perceptual linear predictive (PLP) analysis of speech*. J. Acoust. Soc. Am. 87, 1738–1752. <https://doi.org/10.1121/1.399423>
- Hifny, Y., Ali, A., 2019. *Efficient Arabic Emotion Recognition Using Deep Neural Networks*, in: ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). Presented at the ICASSP 2019 - 2019 IEEE International

- Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 6710–6714.
<https://doi.org/10.1109/ICASSP.2019.8683632>
- Hinton, G., Deng, L., Yu, D., Dahl, G.E., Mohamed, A., Jaitly, N., Senior, A., Vanhoucke, V., Nguyen, P., Sainath, T.N., Kingsbury, B., 2012. *Deep Neural Networks for Acoustic Modeling in Speech Recognition: The Shared Views of Four Research Groups*. IEEE Signal Process. Mag. 29, 82–97. <https://doi.org/10.1109/MSP.2012.2205597>
- Hochreiter, S., Schmidhuber, J., 1997. *Long Short-Term Memory*. Neural Comput. 9, 1735–1780.
<https://doi.org/10.1162/neco.1997.9.8.1735>
- Holes, C., 2004. *Modern Arabic: Structures, Functions, and Varieties*. Georgetown University Press.
- HOW DIFFERENT ARE ARABIC DIALECTS? [WWW Document], n.d. URL
<https://www.etoninstitute.com/blog/different-arabic-dialects> (accessed 11.1.23).
- Hozjan, V., Kačič, Z., 2003. *Context-Independent Multilingual Emotion Recognition from Speech Signals*. Int. J. Speech Technol. 6, 311–320.
<https://doi.org/10.1023/A:1023426522496>
- Jack, R.E., Garrod, O.G.B., Schyns, P.G., 2014. *Dynamic Facial Expressions of Emotion Transmit an Evolving Hierarchy of Signals over Time*. Curr. Biol. 24, 187–192.
<https://doi.org/10.1016/j.cub.2013.11.064>
- Jin, Y., 2017. *Development of Word Cloud Generator Software Based on Python*. Procedia Eng., 13th Global Congress on Manufacturing and Management Zhengzhou, China 28-30 November, 2016 174, 788–792. <https://doi.org/10.1016/j.proeng.2017.01.223>
- Jurafsky, D., Martin, J., 2008. *Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition*.

- Katz, S., 1987. *Estimation of probabilities from sparse data for the language model component of a speech recognizer*. IEEE Trans. Acoust. Speech Signal Process. 35, 400–401.
<https://doi.org/10.1109/TASSP.1987.1165125>
- Kaur, K., Singh, P., 2023. *Trends in speech emotion recognition: a comprehensive survey*. Multimed. Tools Appl. 82, 29307–29351. <https://doi.org/10.1007/s11042-023-14656-y>
- Kaya, H., Gürpınar, F., Salah, A.A., 2017. *Video-based emotion recognition in the wild using deep transfer learning and score fusion*. Image Vis. Comput., Multimodal Sentiment Analysis and Mining in the Wild Image and Vision Computing 65, 66–75.
<https://doi.org/10.1016/j.imavis.2017.01.012>
- Kaye, A.S., 1970. *Modern standard Arabic and the colloquials*. Lingua 24, 374–391.
[https://doi.org/10.1016/0024-3841\(70\)90089-6](https://doi.org/10.1016/0024-3841(70)90089-6)
- Kirchhoff, K., Bilmes, J.A., Das, S., Duta, N., Egan, M., Ji, G., He, F., Henderson, J., Liu, D., Noamany, M., Schone, P., Schwartz, R.M., Vergyri, D., 2003. *Novel approaches to Arabic speech recognition: report from the 2002 Johns-Hopkins Summer Workshop*, in: 2003 IEEE International Conference on Acoustics, Speech, and Signal Processing, ICASSP '03, Hong Kong, April 6-10, 2003. IEEE, pp. 344–347.
<https://doi.org/10.1109/ICASSP.2003.1198788>
- Kneser, R., Ney, H., 1995. *Improved backing-off for M-gram language modeling*, in: 1995 International Conference on Acoustics, Speech, and Signal Processing. Presented at the 1995 International Conference on Acoustics, Speech, and Signal Processing, pp. 181–184 vol.1. <https://doi.org/10.1109/ICASSP.1995.479394>

- Krizhevsky, A., Sutskever, I., Hinton, G.E., 2012. *ImageNet Classification with Deep Convolutional Neural Networks*, in: *Advances in Neural Information Processing Systems*. Curran Associates, Inc.
- Kron, A., Pilkiw, M., Banaei, J., Goldstein, A., Anderson, A.K., 2015. *Are valence and arousal separable in emotional experience?* *Emotion* 15, 35–44.
<https://doi.org/10.1037/a0038474>
- Kuchibhotla, S., Vankayalapati, H.D., Anne, K.R., 2016. *An optimal two stage feature selection for speech emotion recognition using acoustic features*. *Int. J. Speech Technol.* 19, 657–667. <https://doi.org/10.1007/s10772-016-9358-0>
- Landau, M.J., 2008. *Acoustical Properties of Speech as Indicators of Depression and Suicidal Risk*. *Vanderbilt Undergrad. Res. J.* 4. <https://doi.org/10.15695/vurj.v4i0.2783>
- Laput, G., Xiao, R., Harrison, C., 2016. *ViBand: High-Fidelity Bio-Acoustic Sensing Using Commodity Smartwatch Accelerometers*, in: *Proceedings of the 29th Annual Symposium on User Interface Software and Technology, UIST '16*. Association for Computing Machinery, New York, NY, USA, pp. 321–333.
<https://doi.org/10.1145/2984511.2984582>
- Latif, S., Rana, R., Qadir, J., Epps, J., 2020. *Variational Autoencoders for Learning Latent Representations of Speech Emotion: A Preliminary Study*.
<https://doi.org/10.48550/arXiv.1712.08708>
- Lecorvé, G., Motlicek, P., 2012. *Conversion of recurrent neural network language models to weighted finite state transducers for automatic speech recognition*, in: *Interspeech 2012*. Presented at the Interspeech 2012, ISCA, pp. 1668–1671.
<https://doi.org/10.21437/Interspeech.2012-457>

- LeDoux, J.E., 2012. Chapter 21 - *Evolution of human emotion: A view through fear*, in: Hofman, M.A., Falk, D. (Eds.), *Progress in Brain Research, Evolution of the Primate Brain*. Elsevier, pp. 431–442. <https://doi.org/10.1016/B978-0-444-53860-4.00021-0>
- Lee, J., Tashev, I., n.d. *High-level Feature Representation using Recurrent Neural Network for Speech Emotion Recognition*.
- Li, H., Tu, M., Huang, J., Narayanan, S., Georgiou, P., 2021. *Speaker-invariant Affective Representation Learning via Adversarial Training*.
<https://doi.org/10.48550/arXiv.1911.01533>
- Li, Y., Chao, L., Liu, Y., Bao, W., Tao, J., 2015. *From simulated speech to natural speech, what are the robust features for emotion recognition?*, in: 2015 International Conference on Affective Computing and Intelligent Interaction (ACII). Presented at the 2015 International Conference on Affective Computing and Intelligent Interaction (ACII), pp. 368–373. <https://doi.org/10.1109/ACII.2015.7344597>
- Lim, N., 2016. *Cultural differences in emotion: differences in emotional arousal level between the East and the West*. *Integr. Med. Res.* 5, 105–109.
<https://doi.org/10.1016/j.imr.2016.03.004>
- Lippmann, R.P., 1989. *Review of Neural Networks for Speech Recognition*. *Neural Comput.* 1, 1–38. <https://doi.org/10.1162/neco.1989.1.1.1>
- Liu, Z., Huang, J., Wang, Y., Chen, T., 1997. *Audio feature extraction and analysis for scene classification*, in: *Proceedings of First Signal Processing Society Workshop on Multimedia Signal Processing*. Presented at the *Proceedings of First Signal Processing Society Workshop on Multimedia Signal Processing*, pp. 343–348.
<https://doi.org/10.1109/MMSP.1997.602659>

- Lotfian, R., Busso, C., 2018. *Predicting Categorical Emotions by Jointly Learning Primary and Secondary Emotions through Multitask Learning*, in: Interspeech 2018. Presented at the Interspeech 2018, ISCA, pp. 951–955. <https://doi.org/10.21437/Interspeech.2018-2464>
- Luo, D., Zou, Y., Huang, D., 2018. *Investigation on Joint Representation Learning for Robust Feature Extraction in Speech Emotion Recognition*, in: Interspeech 2018. Presented at the Interspeech 2018, ISCA, pp. 152–156. <https://doi.org/10.21437/Interspeech.2018-1832>
- Ma, X., Wu, Z., Jia, J., Xu, M., Meng, H., Cai, L., 2018. *Emotion Recognition from Variable-Length Speech Segments Using Deep Learning on Spectrograms*, in: Interspeech 2018. Presented at the Interspeech 2018, ISCA, pp. 3683–3687. <https://doi.org/10.21437/Interspeech.2018-2228>
- Maamouri, M., Bies, A., Buckwalter, T., Diab, M.T., Habash, N., Rambow, O., Tabessi, D., 2006. *Developing and Using a Pilot Dialectal Arabic Treebank*, in: LREC. pp. 443–448.
- Maaoui, C., Pruski, A., Maaoui, C., Pruski, A., 2010. *Emotion Recognition through Physiological Signals for Human-Machine Communication*, in: Cutting Edge Robotics 2010. IntechOpen. <https://doi.org/10.5772/10312>
- Madanian, S., Chen, T., Adeleye, O., Templeton, J.M., Poellabauer, C., Parry, D., Schneider, S.L., 2023. *Speech emotion recognition using machine learning — A systematic review*. *Intell. Syst. Appl.* 20, 200266. <https://doi.org/10.1016/j.iswa.2023.200266>
- Mahony, N.O., Campbell, S., Carvalho, A., Harapanahalli, S., Velasco-Hernandez, G., Krpalkova, L., Riordan, D., Walsh, J., 2020. *Deep Learning vs. Traditional Computer Vision*. <https://doi.org/10.1007/978-3-030-17795-9>

- Mäntylä, M.V., Graziotin, D., Kuuttila, M., 2018. *The evolution of sentiment analysis—A review of research topics, venues, and top cited papers*. Comput. Sci. Rev. 27, 16–32.
<https://doi.org/10.1016/j.cosrev.2017.10.002>
- Mao, X., Zhang, B., Luo, Y., n.d. *SPEECH EMOTION RECOGNITION BASED ON A HYBRID OF HMM/ANN*.
- Marechal, C., Mikołajewski, D., Tyburek, K., Prokopowicz, P., Bougueroua, L., Ancourt, C., Węgrzyn-Wolska, K., 2019. *Survey on AI-Based Multimodal Methods for Emotion Detection*, in: Kołodziej, J., González-Vélez, H. (Eds.), *High-Performance Modelling and Simulation for Big Data Applications: Selected Results of the COST Action IC1406 cHiPSet*, Lecture Notes in Computer Science. Springer International Publishing, Cham, pp. 307–324. https://doi.org/10.1007/978-3-030-16272-6_11
- Matrane, Y., Benabbou, F., Sael, N., 2023. *A systematic literature review of Arabic dialect sentiment analysis*. J. King Saud Univ. - Comput. Inf. Sci. 35.
<https://doi.org/10.1016/j.jksuci.2023.101570>
- Mcculloch, W.S., Pitts, W., n.d. *A LOGICAL CALCULUS OF THE IDEAS IMMANENT IN NERVOUS ACTIVITY*.
- Medhat, W., Hassan, A., Korashy, H., 2014. *Sentiment analysis algorithms and applications: A survey*. Ain Shams Eng. J. 5, 1093–1113. <https://doi.org/10.1016/j.asej.2014.04.011>
- Meftah, A.H., Qamhan, M., Alotaibi, Y.A., Zakariah, M., 2020. *Arabic Speech Emotion Recognition Using KNN and KSUEmotions Corpus*. Int. J. Simul. Syst. Sci. Technol.
<https://doi.org/10.5013/IJSSST.a.21.02.21>

- Meftah, A.H., Qamhan, M.A., Seddiq, Y., Alotaibi, Y.A., Selouani, S.A., 2021. *King Saud University Emotions Corpus: Construction, Analysis, Evaluation, and Comparison*. IEEE Access 9, 54201–54219. <https://doi.org/10.1109/ACCESS.2021.3070751>
- Mehrabian, A., 1981. *Silent messages: implicit communication of emotions and attitudes*, 2d ed. ed. Wadsworth Pub. Co., Belmont, Calif.
- Mehrabian, A., 1969. *Significance of posture and position in the communication of attitude and status relationships*. Psychol. Bull. 71, 359–372. <https://doi.org/10.1037/h0027349>
- Mikolov, T., n.d. *Recurrent neural network based language model*.
- Miotto, R., Wang, F., Wang, S., Jiang, X., Dudley, J.T., 2018. *Deep learning for healthcare: review, opportunities and challenges*. Brief. Bioinform. 19, 1236–1246. <https://doi.org/10.1093/bib/bbx044>
- Mishra, P., Sharma, R., 2020. *Gender Differentiated Convolutional Neural Networks for Speech Emotion Recognition*, in: 2020 12th International Congress on Ultra Modern Telecommunications and Control Systems and Workshops (ICUMT). Presented at the 2020 12th International Congress on Ultra Modern Telecommunications and Control Systems and Workshops (ICUMT), pp. 142–148. <https://doi.org/10.1109/ICUMT51630.2020.9222412>
- Mishra, S., Bhatnagar, N., P, P., T. R, S., 2023. *Speech emotion recognition and classification using hybrid deep CNN and BiLSTM model*. Multimed. Tools Appl. <https://doi.org/10.1007/s11042-023-16849-x>
- Mishra, S.P., Warule, P., Deb, S., 2023. *Speech emotion recognition using MFCC-based entropy feature*. Signal Image Video Process. <https://doi.org/10.1007/s11760-023-02716-7>
- Mohammed, T.A.E., 2016. *Review of Sentiment Analysis for Classification Arabic Tweets*.

- Moon, T.K., 1996. The expectation-maximization algorithm. *IEEE Signal Process. Mag.* 13, 47–60. <https://doi.org/10.1109/79.543975>
- Moskowitz, J.T., Cheung, E.O., Freedman, M., Fernando, C., Zhang, M.W., Huffman, J.C., Addington, E.L., 2021. *Measuring positive emotion outcomes in positive psychology interventions: A literature review*. *Emot. Rev.* 13, 60–73. <https://doi.org/10.1177/1754073920950811>
- Mourad, A., Darwish, K., 2013. *Subjectivity and Sentiment Analysis of Modern Standard Arabic and Arabic Microblogs*, in: Balahur, A., van der Goot, E., Montoyo, A. (Eds.), *Proceedings of the 4th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*. Presented at the WASSA 2013, Association for Computational Linguistics, Atlanta, Georgia, pp. 55–64.
- Mozetič, I., Torgo, L., Cerqueira, V., Smailović, J., 2018. *How to evaluate sentiment classifiers for Twitter time-ordered data?* *PLoS ONE* 13, e0194317. <https://doi.org/10.1371/journal.pone.0194317>
- Mubarak, H., Hussein, A., Chowdhury, S.A., Ali, A., 2021. *QASR: QCRI Aljazeera Speech Resource A Large Scale Annotated Arabic Speech Corpus*, in: Zong, C., Xia, F., Li, W., Navigli, R. (Eds.), *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*. Presented at the ACL-IJCNLP 2021, Association for Computational Linguistics, Online, pp. 2274–2285. <https://doi.org/10.18653/v1/2021.acl-long.177>

- Mustafa, M.B., Yusoof, M.A.M., Don, Z.M., Malekzadeh, M., 2018. *Speech emotion recognition research: an analysis of research focus*. Int. J. Speech Technol. 21, 137–156.
<https://doi.org/10.1007/s10772-018-9493-x>
- Nelci, J., Salloum, S.A., Shaalan, K., 2022. *An Overview of Sentiment Analysis with Dialectical Processing*, in: Al-Emran, M., Al-Sharafi, M.A., Al-Kabi, M.N., Shaalan, K. (Eds.), Proceedings of International Conference on Emerging Technologies and Intelligent Systems, Lecture Notes in Networks and Systems. Springer International Publishing, Cham, pp. 3–12. https://doi.org/10.1007/978-3-030-82616-1_1
- Nfissi, A., Bouachir, W., Bouguila, N., Mishara, B.L., 2022. *CNN-n-GRU: end-to-end speech emotion recognition from raw waveform signal using CNNs and gated recurrent unit networks*, in: 2022 21st IEEE International Conference on Machine Learning and Applications (ICMLA). Presented at the 2022 21st IEEE International Conference on Machine Learning and Applications (ICMLA), pp. 699–702.
<https://doi.org/10.1109/ICMLA55696.2022.00116>
- Nicholson, J., Takahashi, K., Nakatsu, R., 1999. *Emotion recognition in speech using neural networks*, in: ICONIP'99. ANZIS'99 & ANNES'99 & ACNN'99. 6th International Conference on Neural Information Processing. Proceedings (Cat. No.99EX378). Presented at the ICONIP'99. ANZIS'99 & ANNES'99 & ACNN'99. 6th International Conference on Neural Information Processing. Proceedings (Cat. No.99EX378), pp. 495–501 vol.2. <https://doi.org/10.1109/ICONIP.1999.845644>
- Noroozi, F., Marjanovic, M., Njegus, A., Escalera, S., Anbarjafari, G., 2019. *Audio-Visual Emotion Recognition in Video Clips*. IEEE Trans. Affect. Comput. 10, 60–75.
<https://doi.org/10.1109/TAFFC.2017.2713783>

- Otter, D.W., Medina, J.R., Kalita, J.K., 2021. *A Survey of the Usages of Deep Learning for Natural Language Processing*. IEEE Trans. Neural Netw. Learn. Syst. 32, 604–624.
<https://doi.org/10.1109/TNNLS.2020.2979670>
- Pang, B., Lee, L., n.d. *Opinion mining and sentiment analysis*.
- Parry, J., Palaz, D., Clarke, G., Lecomte, P., Mead, R., Berger, M., Hofer, G., 2019. Analysis of Deep Learning Architectures for Cross-Corpus Speech Emotion Recognition, in: Interspeech 2019. Presented at the Interspeech 2019, ISCA, pp. 1656–1660.
<https://doi.org/10.21437/Interspeech.2019-2753>
- Passricha, V., Aggarwal, R.K., 2020. *A Hybrid of Deep CNN and Bidirectional LSTM for Automatic Speech Recognition*. J. Intell. Syst. 29, 1261–1274.
<https://doi.org/10.1515/jisys-2018-0372>
- Perkel, J.M., 2018. *Why Jupyter is data scientists' computational notebook of choice*. Nature 563, 145–146. <https://doi.org/10.1038/d41586-018-07196-1>
- Petridis, S., Li, Z., Pantic, M., 2017. End-To-End Visual Speech Recognition With LSTMs.
<https://doi.org/10.48550/arXiv.1701.05847>
- Petridis, S., Stafylakis, T., Ma, P., Cai, F., Tzimiropoulos, G., Pantic, M., 2018. *End-to-End Audiovisual Speech Recognition*, in: 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). Presented at the 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 6548–6552.
<https://doi.org/10.1109/ICASSP.2018.8461326>
- Petrushin, V.A., 2000. *Emotion recognition in speech signal: experimental study, development, and application*, in: 6th International Conference on Spoken Language Processing (ICSLP 2000). Presented at the 6th International Conference on Spoken Language

Processing (ICSLP 2000), ISCA, vols. vols. 2, 222-225–0.

<https://doi.org/10.21437/ICSLP.2000-791>

Population Levels, Trends and Policies in the Arab Region: Challenges and Opportunities |

United Nations Development Programme [WWW Document], n.d. URL

<https://www.undp.org/arab-states/publications/population-levels-trends-and-policies-arab-region-challenges-and-opportunities> (accessed 11.2.23).

Pozzi, F.A., Fersini, E., Messina, E., Liu, B., 2016. *Sentiment Analysis in Social Networks*, 1st ed. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA.

Prabowo, R., Thelwall, M., 2009. *Sentiment analysis: A combined approach*. J. Informetr. 3, 143–157. <https://doi.org/10.1016/j.joi.2009.01.003>

Probst, P., Boulesteix, A.-L., Bischl, B., 2019. *Tunability: Importance of Hyperparameters of Machine Learning Algorithms*. J. Mach. Learn. Res. 20, 1–32.

Qasim, H., Abdulbaqi, H.A., 2022. *Arabic speech recognition using deep learning methods: Literature review*. AIP Conf. Proc. 2398, 050029. <https://doi.org/10.1063/5.0094741>

Rabiner, L.R., 1989. *A tutorial on hidden Markov models and selected applications in speech recognition*. Proc. IEEE 77, 257–286. <https://doi.org/10.1109/5.18626>

Ramachandram, D., Taylor, G.W., 2017. *Deep Multimodal Learning: A Survey on Recent Advances and Trends*. IEEE Signal Process. Mag. 34, 96–108.
<https://doi.org/10.1109/MSP.2017.2738401>

Renals, S., Morgan, N., Boulard, H., Cohen, M., Franco, H., 1994. *Connectionist probability estimators in HMM speech recognition*. IEEE Trans. Speech Audio Process. 2, 161–174.
<https://doi.org/10.1109/89.260359>

- Ringeval, F., Schuller, B., Valstar, M., Jaiswal, S., Marchi, E., Lalanne, D., Cowie, R., Pantic, M., 2015. AV+EC 2015: *The First Affect Recognition Challenge Bridging Across Audio, Video, and Physiological Data*, in: Proceedings of the 5th International Workshop on Audio/Visual Emotion Challenge, AVEC '15. Association for Computing Machinery, New York, NY, USA, pp. 3–8. <https://doi.org/10.1145/2808196.2811642>
- Robinson, A.J., 1994. *An application of recurrent nets to phone probability estimation*. IEEE Trans. Neural Netw. 5, 298–305. <https://doi.org/10.1109/72.279192>
- Rubin, A., ER, B., 2009. Research Methods for Social Work.
- Rubin, D.C., Talarico, J.M., 2009. *A Comparison of Dimensional Models of Emotion: Evidence from Emotions, Prototypical Events, Autobiographical Memories, and Words*. Mem. Hove Engl. 17, 802–808. <https://doi.org/10.1080/09658210903130764>
- Rumelhart, D.E., McClelland, J.L., 1987. *Learning Internal Representations by Error Propagation*, in: Parallel Distributed Processing: Explorations in the Microstructure of Cognition: Foundations. Presented at the Parallel Distributed Processing: Explorations in the Microstructure of Cognition: Foundations, MIT Press, pp. 318–362.
- Rybka, J., Janicki, A., 2013. *Comparison of speaker dependent and speaker independent emotion recognition*. Int. J. Appl. Math. Comput. Sci. 23, 797–808.
- Sadat, F., Kazemi, F., Farzindar, A., 2014. *Automatic Identification of Arabic Language Varieties and Dialects in Social Media*, in: Lin, S., Ku, L.-W., Cambria, E., Kuo, T.-T. (Eds.), Proceedings of the Second Workshop on Natural Language Processing for Social Media (SocialNLP). Presented at the SocialNLP 2014, Association for Computational Linguistics and Dublin City University, Dublin, Ireland, pp. 22–27. <https://doi.org/10.3115/v1/W14-5904>

- Sander, D., Scherer, K., 2009. *Oxford Companion to Emotion and the Affective Sciences*. OUP Oxford.
- Sarker, I.H., 2021. *Deep Learning: A Comprehensive Overview on Techniques, Taxonomy, Applications and Research Directions*. *SN Comput. Sci.* 2, 420.
<https://doi.org/10.1007/s42979-021-00815-1>
- Scherer, E. by K.R., Schorr, A., Johnstone, and T. (Eds.), 2001. *Appraisal Processes in Emotion: Theory, Methods, Research, Series in Affective Science*. Oxford University Press, Oxford, New York.
- Schröder, M., Devillers, L., Karpouzis, K., Martin, J.-C., Pelachaud, C., Peter, C., Pirker, H., Schuller, B., Tao, J., Wilson, I., 2007. *What should a generic emotion markup language be able to represent?*
- Schuller, B., Steidl, S., & Batliner, A. (2009). *The Interspeech 2009 Emotion Challenge*. 312–315. <https://doi.org/10.21437/Interspeech.2009-103>
- Schuller, B.W., 2018. *Speech emotion recognition: two decades in a nutshell, benchmarks, and ongoing trends*. *Commun. ACM* 61, 90–99. <https://doi.org/10.1145/3129340>
- Schuster, M., Paliwal, K.K., 1997. *Bidirectional recurrent neural networks*. *IEEE Trans. Signal Process.* 45, 2673–2681. <https://doi.org/10.1109/78.650093>
- Seirafi, M., De Weerd, P., de Gelder, B.L., 2013. *Emotion categorization does not depend on explicit face categorization*. *J. Vis.* 13, 12. <https://doi.org/10.1167/13.2.12>
- Shalan, K., Siddiqui, S., Alkhatib, M., Abdel Monem, A., 2019. *Challenges in Arabic natural language processing*, in: *Computational Linguistics, Speech and Image Processing for Arabic Language*. World Scientific, pp. 59–83.

- Sharma, S., 2021. *Emotion Recognition from Speech using Artificial Neural Networks and Recurrent Neural Networks*, in: 2021 11th International Conference on Cloud Computing, Data Science & Engineering (Confluence). Presented at the 2021 11th International Conference on Cloud Computing, Data Science & Engineering (Confluence), pp. 153–158. <https://doi.org/10.1109/Confluence51648.2021.9377192>
- Shearlaw, M., 2016. *Egypt five years on: was it ever a “social media revolution”?* The Guardian.
- Shen, Y., Yang, H., Lin, L., 2022. *Automatic Depression Detection: an Emotional Audio-Textual Corpus and A Gru/Bilstm-Based Model*, in: ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). Presented at the ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 6247–6251. <https://doi.org/10.1109/ICASSP43922.2022.9746569>
- Shewalkar, A., Nyavanandi, D., Ludwig, S.A., 2019. *Performance Evaluation of Deep Neural Networks Applied to Speech Recognition: RNN, LSTM and GRU*. J. Artif. Intell. Soft Comput. Res. 9, 235–245. <https://doi.org/10.2478/jaiscr-2019-0006>
- Shih, D.-H., Liao, C.-H., Wu, T.-W., Xu, X.-Y., Shih, M.-H., 2022. *Dysarthria Speech Detection Using Convolutional Neural Networks with Gated Recurrent Unit*. Healthc. Basel Switz. 10, 1956. <https://doi.org/10.3390/healthcare10101956>
- Shu, J., Chiu, M., Hui, P., 2019. *Emotion Sensing for Mobile Computing*. IEEE Commun. Mag. 57, 84–90. <https://doi.org/10.1109/MCOM.001.1800834>

- Simonton, K.L., Richards, K.A.R., Washburn, N., 2021. *Understanding Emotion in Physical Education Teaching: A Conceptual Framework for Research and Practice*. Quest 73, 306–322. <https://doi.org/10.1080/00336297.2021.1915352>
- Simonyan, K., Zisserman, A., 2015. *Very Deep Convolutional Networks for Large-Scale Image Recognition*. <https://doi.org/10.48550/arXiv.1409.1556>
- Sivanagaraja, T., Ho, M.K., Khong, A.W.H., Wang, Y., 2017. *End-to-end speech emotion recognition using multi-scale convolution networks*, in: 2017 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC). Presented at the 2017 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC), pp. 189–192. <https://doi.org/10.1109/APSIPA.2017.8282026>
- Staroniewicz, P., 2009. *Recognition of emotional state in Polish speech: comparison between human and automatic efficiency*, in: Proceedings of the 2009 Joint COST 2101 and 2102 International Conference on Biometric ID Management and Multimodal Communication, BioID_MultiComm'09. Springer-Verlag, Berlin, Heidelberg, pp. 33–40.
- Stolar, M.N., Lech, M., Bolia, R.S., Skinner, M., 2017. *Real time speech emotion recognition using RGB image classification and transfer learning*, in: 2017 11th International Conference on Signal Processing and Communication Systems (ICSPCS). Presented at the 2017 11th International Conference on Signal Processing and Communication Systems (ICSPCS), pp. 1–8. <https://doi.org/10.1109/ICSPCS.2017.8270472>
- Swain, M., Routray, A., Kabisatpathy, P., 2018. *Databases, features and classifiers for speech emotion recognition: a review*. Int. J. Speech Technol. 21, 93–120. <https://doi.org/10.1007/s10772-018-9491-z>

The Handbook of Computational Linguistics and Natural Language Processing | Wiley [WWW Document], n.d. . Wiley.com. URL [https://www.wiley.com/en-](https://www.wiley.com/en-us/The+Handbook+of+Computational+Linguistics+and+Natural+Language+Processing-p-9781444324044)

[us/The+Handbook+of+Computational+Linguistics+and+Natural+Language+Processing-p-9781444324044](https://www.wiley.com/en-us/The+Handbook+of+Computational+Linguistics+and+Natural+Language+Processing-p-9781444324044) (accessed 11.5.23).

- Tian, L., Moore, J.D., Lai, C., 2015. *Emotion recognition in spontaneous and acted dialogues*, in: 2015 International Conference on Affective Computing and Intelligent Interaction (ACII). Presented at the 2015 International Conference on Affective Computing and Intelligent Interaction (ACII), pp. 698–704. <https://doi.org/10.1109/ACII.2015.7344645>
- Trigeorgis, G., Ringeval, F., Brueckner, R., Marchi, E., Nicolaou, M.A., Schuller, B., Zafeiriou, S., 2016. *Adieu features? End-to-end speech emotion recognition using a deep convolutional recurrent network*, in: 2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). Presented at the 2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 5200–5204. <https://doi.org/10.1109/ICASSP.2016.7472669>
- Tsai, W.-C., Shih, Y.-J., Huang, N.-T., 2019. *Hardware-Accelerated, Short-Term Processing Voice and Nonvoice Sound Recognitions for Electric Equipment Control*. Electronics 8, 924. <https://doi.org/10.3390/electronics8090924>
- Turney, P.D., 2002. *Thumbs Up or Thumbs Down? Semantic Orientation Applied to Unsupervised Classification of Reviews*. <https://doi.org/10.48550/arXiv.cs/0212032>
- Tzinis, E., Potamianos, A., 2017. *Segment-based speech emotion recognition using recurrent neural networks*, in: 2017 Seventh International Conference on Affective Computing and Intelligent Interaction (ACII). Presented at the 2017 Seventh International Conference on

Affective Computing and Intelligent Interaction (ACII), pp. 190–195.

<https://doi.org/10.1109/ACII.2017.8273599>

Tzirakis, P., Trigeorgis, G., Nicolaou, M.A., Schuller, B.W., Zafeiriou, S., 2017. *End-to-End Multimodal Emotion Recognition Using Deep Neural Networks*. IEEE J. Sel. Top. Signal Process. 11, 1301–1309. <https://doi.org/10.1109/JSTSP.2017.2764438>

Tzirakis, P., Zhang, J., Schuller, B.W., 2018. *End-to-End Speech Emotion Recognition Using Deep Neural Networks*, in: 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). Presented at the 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 5089–5093. <https://doi.org/10.1109/ICASSP.2018.8462677>

Vaz, P.J., Schütz, G., Guerrero, C., Cardoso, P.J.S., 2023. *Hybrid Neural Network Based Models for Evapotranspiration Prediction Over Limited Weather Parameters*. IEEE Access 11, 963–976. <https://doi.org/10.1109/ACCESS.2022.3233301>

Wadley, G., Kostakos, V., Koval, P., Smith, W., Webber, S., Cox, A., Gross, J.J., Höök, K., Mandryk, R., Slovak, P., 2022. *The Future of Emotion in Human-Computer Interaction*, in: Extended Abstracts of the 2022 CHI Conference on Human Factors in Computing Systems, CHI EA '22. Association for Computing Machinery, New York, NY, USA, pp. 1–6. <https://doi.org/10.1145/3491101.3503729>

Waibel, A., Hanazawa, T., Hinton, G., Shikano, K., Lang, K.J., 1989. *Phoneme recognition using time-delay neural networks*. IEEE Trans. Acoust. Speech Signal Process. 37, 328–339. <https://doi.org/10.1109/29.21701>

Watson, J.C.E., 2007. *The Phonology and Morphology of Arabic*. OUP Oxford.

What Is a Convolutional Neural Network? | 3 things you need to know [WWW Document], n.d.

URL <https://www.mathworks.com/discovery/convolutional-neural-network-matlab.html>

(accessed 11.11.23).

Why Muslims are the world's fastest-growing religious group, n.d. . Pew Res. Cent. URL

[https://www.pewresearch.org/short-reads/2017/04/06/why-muslims-are-the-worlds-](https://www.pewresearch.org/short-reads/2017/04/06/why-muslims-are-the-worlds-fastest-growing-religious-group/)

[fastest-growing-religious-group/](https://www.pewresearch.org/short-reads/2017/04/06/why-muslims-are-the-worlds-fastest-growing-religious-group/) (accessed 11.25.23).

Xie, L., Liu, Z., 2006. *A Comparative Study of Audio Features for Audio-to-Visual Conversion in Mpeg-4 Compliant Facial Animation*, in: 2006 International Conference on Machine Learning and Cybernetics. Presented at the 2006 International Conference on Machine Learning and Cybernetics, pp. 4359–4364. <https://doi.org/10.1109/ICMLC.2006.259085>

Yang, Z., Hirschberg, J., 2018. *Predicting Arousal and Valence from Waveforms and Spectrograms Using Deep Neural Networks*, in: Interspeech 2018. Presented at the Interspeech 2018, ISCA, pp. 3092–3096. <https://doi.org/10.21437/Interspeech.2018-2397>

Yin, W., Kann, K., Yu, M., Schütze, H., 2017. *Comparative Study of CNN and RNN for Natural Language Processing*. <https://doi.org/10.48550/arXiv.1702.01923>

Yousif, S.A., Elkabani, I., 2015. The Effect of Combining Different Semantic Relations on Arabic Text Classification.

Zaidan, O.F., Callison-Burch, C., 2014. *Arabic Dialect Identification*. Comput. Linguist. 40, 171–202. https://doi.org/10.1162/COLI_a_00169

Zerari, N., Abdelhamid, S., Bouzgou, H., Raymond, C., 2019. *Bidirectional deep architecture for Arabic speech recognition*. Open Comput. Sci. 9, 92–102.

<https://doi.org/10.1515/comp-2019-0004>

- Zhang, B., Quan, C., Ren, F., 2016. *Study on CNN in the recognition of emotion in audio and images*, in: 2016 IEEE/ACIS 15th International Conference on Computer and Information Science (ICIS). Presented at the 2016 IEEE/ACIS 15th International Conference on Computer and Information Science (ICIS), pp. 1–5.
<https://doi.org/10.1109/ICIS.2016.7550778>
- Zhang, T., Zheng, W., Cui, Z., Zong, Y., Yan, J., Yan, K., 2016. *A Deep Neural Network-Driven Feature Learning Method for Multi-view Facial Expression Recognition*. IEEE Trans. Multimed. 18, 2528–2536. <https://doi.org/10.1109/TMM.2016.2598092>
- Zhang, Z., Geiger, J., Pohjalainen, J., Mousa, A.E.-D., Jin, W., Schuller, B., 2018. *Deep Learning for Environmentally Robust Speech Recognition: An Overview of Recent Developments*. ACM Trans. Intell. Syst. Technol. 9, 49:1-49:28.
<https://doi.org/10.1145/3178115>
- Zhang, Z., Weninger, F., Wollmer, M., Schuller, B., 2011. *Unsupervised learning in cross-corpus acoustic emotion recognition*, in: 2011 IEEE Workshop on Automatic Speech Recognition & Understanding. Presented at the Understanding (ASRU), IEEE, Waikoloa, HI, USA, pp. 523–528. <https://doi.org/10.1109/ASRU.2011.6163986>