REPUBLIC OF TÜRKİYE

ALTINBAŞ UNIVERSITY

Institute of Graduate Studies

Electrical and Computer Engineering

DESIGNING A SYSTEM TO DISTINGUISH AMONGST MAIN ARABIC DIALECTS

Dheyaa Hussein Hammad ALHELAL

Doctor of Philosophy

Supervisor

Asst. Prof. Dr. Timur İNAN

İstanbul, 2024

DESIGNING A SYSTEM TO DISTINGUISH AMONGST MAIN ARABIC DIALECTS

Dheyaa Hussein Hammad ALHELAL

Electrical and Computer Engineering

Doctor of Philosophy

ALTINBAŞ UNIVERSITY

2024

The dissertation titled DESIGNING A SYSTEM TO DISTINGUISH AMONGST MAIN ARABIC DIALECTS prepared by Dheyaa Hussein Hammad ALHELAL and submitted on 05/12/2024 has been **accepted unanimously** for the PhD in Electrical and Computer Engineering.

_____
Asst. Prof. Dr. Timur İNAN
Supervisor

Thesis Defense Committee Members:

Asst. Prof. Dr. Timur İNAN	Department of Software Engineering, Altınbaş University	_____
Prof. Dr. Serhat ÖZEKEŞ	Department of Computer Engineering, Marmara University	_____
Assoc. Prof. Dr. Oğuz ATA	Department of Software Engineering, Altınbaş University	_____
Asst. Prof. Dr. Mesut ÇEVİK	Department of Elektric and Elektronic Engineering, Altınbaş University	_____
Asst. Prof. Dr. Veysel Gökhan BÖCEKÇİ	Department of Elektric and Elektronic Engineering, Marmara University	_____

I hereby declare that this dissertation meets all format and submission requirements for a PhD dissertation.

I hereby declare that all information/data presented in this graduation project has been obtained in full accordance with academic rules and ethical conduct. I also declare all unoriginal materials and conclusions have been cited in the text and all references mentioned in the Reference List have been cited in the text, and vice versa as required by the abovementioned rules and conduct.

Dheyaa Hussein Hammad ALHELAL

Signature

DEDICATION

To those who sacrificed their lives for our comfort

To those who convicted them with my life

To my parents: Hussein Alhelal and Fatehiya Abd

To my wife for her sacrifices throughout my study

To those who support me

To my brothers and my sisters

I dedicate this work to you

PREFACE

I thank and praise the all-mighty god-Allah- for his countless bounties.

I am using this opportunity to thank everyone who helped me throughout this work, especially my family who has suffered for me and my success.

Also, I would like to express my gratitude to everyone who taught me since I was little until I got to this point. I wish to thank my advisor, Dr Timur INAN, for all his guidance, patience, and encouragement and my committee members for their discussion and valuable input. I also like to thank everyone who has helped and supported me even with a simple encouraging word.

Finally, I express my warmest gratitude to the
Ministry of Higher Education and Scientific Research in Iraq, and Altınbaş University for their valuable efforts.

ABSTRACT

DESIGNING A SYSTEM TO DISTINGUISH AMONGST MAIN ARABIC DIALECTS

ALHELAL, Dheyaa Hussein Hammad

Ph.D., Electrical and Computer Engineering, Altınbaş University,

Supervisor: Asst. Prof. Dr. Timur İNAN

Date: 12/2024

Pages: 92

Dialect identification is one of the more recent areas of interest for scholars. This study focuses on distinguishing between well-known Arabic dialects through conversations or speeches. What distinguishes this study from others is the way it approaches the topic, as speech was treated as if it were a sound, such as the sound of birds or even music. In other words, the networks were trained on everyday speech without delving into the details of the language, which represents a strong challenge for researchers. With this approach, all the difficulties and challenges facing researchers are overcome. Here, three Arabic speech patterns are taken into consideration: Levantine Dialect (LD), Egyptian Dialect (ED), and Arabian Peninsula Dialect (APD). To distinguish between the three talking styles, we suggest three models for artificial neural networks. The Multi-Layer Perceptron (MLP), Convolutional Neural Network (CNN), and Deep Recurrent Neural Network (DRNN) networks serve as the foundation for these models. This work presents a thorough analysis of the three suggested models, along with comparisons between them. Spoken Arabic Regional Archive (SARA) dataset is employed. It has been prepared and split up into three sections. The Original SARA (OSARA), Filtered SARA (FSARA), and Mixed SARA (MSARA), which combines the OSARA and FSARA, are these. The suggested DRNN model using the MSARA group of the used dataset has the highest accuracy, 90.70%, according to the results.

Keywords: Arabic Dialects, Convolutional Neural Network, Deep Recurrent Neural Network, Multi-Layer Perceptron, FSARA.

ÖZET

ANA ARAP LEHÇELERİNİ AYIRT ETMEK İÇİN BİR SİSTEM GELİŞTİRİLMESİ

ALHELAL, Dheyaa Hussein Hammad

Doktora, Elektrik ve Bilgisayar Mühendisliği Bölümü, Altınbaş Üniversitesi

Danışman: Dr. Öğr. Üyesi Timur İNAN

Tarih: 12/2024

Sayfa: 92

Lehçe tanımlama, bilim insanları için yeni ilgi alanlarından biridir. Bu çalışma, konuşmalar veya konuşmalar aracılığıyla iyi bilinen Arap lehçeleri arasında ayrım yapmaya odaklanmaktadır. Bu çalışmayı diğerlerinden ayıran şey, konuya yaklaşım biçimidir, çünkü konuşma kuş sesi veya müzik gibi bir sesmiş gibi ele alınmıştır. Başka bir deyişle, ağlar dilin ayrıntılarına dalmadan günlük konuşma konusunda eğitilmiştir, bu da araştırmacılar için güçlü bir meydan okumadır. Bu yaklaşımla, araştırmacıların karşılaştığı tüm zorluklar ve güçlükler aşılmıştır. Burada, üç Arapça konuşma kalıbı dikkate alınmaktadır: Levantine Lehçesi (LD), Mısır Lehçesi (ED) ve Arap Yarımadası Lehçesi (APD). Üç konuşma stilini birbirinden ayırt etmek için yapay sinir ağları için üç model önerilmektedir. Çok Katmanlı Algılayıcı (MLP), Evrişimli Sinir Ağı (CNN) ve Derin Yinelemeli Sinir Ağı (DRNN) ağları bu modeller için temel oluşturmaktadır. Bu çalışma, önerilen üç modelin kapsamlı bir analizini ve bunlar arasındaki karşılaştırmaları sunmaktadır. Konuşulan Arapça Bölgesel Arşivi (SARA) veri kümesi kullanılmıştır. Veriseti üç bölüme ayrılmaktadır. Orijinal SARA (OSARA), Filtrelenmiş SARA (FSARA) ve OSARA ile FSARA'yı birleştiren Karma SARA (MSARA). Sonuçlara göre, kullanılan veri setinin MSARA grubu kullanılarak önerilen DRNN modeli %90,70 ile en yüksek doğruluğa sahiptir.

Anahtar Kelimeler: Arap Lehçeleri, Evrişimli Sinir Ağı, Derin Tekrarlamalı Sinir Ağı, Çok Katmanlı Algılayıcı, FSARA.

TABLE OF CONTENTS

Pages

LIST OF TABLES

Pages

LIST OF FIGURES

Pages

ABBREVIATIONS

1D	:	One-Dimensional
2D	:	Two-Dimensional
AF	:	Activation Function
APD	:	Arabian Peninsula Dialect
ASR	:	Automatic Speech Recognition
Bi-LSTM	:	Bidirectional Long Short-Term Memory
CA	:	Classical Arabic
CNN	:	Convolutional Neural Network
CPU	:	Central Processing Units
DCT	:	Discrete Cosine Transform
DFT	:	Discrete Fourier Transform
DID	:	Dialect Identification
DRNN	:	Deep Recurrent Neural Network
ED	:	Egyptian Dialect
ERB	:	Equivalent Rectangular Bandwidth
FFT	:	Fast Fourier Transform
FIR	:	Finite Impulse Response
FSARA	:	Filtered Spoken Arabic Regional Archive
GMM	:	Gaussian Mixture Model
GPU	:	Graphical Processing Unit
ISTFT	:	Inverse Short-Time Fourier Transform

JSON	:	JavaScript Object Notation
k-NN	:	k-Nearest Neighbours
LD	:	Levantine Dialect
LID	:	Language Identification
LPC	:	Linear Prediction Coding
LPCC	:	Linear Predictive Cepstral Coefficients
LSTM	:	Long Short-Term Memory
MFCC	:	Mel-Frequency Cepstral Coefficients
MGB-3	:	Multi-Genre Broadcast-3rd edition
MLP	:	Multi-Layer Perceptron
MSA	:	Modern Standard Arabic
MSARA	:	Mixed Spoken Arabic Regional Archive
NLP	:	Natural Language Processing
NPC	:	Neural Predictive Coding
OSARA	:	Original Spoken Arabic Regional Archive
PC	:	Personal Computer
PLP	:	Perceptual Linear Prediction
PNCC	:	Power Normalized Cepstral coefficient
QoS	:	Quality of Service
ReLU	:	Rectified Linear Unit
RNN	:	Recurrent Neural Network
SARA	:	Spoken Arabic Regional Archive
SASSC	:	Standard Arabic Single Speaker Corpus
SNN	:	Siamese Neural Network
STFT	:	Short-Time Fourier Transform

SVM : Support Vector Machines

TF : Time-Frequency

LIST OF SYMBOLS

n : Number of Frames

N : Number of Samples per Frame

α : Fixed Value

π : Fixed Value (22/7)

f : A Given Frequency in Hz

1. INTRODUCTION

1.1 SPEECH AND LANGUAGE

1.1.1 Fundamentals of Speech

Speech is the primary method of human communication, consisting of a rich set of sounds through which thoughts, emotions, and complex social signals are conveyed. The speech represents one of the most complex forms of human expression because it represents the complex coordination between cognitive processes and bodily. Speech production is the generation of sounds through the respiratory system and the vibration of the vocal cords while modifying the airflow by the organs responsible for speech, such as the tongue, lips, and palate. Together, these processes produce the smallest units of sound that give meaning in language and are called phonemes [1] [2].

The study of speech sounds is called Phonetics which can be divided into three branches: The first one is articulatory phonetics which deals with the study of how sound is produced. The second branch is acoustic phonetics which deals with studying how sound is transmitted. The last one is auditory phonetics which deals with studying how sound is perceived. The study of how these sounds function within a specific language or dialect to convey meaning is called phonology. The interaction between these phonological elements and phonetics represents the basis of linguistic structure and diversity [3].

Prosody is extremely important for the expression of speech because it includes features such as intonation, stress, and rhythm which in turn help in conveying emotion, emphasis and meaning beyond the literal meaning of the words. The more these prosodic elements are used correctly, the more effective communication will be, which varies greatly depending on languages and dialects [4].

1.1.2 Language Diversity and Dialects

Languages are not static entities, i.e., they evolve over time due to being affected by several factors (cultural, social, and geographical). Dialects are usually formed as a result of this development (Language changes based on a social or regional basis). Each dialect reflects

the identity and unique history of its speakers, which makes a big difference between dialects in their morphology, phonology, lexicon, and syntax of the same language [5] [6].

The study of dialects that provides a broad view of the patterns of change and variation of a language is called dialectology. This science is concerned with researching how dialects emerged and how they interact with each other, in addition to how they reflect identities and social relations. Various applications are mainly based on understanding the variation in dialects such as sociolinguistics, the development of language technologies, and language teaching [7] [8].

Dialectal differences are often exhibited in:

a. Phonology: Differences in the pronunciation of consonants and vowels, i.e., in the sound systems.
b. Morphology: Variations in word formation, including inflectional and derivational processes. Variations in derivational and inflectional processes, i.e., in the word formation.
c. Syntax: Differences in grammatical rules and sentence structure.
d. Lexicon: Unique expressions and vocabulary used in each dialect.

These differences may be the result of social interactions, historical migrations, and geographical isolation, which in turn shape the linguistic landscape of the region [9] [6].

1.2 THE ARABIC LANGUAGE

1.2.1 Historical and Sociolinguistic Context

The Arabic language has a rich history dating back more than a thousand years. It was a major pivot in the religious and cultural life of the Arab world. It gained this importance because it is the language of the Quran and Islamic sciences. Classical Arabic (CA), the form used in the Holy Quran, has remained cohesive and stable over the centuries [4]. Modern Standard Arabic (MSA), originally derived from CA, is the primary lingua franca throughout the Arab world which is used in formal communication, education, media, and literature [8].

Despite the linguistic unity that MSA played in the Arab world, the daily language used in most of the Arab world is regional dialects that differ according to the geographical locations of their speakers. These dialects developed from classical Arabic while being influenced by cultures, local languages, and historical events. The phenomenon of social diglossia, where regional dialects and MSA coexist with distinct functions, is what makes the linguistic landscape of the Arab world distinctive [5] [7].

1.2.2 Classification of Arabic Dialects

In general, Arabic dialects can be classified into four main groups, each with its linguistic features:

1.2.2.1 Levantine dialect

Levantine Dialect is used in countries such as Syria, Lebanon, Palestine, and Jordan. It is characterized by its soft consonants and melodic tones, which may be due to the multiple cultural influences and historical interactions in the Levant [8] [10]. To clarify, some classical Arabic sounds are often simplified in pronunciation in the Levantine dialect, such as /q/ is pronounced as /ʔ/ or /k/ [6].

1.2.2.2 Gulf dialect

Gulf Dialect is spoken in the Arabian Peninsula, such as Saudi Arabia, Qatar, Kuwait, the United Arab Emirates, and Bahrain. This dialect is distinguished from others by the fact that it uses emphatic consonants and its preference for lengthening vowels. Nomadic traditions in the Gulf region and maritime trade are reflected in the linguistic features of the Peninsula dialect [8] [11].

1.2.2.3 Egyptian dialect

Egyptian Dialect is the most widespread and understood dialect in the Arab world due to the popularity of Egyptian media, which often uses the Cairene dialect, in the Arab world. The Egyptian dialect is distinguished by its use of a special set of phonetic and morphological patterns, such as the pronunciation of the classical /q/ as /ʔ/ and the patterns of verb conjugation that are unique to it [8] [10]. The spread of understanding of the Egyptian dialect

throughout the Arab world can be attributed to the popularity of Egyptian cinema, television, and music throughout the Arab world [12].

1.2.2.4 Maghrebi dialect

Maghrebi Dialect is used in North African countries, such as Algeria, Tunisia, and Morocco. It is known for its great influence from Berber and French, which as a result leads to its own phonetic, grammatical, and lexical features that distinguish it from other dialects. For instance, the Maghrebi dialect often contains words from Berber and French and shows complex combinations of consonants [11] [10].

Each language has a distinct linguistic identity represented by a group of dialects, which are usually formed by the influence of social, cultural, and historical factors. Mutual understanding between these dialects differs, with some being closer than others and vice versa. The diversity in the dialects poses challenges to communication between users of these dialects and language learning in general, in addition to the development of language technologies [13].

1.3 LINGUISTIC AND COMPUTATIONAL CHALLENGES

1.3.1 Linguistic Complexity of Dialects

The diversity of Arabic dialects and their informal nature create major challenges for computational applications and linguistic research. These challenges come from the lack of formal orthography in dialects, in addition to the great diversity in spoken and written forms, unlike MSA, which benefits from uniform grammar and vocabulary [6]. Consequently, these challenges complicate tasks like dialect identification, speech recognition, and text normalization [14].

1.3.1.1 Phonological variation

Phonological variation among Arabic dialects represents one of the major challenges. This variation is represented in pronunciation, consonant combinations, and vowel length. For example, the pronunciation of the classic letter /q/ differs between dialects; it is pronounced in the Egyptian dialect as /ʔ/, in the Levantine dialect as /q/, and in the Gulf dialect as /g/ [6]

[10]. This variation or phonetic differences will have an impact on speech recognition systems that need to consider multiple pronunciations and accents.

1.3.1.2 Morphological variation

Morphological differences can be defined as variations in word formation and inflection. Dialects often modify or simplify the complex morphological rules of MSA, producing unique noun pluralization patterns and verb conjugations for that dialect. For example, there is a different pattern when compared to MSA that the Egyptian dialect uses to form the plural, which poses a challenge for morphological analysis and processing [10] [14].

1.3.1.3 Syntactic variation

Syntactic variation is represented by differences in word order and sentence structure in general. Dialects may introduce new syntactic constructions that are not present in MSA, which has an impact on how sentences are formed and interpreted. To illustrate, word order in the Maghrebi dialect is often different from what it is in MSA, with a preference for placing verbs at the end of sentences [8] [10].

1.3.1.4 Lexical variation

Lexical variation involves differences in vocabulary. Dialects often include words or expressions unique to a particular dialect for shared common concepts. This lexical diversity results from regional innovations and borrowing from other languages. For example, the Maghrebi dialect incorporates many French and Berber terms, while the Gulf dialect includes many loanwords from Persian and English [8] [11].

1.3.2 Computational Challenges

The big challenges facing Natural Language Processing (NLP) and computational linguistics are the diversity of Arabic dialects and their informality.

1.3.2.1 Text normalization

One of the reasons for the complexity of text normalization is the lack of standardized spelling among dialects. Text normalization is considered an essential and decisive step for tasks like machine translation and sentiment analysis. Therefore, it often requires using

sophisticated algorithms to standardize abbreviations, informal spellings, and code-switching that are used by dialects [15] [16].

1.3.2.2 Speech recognition

The diversity of Arabic dialects in pronunciation, accents, and speaking styles must be considered in building any system for Automatic Speech Recognition (ASR). Building a robust and efficient ASR system requires large collections of annotated speech representing all dialects, which is a major challenge due to its scarcity [15].

1.3.2.3 Dialect identification

The task of dialect identification is to differentiate among different Arabic dialects in both written and spoken forms. What makes this task difficult, and challenging is the great diversity within the dialects on one side and the overlapping features between the dialects on the other. For dialect identification systems to be effective, it must pick up contextual information and subtle linguistic cues [17] [15].

1.3.2.4 Machine translation

The challenges facing machine translation systems for Arabic dialects are translating regional and informal expressions into the intended language accurately. The training of translation models will be complicated due to the lack of parallel dialect groups, which requires innovative methods to deal with dialectal data [18] [19].

1.4 ARABIC DIALECT IDENTIFICATION

1.4.1 Significance and Application

Various fields will be significantly impacted by accurate Arabic dialect identification:

a. Sociolinguistics: Understanding dialect diversity in the Arabic language helps in the study of social identity, language evolution, and patterns of communication within the Arab world [5] [6].
b. Language Technologies: The performance of NLP applications (machine translation, ASR, and sentiment analysis)is enhanced by dialect identification, which allows systems to deal with dialect-specific data more effectively [15].

c. Education: Dialect identification has an important role in supporting language learning by providing dedicated resources for specific dialects that help learners understand linguistic regional differences [20].

1.4.2 Approaches and Techniques

1.4.2.1 Linguistic features

Linguistic features utilized for dialect identification involve morphological, phonological, lexical, and syntactic characteristics. Morphological features deal with word formation rules while phonological features may include specific consonant and vowel sounds. Lexical features involve dialect-specific vocabulary, and syntactic features include sentence structure [10] [14].

1.4.2.2 Machine learning

Machine learning methods for dialect identification include training classifiers on predefined datasets. Feature extraction techniques, such as tf-idf and n-grams, are used to pick up linguistic patterns, while models, like neural networks and Support Vector Machines (SVM) classify the input data, depending on these features [15].

1.4.2.3 Deep learning

Deep learning methods (Convolutional Neural Networks (CNN) and Recurrent Neural Networks (RNN)) use massive quantities of data to develop intricate representations of dialectal features. These models have demonstrated encouraging outcomes in managing the diversity and intricacy of dialectal data [21].

1.4.2.4 Hybrid approaches

Hybrid approaches incorporate machine learning with linguistic insight techniques to enhance accuracy. For example, combining handcrafted features and linguistic rules into machine learning techniques can improve their ability to pick up dialect-specific nuances [17] [19].

1.5 MOTIVATION

The motivation behind developing a system to differentiate between the main Arabic dialects is stemming from three factors:

a. Linguistic Research: A reliable dialect identification system can supply worthy data for studying sociolinguistic dynamics, language diversity, and evolution within the Arab world [5] [6].
b. Technological Advancements: Enhancement of NLP applications' ability to treat dialectal data can improve the user experience and expand the scope of language technologies, involving ASR, sentiment analysis, and machine translation [15].
c. Educational Support: Thoroughly understanding of dialectal variations can help guide language teaching materials and methodologies, Support for learners in dealing with the linguistic variation of the Arab world [20].

1.6 OBJECTIVES

Because of the Challenges of distinguishing among Arab dialects, it seems appropriate to achieve a system that utilizes artificial intelligence to handle such issues. Therefore, the objective of this work is to design a machine learning system to identify or recognize among the three main Arabic dialects Egyptian Dialect (ED), Arabian Peninsula Dialect (APD), and Levantine Dialect (LD). Many contributions are provided in this work which can be highlighted as follows:

a. Collecting and preparing the dataset of the Spoken Arabic Regional Archive (SARA) [22].

b. Proposing three artificial neural network models based on the Multi-Layer Perceptron (MLP) network, Convolutional Neural Network (CNN), and Deep Recurrent Neural Network (DRNN) to differentiate among the three main Arabic dialects.

c. Providing comprehensive study here including establishing comparisons between the three proposed models with three datasets.

2. GENERAL INFORMATION

2.1 HISTORICAL AND SOCIOLINGUISTIC CONTEXT OF ARABIC

2.1.1 Evolution and Classification

A rich tapestry of Arabic dialects has been created as a result of the development of the Arabic language, and each of these dialects has its linguistic features. MSA is a product of the development of CA, which is used as the official written language throughout the Arab world [4]. As time goes by, the spoken varieties have diverged due to social and regional influences which referred to collectively as Arabic dialects [8].

Main Dialect Groups

a. Levantine Dialect: Used in Syria, Lebanon, Palestine, and Jordan. It has been influenced by migrations in the Eastern Mediterranean region and historical interactions [23] [24].
b. Gulf Dialect: The Arabian Peninsula Dialect, involving Saudi Arabia, Qatar, Kuwait, the United Arab Emirates, and Bahrain. This Arabic Dialect group is characterized by lexical variations and phonetic traits. It has been influenced by maritime trade and Bedouin traditions [25] [26].
c. Egyptian Dialect: Spoken in Egypt, especially Cairo, this dialect has become widespread due to Egypt's dominance of the Arab world in the media and culture [27] [28].
d. Maghrebi Dialect: Used in North Africa, including Morocco, Libya, Tunisia, and Algeria. It is combining elements from French languages and Berber, leading to syntactic and distinct phonological features [5].

2.1.2 Sociolinguistic Factors

The Arabic language is characterized by its coexistence with being diglossic, as the MSA is used for formal communication, while regional dialects are used for daily life [7]. The diglossia in the Arabic language affects language utilization throughout various contexts, affected by factors such as social class, education, and media exposure [14].

Migration and urbanization have contributed to the growth of new linguistic forms and dialect levelling, reflecting interactions and social changes across and within Arab-speaking communities [29] [30].

2.2 COMPUTATIONAL APPROACHES TO DIALECTS IDENTIFICATION

2.2.1 Linguistic Features and Manual Annotation

Linguistic features (morphology, phonology, syntax) have been the focus of researchers on dialect identification in early methods. Where differences among dialects and characteristic patterns are identified by manually annotating corpora by researchers [10] [31].

Morphological variations involve noun formations and verb conjugations, in addition to providing clues for dialect classification [32]. Phonological differences are used as distinguishing markers, like the classical /q/ sound that is pronounced in different ways depending on the dialect[33].

These are two examples with some details for this approach:

In 2018, Najafian et al. examined some dialect identification techniques for the Arabic language, such as n-gram and i-vector phonotactic features with the SVM classifier. The work verifies the usefulness of using CNN for directly mapping acoustic and phonotactic features to one of the Arabic dialects. In addition, a new phonotactic dialect identification technique with a new feature representation methodology was suggested. It acquired phone duration, probability statistics, and phone sequences. The suggested technique achieved 73.27% accuracy after using a mixture of multi-lingual phonotactics of a group of languages involving English, Arabic, Hungarian, Czech, and Russian. Confusion matrix was exploited to examine the dialect identification errors. The suggested phonotactic technique achieved 24.7% and 19% relative error rate reduction as the final system fusion [34].

In the same year, Shon et al. suggested an end-to-end dialect identification method with a Siamese Neural Network (SNN) to elicit language embedding. Acoustic and linguistic features were used. Three types of acoustic features were utilized to train the end-to-end system. Acoustic features were the log mel-scale filter bank energies, Mel-Frequency Cepstral Coefficients (MFCCs), and spectrogram energies. The linguistic feature research

focused on learning similarities and dissimilarities between dialects by using SNN. Therefore, system performance is improved in addition to reducing the feature dimensionality. The final result achieved 73% accuracy when the system utilized a single feature set, while 78% accuracy was achieved when multiple feature sets were used. The end-to-end system accomplished a good result compared to the SNN with language embeddings [35].

2.2.2 Machine Learning Approaches

Automating feature extraction and classification in machine learning methods has improved dialect identification. Common approaches involve:

a. Feature Extraction: Methods such as tf-idf, n-grams, and phonetic transcriptions pick up linguistic patterns. N-grams are particularly beneficial for capturing word sequences and character [36] [37].
b. Classification Algorithms: Algorithms like decision trees, SVM, and k-Nearest Neighbours (k-NN) classify text depending on extracted features [17] [38].

These are two examples with more details for dialect identification that used the machine learning approach:

In 2015, Ali *et al.* applied approaches for dialect identification of Arabic speech. These approaches are based on lexical and phonetic features that are extracted from bottleneck features. The suggested speech recognition system is based on the i-vector framework. Then, two classifiers which are generative and discriminative were examined. Multi-class SVM was utilized to gather between the two classifiers. When the features were tested by utilizing a binary classifier to differentiate between MSA and dialect Arabic, the accuracy was 100%. The suggested methods were used to distinguish among the Arabic dialects (Egyptian, Gulf, Levantine, Maghrebi, and MSA), and the accuracy of the final result was recorded as 59.2% [39].

In 2020, Hanani and Naser acclimatized the X-vector framework for Arabic dialect identification. At the same time, the phonetic features, traditional i-vectors, Gaussian Mixture Model (GMM)-tokens, bottleneck features, and word transcriptions were also utilized. The X-vector achieved performance improvements of 16.60%. When it is combined

with the bottleneck, i-vectors features and word uni-gram performance has been further improved to 69.70% [40].

2.2.3 Deep Learning Approaches

Deep learning has shown important promise in dialect identification by using large datasets to learn complex patterns. Key techniques involve:

a. Convolutional Neural Networks (CNN): Efficient for capturing spatial hierarchies in text, CNN is utilized to extract features from words or sequences of characters [41] [42].
b. Recurrent Neural Networks (RNN): Appropriate for sequential data, RNN and its variants, like Long Short-Term Memory (LSTM) networks, capture contextual information and temporal dependencies [43] [44].
c. Transformers: Transformers, with its attention mechanisms, have further improved the models of complex dependencies in dialectal data [45] [46].

Two examples with more details for dialect identification that used the deep learning approach are presented below:

In 2022, Alsayadi *et al.* suggested a hybrid acoustic model that combines the CNN and LSTM. Attention-based encoder-decoder methods were used to construct the model. Two datasets of the Standard Arabic Single Speaker Corpus (SASSC) and Multi-Genre Broadcast-3rd edition (MGB-3) were utilized. The suggested model obtained a 57.02% word error rate [47].

In 2023, Nasr *et al.* proposed a new transcribed corpus for Yamani, Jordanian, and multi-dialectal Arabic. Numerous baselines deep neural models have been designed like Bidirectional LSTM (Bi-LSTM) and DeepSpeech2 with attention model. Yamani speech corpus obtained a word error rate of 59% by using the Bi-LSTM with attention model. The Jordanian speech corpus obtained a word error rate of 83% by using the same model. The multi-dialectal Jordanian, Yamani, and Arab speech corpus obtained a word error rate of 53% also by using the same model. The DeepSpeech2 model obtained the best performance, where the model obtained 31%, 68%, and 30% for Yamani, Jordanian and multi-dialectal Arabic corpus [48].

2.2.4 Hybrid Approaches

Combining machine learning with linguistic insights has improved the performance of dialect identification systems. Hybrid approaches are hiring automated machine learning methods along with handcrafted linguistic features to pick up the nuances of dialect [49] [37].

This is an example of using the hybrid approach for dialect identification:

In 2017, Shon proposed an Arabic dialect identification system to distinguish among the main Arabic dialects. Dialect variability among the training and testing domains and domain mismatches have been focused on. To build an efficient identification system, SNN was used for understanding similarities and dissimilarities between Arabic dialects. i-vector post-processing was utilized to set domain mismatches. At the same time, both linguistic and acoustic features have been exploited. The system obtained 75% accuracy for 10-hours test set [50].

2.3 CHALLENGES IN ARABIC DIALECT PROCESSING

2.3.1 Data Scarcity and Annotation

One of the most important challenges facing Arabic dialects is the scarcity of annotated data. What makes the process of collecting and annotation of corpora a complex one is the diversity of spelling conventions and the lack of uniform orthography [51] [52].

Efforts to handle this challenge involve: creating a large-scale corpora and utilizing crowdsourcing for data annotation [19].

2.3.2 Text Normalization and Standardization

Text normalization is critical for addressing dialectal diversity, involving code-switching and informal spellings between dialects and MSA. Methods like machine learning models and rule-based methods are utilized to standardize text for NLP applications [53] [18].

2.3.3 Phonological and Morphological Variation

Morphological and phonological diversity between dialects creates challenges for text and speech processing. diversity in word formation and pronunciation influences the

performance of systems that are designed for applications such as part-of-speech tagging and speech recognition [6] [10].

2.4 APPLICATIONS OF DIALECT IDENTIFICATION

2.4.1 Speech Recognition and Synthesis

Dialect identification improves speech synthesis systems and Automatic Speech Recognition (ASR). This is done by allowing these applications to adapt to pronunciations and regional accents. In addition to the enhancement of the naturalness of synthesized speech and transcription accuracy [54] [55].

2.4.2 Machine Translation

Dialect identification is useful for machine translation systems by addressing region-specific and informal expressions accurately, leading to enhanced translation quality [56] [57].

2.4.3 Sentiment Analysis

Dialect identification is very important for sentiment analysis applications that provide an accurate interpretation of sentiments expressed in various dialects. This is especially significant for customer reviews, analysis of social media, and other informal texts [58] [59].

2.4.4 Educational Tools

Educational tools benefit from dialect identification to provide tailored learning resources that help learners understand regional linguistic variations and improve language proficiency [60] [61].

2.4.5 Sociolinguistic Research

Dialect identification accurately enhances sociolinguistic research by providing data for analyzing language diversity, evolution, and social identity in the Arab world. This makes studying dialectal variations and trends easier [5] [62].

3. METHODOLOGY

This study aims to design and implement a machine learning system that can identify or distinguish among the three main Arabic dialects of APD, ED, and LD. To achieve this mission a schematic procedure is proposed. It includes multiple stages starting with collecting the Mixed SARA (MSARA) dataset, which involves both the Original SARA (OSARA) and Filtered SARA (FSARA) sets. After that, pre-processing tasks only for the OSARA are executed. Consequently, the MFCC feature extraction is utilized. Subsequently, all datasets are divided into training and testing groups. For the training group, the three suggested machine learning models of MLP, CNN, and DRNN are trained, and their weights are obtained. For the testing group, all the models are tested after obtaining the resulting training weights. All testing results of all models are recorded. After that, comparisons and evaluations are applied to all of them. The schematic overview for the suggested system of distinguishing among the dialects of APD, ED, and LD is demonstrated in Figure 3.1.

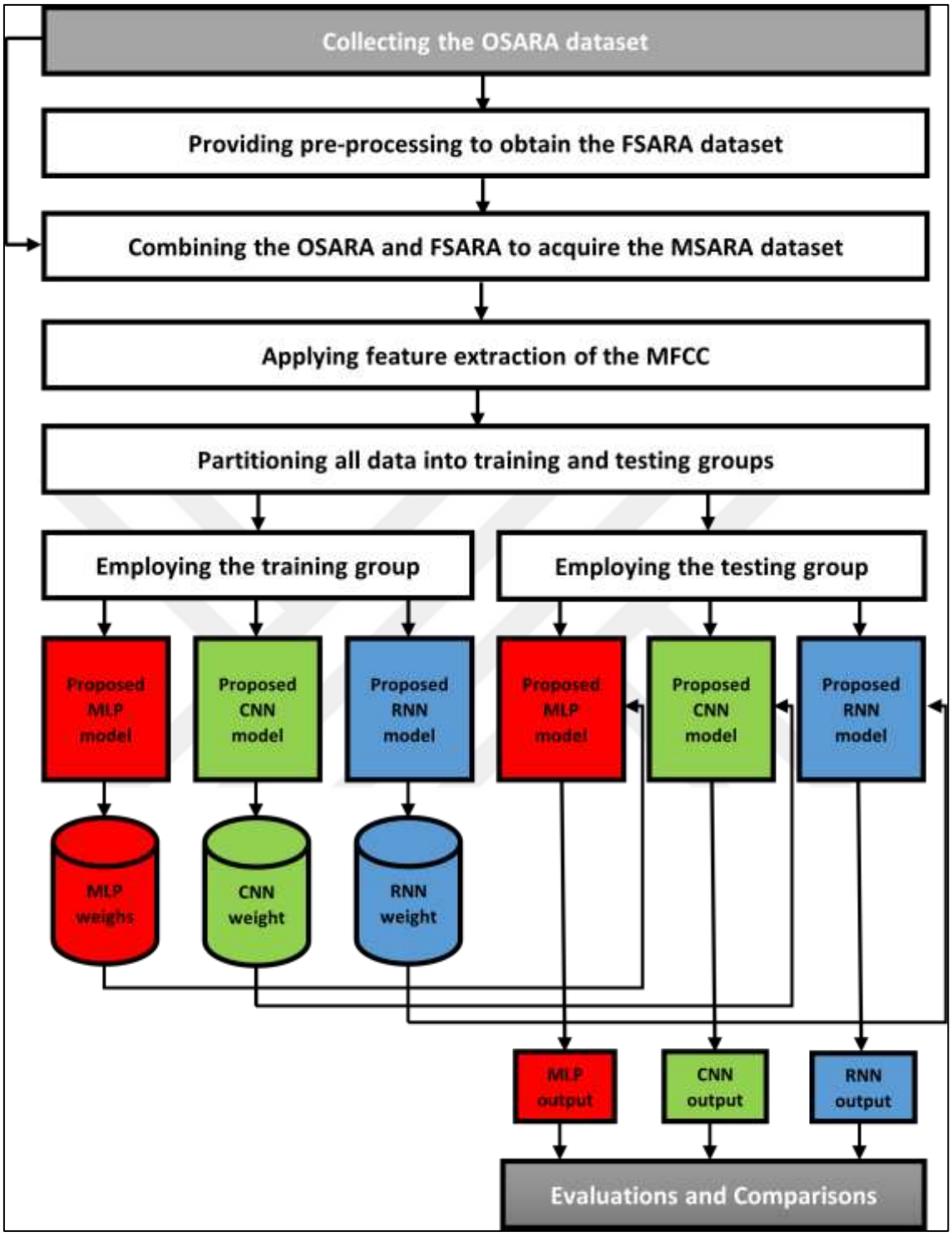

Figure 3.1: Schematic Overview for the Suggested System of Distinguishing Among the Dialects of APD, ED and LD.

3.1 COLLECTING THE OSARA DATASET

The researcher in [22] proposed a dataset involving spontaneous and undefined phrases (spontaneous conversations related to everyday life). This dataset has been obtained from films, series, and acting scenes in three different Arabic dialects. It is defined by the Spoken Arabic Regional Archive (SARA). The dataset is not affected by live transmission Quality of Service (QoS) issues like delay and unreliable packet delivery. These issues are avoided when the media for SARA is downloaded from YouTube and then stored before the sampling process. The SARA dataset is beneficial for many machine learning fields like automatic speech recognition, speaker verification, and accent recognition. In order to avoid wrong pronunciations from children, which influence the detection process, the SARA dataset includes only adult speakers (males and females) for three Arabic dialects: APD, ED, and LD. To add variety to the database, each group of dialects consists of several different accents. Each utterance contains one speaker, while the number of speakers and the words in the same group of dialects are unknown. To illustrate, SARA dataset specifications are described in the following tables [22]:

Table 3.1: Number of Participants in the SARA Dataset.

.

Dialect	No. of Males	No. of Females	Total No. of Participants
ED	877	488	1365
APD	657	573	1230
LD	583	662	1245

Table 3.2: Recording Times and Number of Samples in the SARA Dataset.

Recording Time	No. of Samples in the ED	No. of Samples in the APD	No. of Samples in the LD	Total No. of Samples
3 sec	468	388	398	1254
4 sec	377	409	354	1140
5 sec	271	269	246	786
6 sec	166	117	158	441
7 sec	83	47	89	219
Total	1365	1230	1245	3840

Table 3.3: Considered Accents with Number of Speakers in the SARA Dataset.

Dialect	Accent	No. of Male Speakers	No. of Female Speakers
APD	**Saudi**	143	207
	Hejazi	29	0
	Najdi	8	4
	Bahrani	335	282
	Gulf	64	46
	Omani	78	38
LD	**Syrian**	46	261
	Lebanese	76	341
	Jordanian	362	42
	Palestinian	99	18
ED	**Egyptian**	623	221
	Sa'idi	254	267

These tables show the comprehensiveness of the SARA dataset. So, it is precious to be exploited in this study. However, it has been widely considered and prepared. It has been called in this work the Original SARA (OSARA) dataset.

3.2 DATASET PREPARATION

In actual reality (practical reality), the audio signals are not pure. In fact, it is a mixture between the desirable and undesirable (background music, noise, babble speech, …, etc.) signals. Therefore, filtering out such a mixture of signals at the processing stage has become essential. Generally, Deep Neural Network (DNN) was utilized to enhance and purify the audio signal [63] [64]. It should be noted that deep filtering was first proposed for cleaning and enhancing audio signals by researchers in [65] [66]. One of the benefits of using the deep filter is its capability to restore signal degradation, such as notch-filters or time-frame zeroing, as the filter is utilized for multiple Time-Frequency (TF) bins [65] [66]. In order to product a speech enhancement framework with cogent performance the DeepFilterNet was proposed. It has two-phase speech enhancement frameworks. Because of, the fact that both

noise and speech have a smooth spectral envelope, the first phase was taken this advantage into account. In the second phase, the enhancement (improving) focused on the lower frequencies only; because these frequencies in the periodic speech components have the most energy [67]. After that, DeepFilterNet2 has been proposed as an expanded work by enhancement the network structure, data augmentation, and training procedure [68]. Figure 3.2 illustrates a schematic overview of the DeepFilterNet2 speech enhancement process.

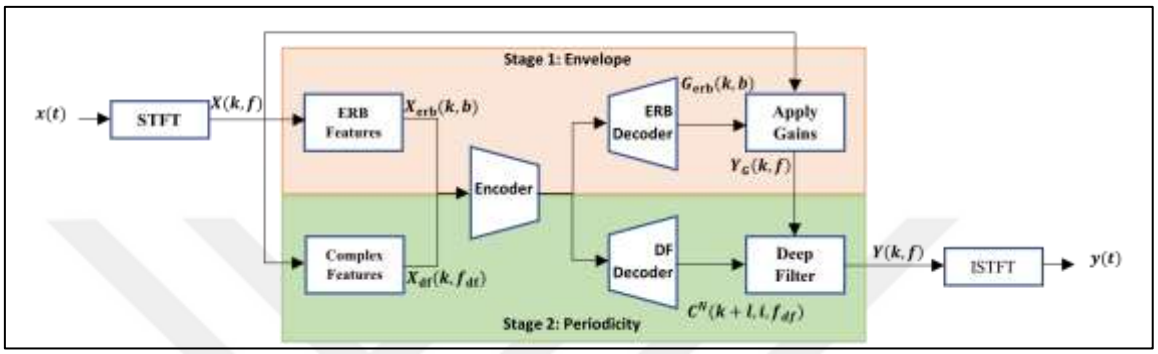

Figure 3.2: Schematic Overview of the DeepFilterNet2 Speech Enhancement Process [68].

From the figure, it can be noticed that an original signal $x(t)$ inputs to the DeepFilterNet2. It has been processed by the Short-Time Fourier Transform (STFT) where $X(k,f)$ has been produced. After that, the $X(k,f)$ produced has been passed to the Equivalent Rectangular Bandwidth (ERB) Features in stage 1 (Envelope), where $X_{\mathrm{erb}}(k,b)$ has been generated. It has been also passed to the Complex Features in stage 2 (Periodicity), where $X_{\mathrm{df}}(k,f_{\mathrm{df}})$ has been generated. Both $X_{\mathrm{erb}}(k,b)$ and $X_{\mathrm{df}}(k,f_{\mathrm{df}})$ have been passed through Encoder. The output of the Encoder goes to the ERB Decoder in stage 1, where $G_{\mathrm{erb}}(k,b)$ has been produced, and DF Decoder in stage 2, where $C^N(k+l,i,f_{\mathrm{df}})$ has been generated. Both $X(k,f)$ and $G_{\mathrm{erb}}(k,b)$ passed to Apply Gains in stage 1, where $Y_{\mathrm{G}}(k,f)$ has been provided. Then, $Y_{\mathrm{G}}(k,f)$ and $C^N(k+l,i,f_{\mathrm{df}})$ has been passed through Deep Filter in stage 2, where $Y(k,f)$ has been produced. Finally, $Y(k,f)$ has been processed by the Inverse Short-Time Fourier Transform (ISTFT) to generate $y(t)$.

DeepFilterNet2 is utilized in this work, where the Filtered SARA (FSARA) dataset has been derived from the OSARA dataset. Nevertheless, both two groups (OSARA and FSARA) datasets have been exploited and concatenated here as one group named the Mixed SARA (MSARA) dataset. The relationship among the three groups of the SARA dataset has been demonstrated as given in Figure 3.3.

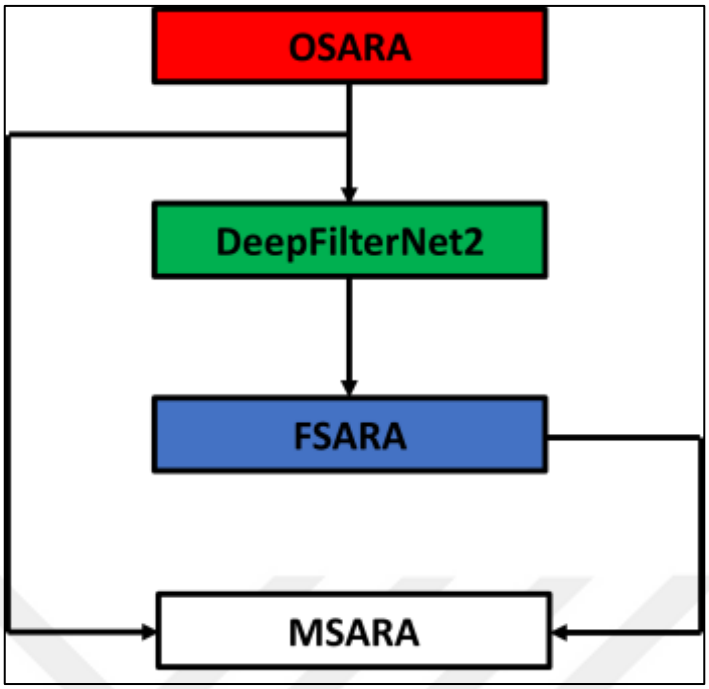

Figure 3.3: Preparing the MSARA Dataset, Which Has Been Utilized in This Work.

3.3 FEATURE EXTRACTION

Although Language Identification (LID) or Dialect Identification (DID) have been started decades ago, it still represents a challenge to researchers. This is because of the acoustic phonetics, similarities of phonotactics, and prosodic cues between different dialects of languages. In addition to that, in many cases of practical application, lack of control over the channel characteristics, noise level, and available speech duration [69]. To the best of acquired knowledge, most research on recognizing Arabic dialects has not exceeded 78% accuracy. So, the goal of this study is to dominate the challenges and develop a system (design and implement) with high performance (high accuracy).

Commonly in any system, extracting benefitable features from an input is significant. There are many techniques utilized for feature extraction Linear Prediction Coding (LPC), Linear Predictive Cepstral Coefficients (LPCC), Perceptual Linear Prediction (PLP), Power Normalized Cepstral coefficient (PNCC), Neural Predictive Coding (NPC), and Mel-Frequency Cepstral Coefficients (MFCC). The MFCC is considered the most popular technique used for audio recognition due MFCC depends on known differences in the critical frequency range of the human ears [70]. Where authors in [71] proved the superiority of MFCC when compared to PNCC. A block diagram of the total processes that are utilized for the MFCC feature extraction is given in Figure 3.4 [72]:

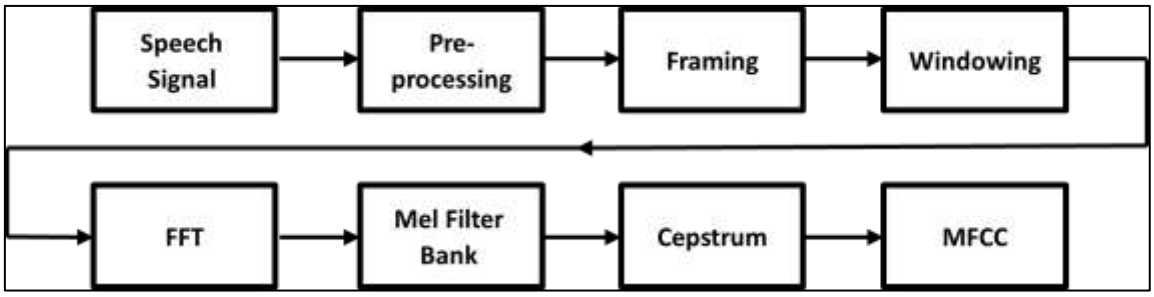

Figure 3.4: Block Diagram of the Total Processes That Are Utilized for the Mfcc Feature Extraction.

The first process is pre-processing. Of course, the entrance audio signal is analog, so it must be changed to digital. So, a set of samples $s(k)$ is acquired depending on the sampling rate. Then, the $s(k)$ signal has been passed through a first-order Finite Impulse Response (FIR) filter to rise its energy (magnitudes in the spectrum of the frequencies).

After that, framing process has been applied since the analog signal has been changed to N number of samples per frame. The neighbouring frames are disjointed by M number of samples per other frame and $(M < N)$, where the overlapping among the 1st frame and 2nd frame is $(N - M)$ number of samples. Each frame is individually synthesized and analyzed to protect information from loss (preventing discontinuity in audio samples).

Subsequently, the windowing process has been performed. This operation aims to achieve continuity among a single frame on one side and all frames on the other. As a result, continuity has been achieved as much as possible from the 1st point in the 1st frame to the last point in the last frame. Mathematically, this process is defined as $s[n] * w[n]$, where $s[n]$ is the signal itself and $w[n]$ represented as follows: [73]

$$w[n, \alpha] = (1 - \alpha) - \alpha cos\left(\frac{2\pi n}{N-1}\right), \quad 0 \leq N \leq N - 1 \tag{3.1}$$

where: $w[n, \alpha]$ is the hamming window, n is the number of frames, α is a fixed value and N is the number of samples per frame.

Consequently, the Fast Fourier Transform (FFT) process has been utilized to change the time domain to the frequency domain. The frequency domain (frequency spectrum for each frame) is valuable as spectral analysis shows that various energy distributions at frequencies are proportional to various timbres in voice signals.

After that, the Mel filter bank process was considered. Essentially, human's understanding of frequencies is non-linear. In other words, a human's ear does not support linear frequency resolution because its sensitivity at high frequencies decreases. The Mel-scale has been used to overcome the limitation of understanding frequencies by a human's ear. The Mel for any frequency has been expressed as follows [74]:

$$Mel(f) = 2595 * log10\left(1 + \frac{f}{700}\right) \quad (3.2)$$

where: $Mel(f)$ is the Mel for any frequency, f is a given frequency in Hz.

Finally, Cepstrum process has been provided. Since the output of the previous process (Mel filter bank) is highly correlated, that is an issue for machine learning algorithms and classification applications to not deal with, it is requisite to de-correlate the output (features) and sort them in descending order of information by using the Discrete Cosine Transform (DCT) [75]. DCT is equivalent to the Discrete Fourier Transform (DFT), except DCT only uses real numbers and its length is double that in the DFT. The MFCC has resulted, it presents a significant representation of the local spectral properties of a speech signal in analyzing a certain frame. This is due to Mel spectrum coefficients and their logarithm are real numbers [76].

3.4 PROPOSED MACHINE LEARNING MODELS

As mentioned before, three models are proposed in this work. The first model is based on the MLP, the second model is based on the CNN, and the third model is based on the DRNN. All the models are designed after many experiments where the number of layers, number of neurons in each layer, and values of parameters have been changed many times to get the best performance for the entire model. These are for presenting appropriate modifications to the task of recognizing the Arab dialects.

3.4.1 Proposed MLP

Fundamentally , the MLP is a kind of machine learning that is utilized as a classifier. So, the MLP in this work is adapted to receive the extracted features of MFCC as inputs. while, its outputs are adapted for distinguishing the three Arabic dialects of APD, ED, and LD. The output essentially involves three nodes (neurons), each class for an Arabic dialect is

represented by a node. The main architecture of the suggested MLP model has been demonstrated as shown in figure 3.5.

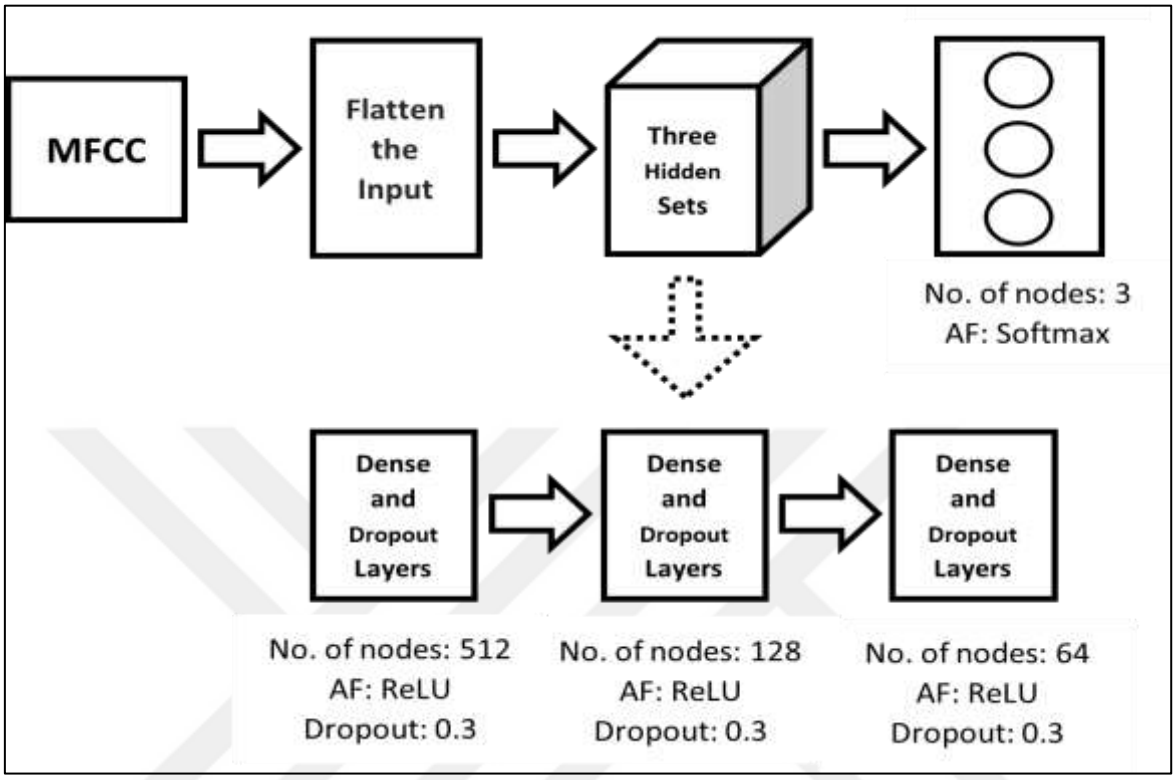

Figure 3.5: The Main Architecture of the Suggested MLP Model.

Since the dataset is prepared and saved as a JavaScript Object Notation (JSON) file (MFCC features), this will lead to a feeding issue because our input is a multi-dimensional array, so it is necessary to convert it to a one-dimensional array (flatten the input out). From the figure, it can be noticed that each MFCC feature has been prepared as a One-Dimensional (1D) vector. Consequently, this vector has been applied as flatten in the input. Three hidden sets have been selected experimentally. They are three dense layers, and all of them use a Rectified Linear Unit (ReLU) as an Activation Function (AF) which is easy to implement and fast. Also, the dropout layer followed each dense layer, which work on reducing the data size. The numbers of nodes in the dense layers are as follows: 512 neurons for the first dense layer, 256 neurons for the second dense layer, and 64 neurons for the third dense layer. Each dropout layer works on reducing the dataset size by 0.3 (or 30%). In the output layer, three nodes are utilized, as mentioned before, and softmax AF has been applied. The softmax AF

helps in selecting the neuron of the highest value to acquire the specific class of the Arabic dialect.

The issue of overfitting is also taken into consideration here, as the network architecture is simple where the number of neurons and layers are as few as possible, and good model performances are maintained. In addition to that, utilizing the dropout of 30% and L2 regularization in the hidden sets is valuable for addressing the overfitting.

3.4.2 Proposed CNN

Principally, CNN is a kind of deep learning which can be used for image processing. Therefore, the CNN in this work is adapted to receive the MFCC as an input image. Likewise, the suggested MLP, outputs are adapted for recognizing among the three Arabic dialects. Again, in the output layer, there are three neurons too, each neuron refers to a specific class for an Arabic dialect. The main architecture of the suggested CNN model has been described as given in Figure 3.6.

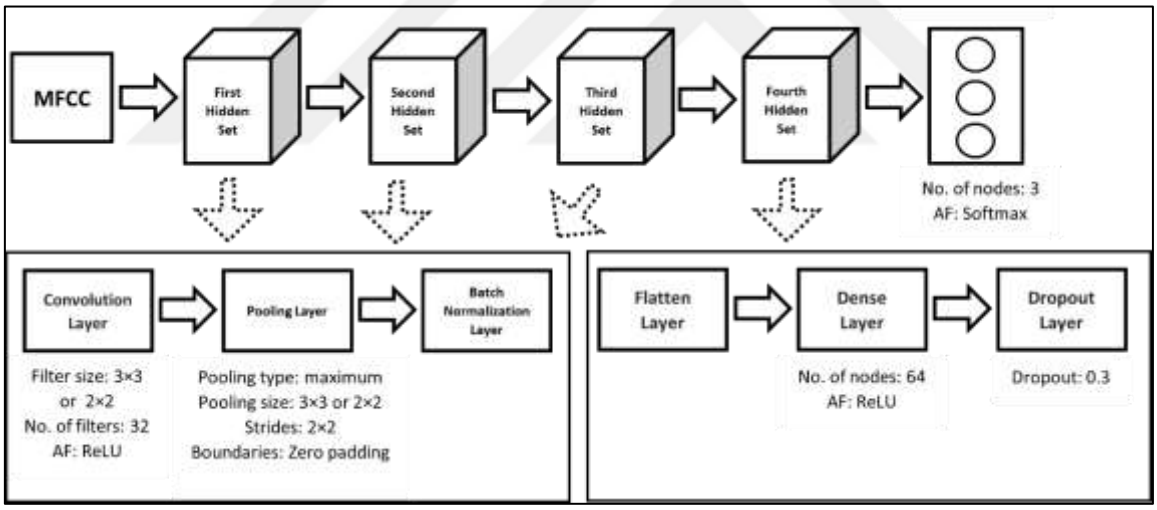

Figure 3.6: The Main Architecture of The Suggested CNN Mode.

From the figure and as mentioned before, it can be obtained that each MFCC feature has been prepared as a Two-Dimensional (2D) vector. This vector has been applied as input. The hidden layers are also empirically selected. As it is clear from the figure, there are three sets of convolution, pooling, and normalization layers that have been sequentially arranged. The first convolution layer has 32 filters (kernels), each of size 3×3 pixels. The first pooling layer is of maximum type. It downsized 2D vectors by utilizing the pooling size of 3×3 pixels, vertically and horizontally stride of 2×2 pixels, and boundaries of zero padding. The second

and third convolution and pooling layers use the exact parameters that have been applied for the first convolution and pooling layers. The third convolution layer considers the kernel size of 2×2 pixels and the third pooling layer utilizes the pooling size of 2×2 pixels. Batch normalization is utilized for all normalization layers. In the last hidden set, it is important to flatten the output to a 1D vector in order to feed it into a dense layer with 64 neurons and (ReLU) AF. The dense layer is followed by the dropout layer of 0.3 (or 30%), which works on reducing the dataset size in order to overcome the overfitting problem. In the output layer, the softmax AF has been applied too for the three nodes. Again, it is useful for selecting the neuron of the highest value to attain the specific class of the Arabic dialect.

3.4.3 Proposed DRNN

The DRNN is a kind of deep recurrent learning that can be employed for pattern association. The DRNN in this work has been also adapted to accept the MFCC as suitable values. Similarly to the suggested MLP and CNN, the outputs have been adapted for distinguishing among the three Arabic dialects. So, there are three neurons too in the output, each neuron leads to a certain class for an Arabic dialect. The main architecture of the suggested DRNN model can be described as demonstrated in Figure 3.7.

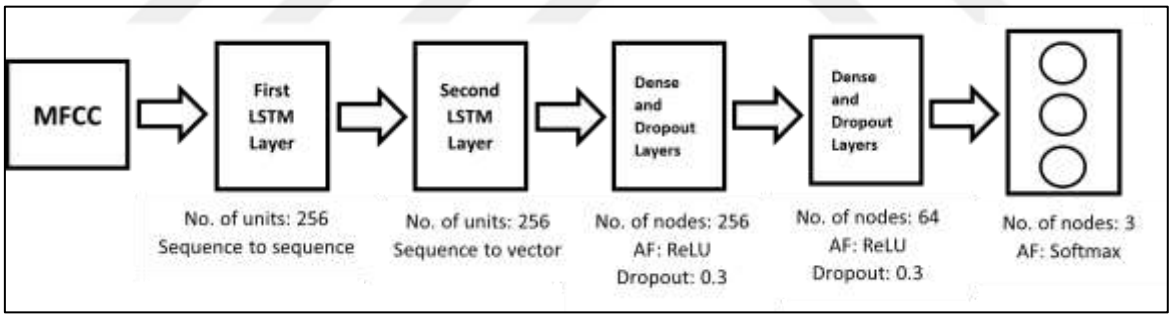

Figure 3.7: The Main Architecture of the Suggested DRNN Model.

From the figure and as mentioned previously, it can be noticed that each MFCC feature has been prepared as a 2D vector. Subsequently, this vector is used as input. The hidden layers are also experimentally determined. They involve two Long Short Term Memory (LSTM) layers followed by two dense layers. The first LSTM layer is a type of sequence to sequence and it consists of 256 processing units. The second LSTM layer is a type of sequence to vector and it consists of 256 processing units. The first dense layer utilized the ReLU as AF followed by a dropout of 0.3 (or 0.3%), to address the overfitting issue, and it consists of 256 neurons. The second dense layer also used the ReLU as AF followed by a dropout of 0.3%,

again to overcome the overfitting problem, but it consists of 64 neurons. In the output layer, the softmax AF is also utilized for all three nodes. As mentioned before, this can be helpful to determine the node of the highest value to reach the specific class of the Arabic dialect.

4. RESULTS

At the beginning, all the experiments are achieved in a Personal Computer (PC) with the following specifications: Asus Laptop, Intel core i7 processor with 2.20 GHz speed, 8 embedded Central Processing Units (CPUs), 8 GB main memory, NVIDIA external graphics card, Graphical Processing Unit (GPU) type GT 630M and 2 GB display memory. In addition, Python version 3.10.9 (64-bits) is utilized as a software tool and Anaconda version 3 is used as an editor.

4.1 EMPLOYED DATASET

As mentioned before, the dataset of SARA [22] is used in this work. It includes 3840 samples for the three Arabic dialects (ARP, ED, and LD). Where SARA has been described in detail in Tables 3.1, 3.2, and 3.3. As a reminder, it has 1230, 1365, and 1245 samples of ARP, ED, and LD, respectively. It is also partitioned into five groups depending on the recording times (3, 4, 5, 6, and 7) seconds, where it involves 1254, 1140, 786, 441, and 219 samples, respectively. Each group includes samples of the three Arabic dialects (ARP, ED, and LD) as described in Table 3.2.

The SARA has been called in this work as the OSARA, which is filtered as the FSARA. Then, both the OSARA and FSARA are concatenated and utilized as the MSARA.

4.2 PRE-PROCESSING RESULTS

In this stage, the DEEPFILTERNET2 is used to purify the OSARA dataset. DEEPFILTERNET2 has been applied to every single sample individually, one by one. This leads to protect the data from missing some information. According to the nature of the dataset, which has been obtained from films, series, and acting scenes that consider the three Arabic dialects, the results are classified into four groups depending on the type of unwanted sounds in the original sample and the effectiveness of the DEEPFILTERNET2 on that sample:

4.2.1 Standard Samples

Most samples fall within this group, as it has a slight noise without influencing the original speaker's voice. The DEEPFILTERNET2 has proved its effectiveness in this kind by

improving the purifications of original samples. Figure 4.1 and figure 4.2 show the original and filtered samples that demonstrate such achieved enhancement.

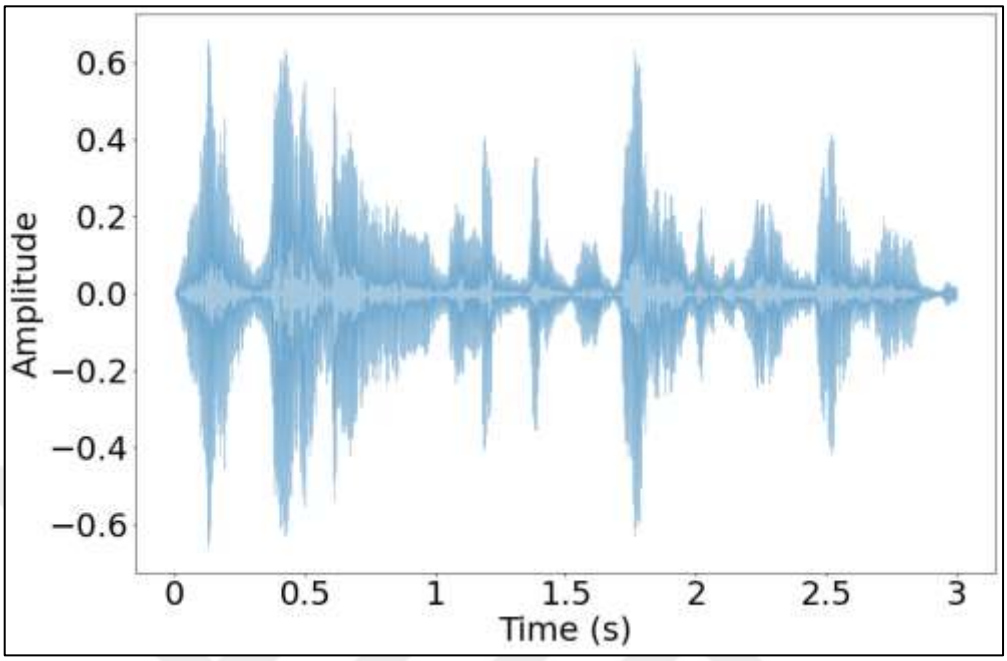

Figure 4.1: The Original Sample Type of Standard Samples.

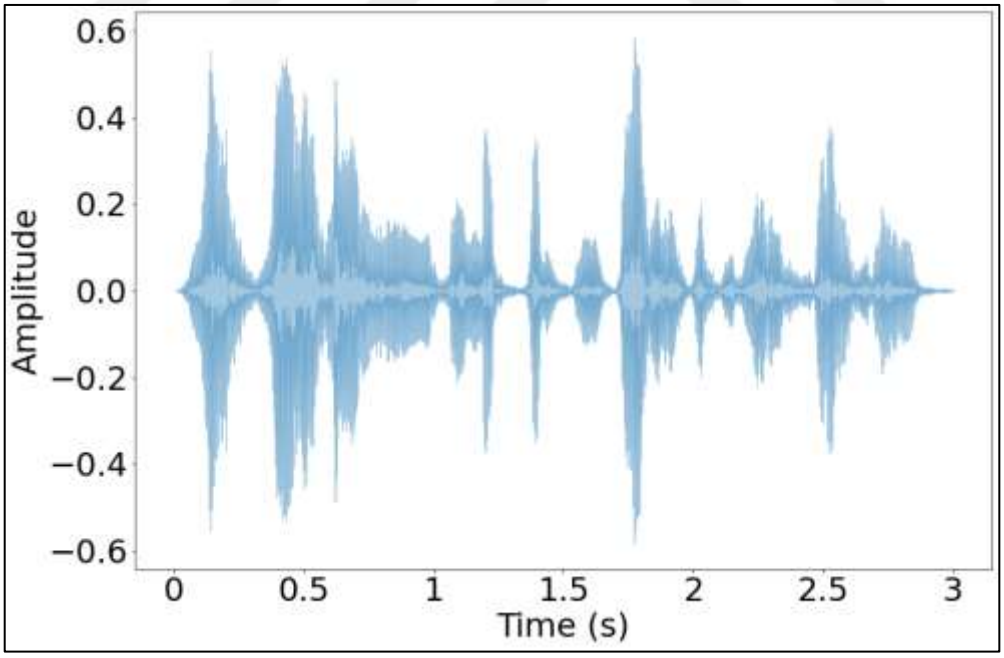

Figure 4.2: The Filtered Sample Type of Standard Samples.

4.2.2 Samples with Background Sounds

According to the nature of artistic environments, background sounds such as music are combined within the speaker's voice. This is expected to affect the speaker's voice. The DEEPFILTERNET2 has also confirmed its efficiency in removing background sounds without affecting the speaker's voice, as given in Figure 4.3 and Figure 4.4.

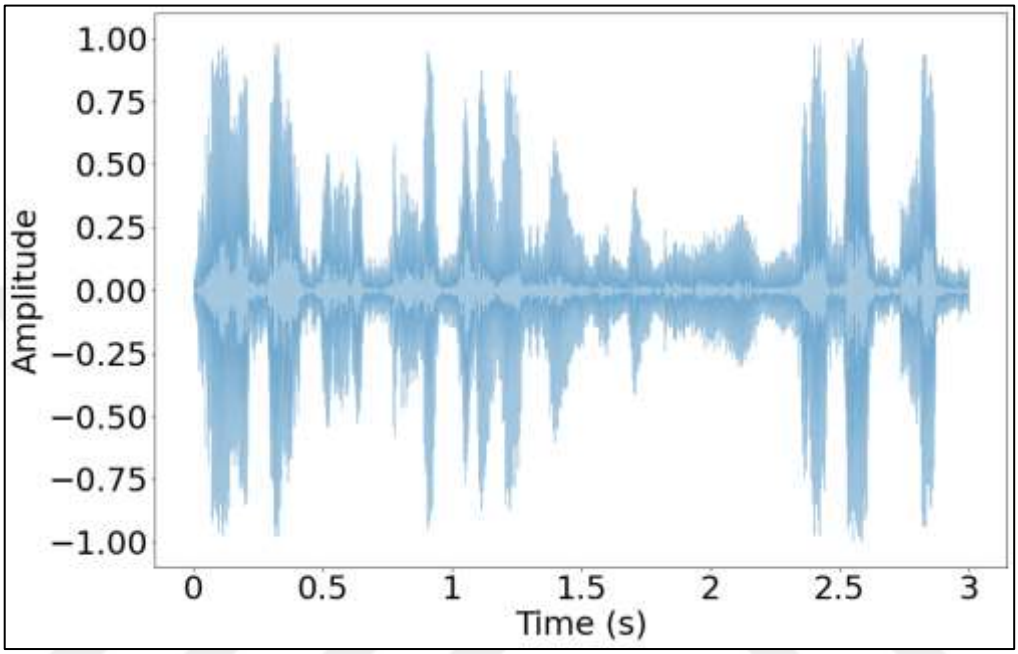

Figure 4.3: The Original Sample Type of Samples with Background Sounds.

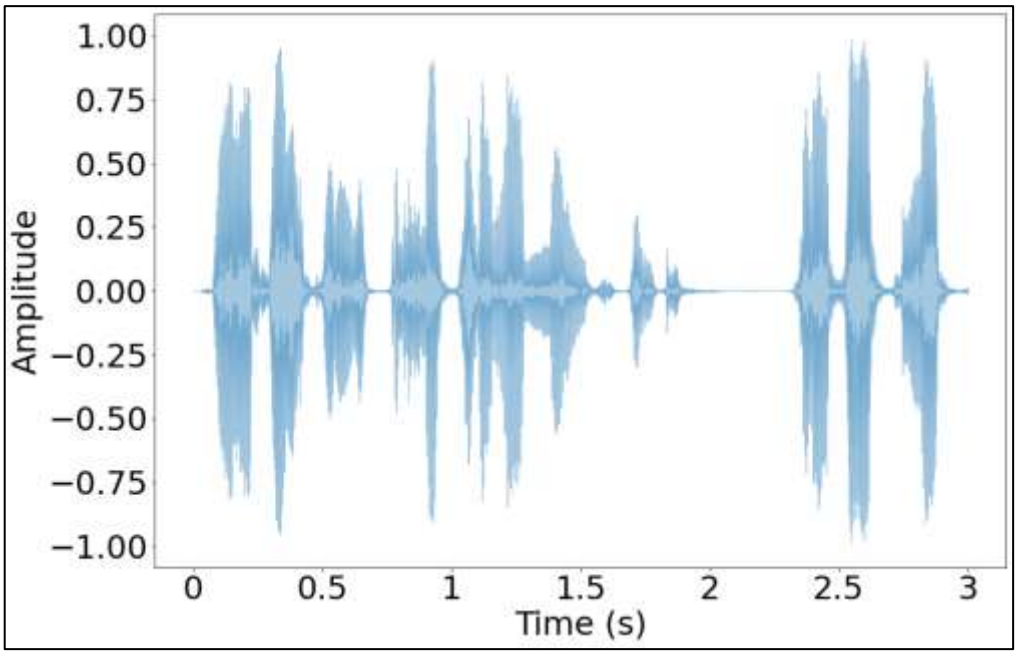

Figure 4.4: The Filtered Sample Type of Samples with Background Sounds.

4.2.3 Samples with Existed Noise

The OSARA dataset had been collected from movies and series. Therefore, it is normal for original samples to have noise that could affect speakers' voices. Some conversations may need to be fully understood by listeners. Again, the DEEPFILTERNET2 has proved its valuableness in dealing with this kind too as shown in figure 4.5 and figure 4.6.

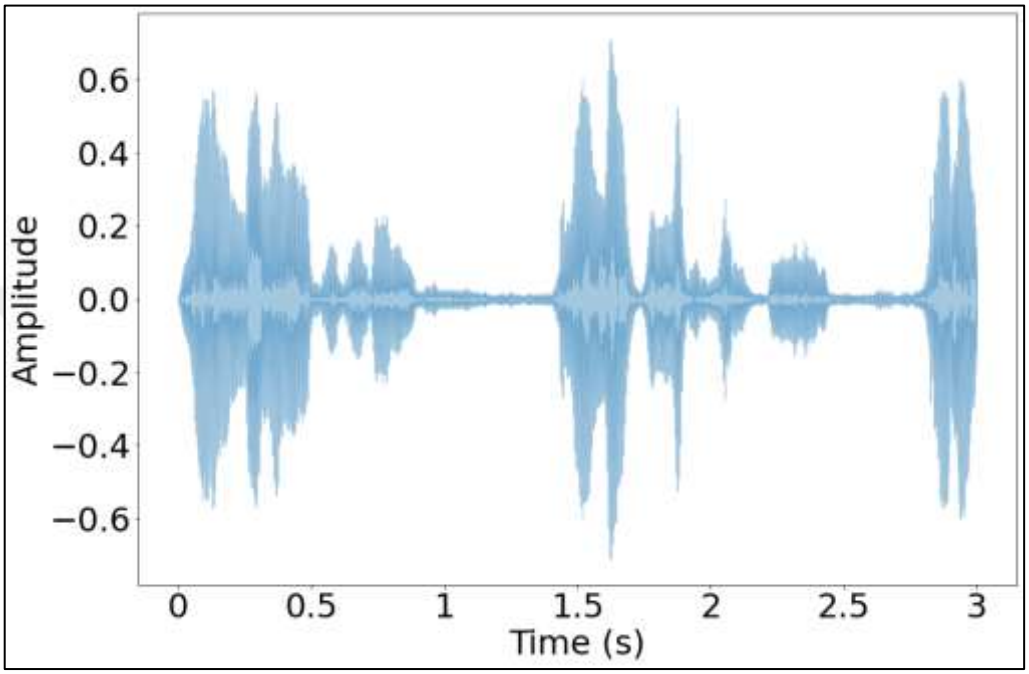

Figure 4.5: The Original Sample Type of Samples with Existed Noise.

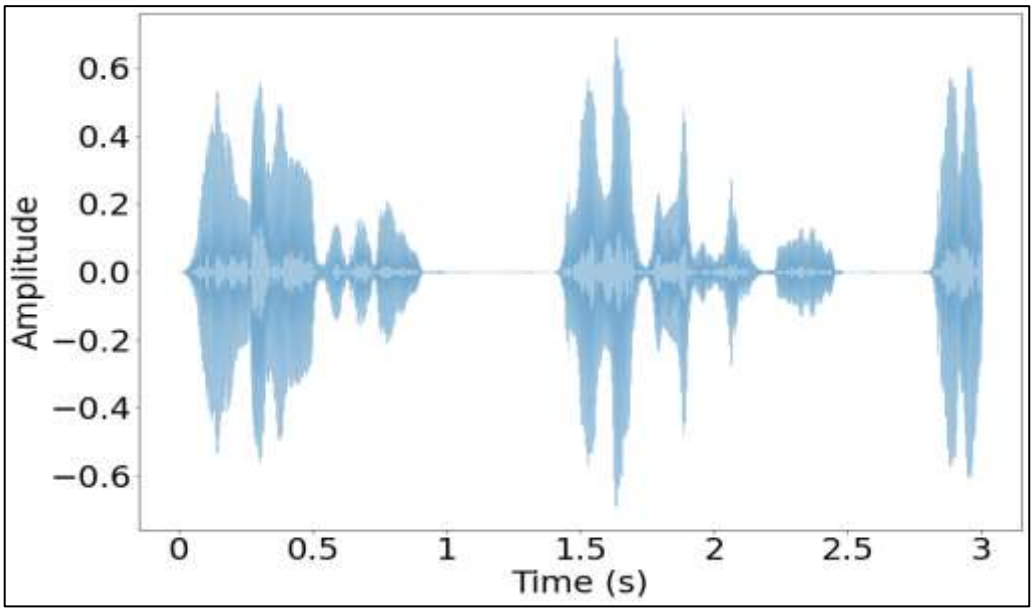

Figure 4.6: The Filtered Sample Type of Samples with Existed Noise.

4.2.4 Samples with Missing Information

Like any other practical application, there is no idealism. Using the DEEPFILTERNET2, information loss appeared in some samples, even in a few samples where it did not exceed 1%. This has been dealt with by maintaining the original samples and preserving the OSARA dataset as much as possible. Figure 4.7 and Figure 4.8 illustrates this case.

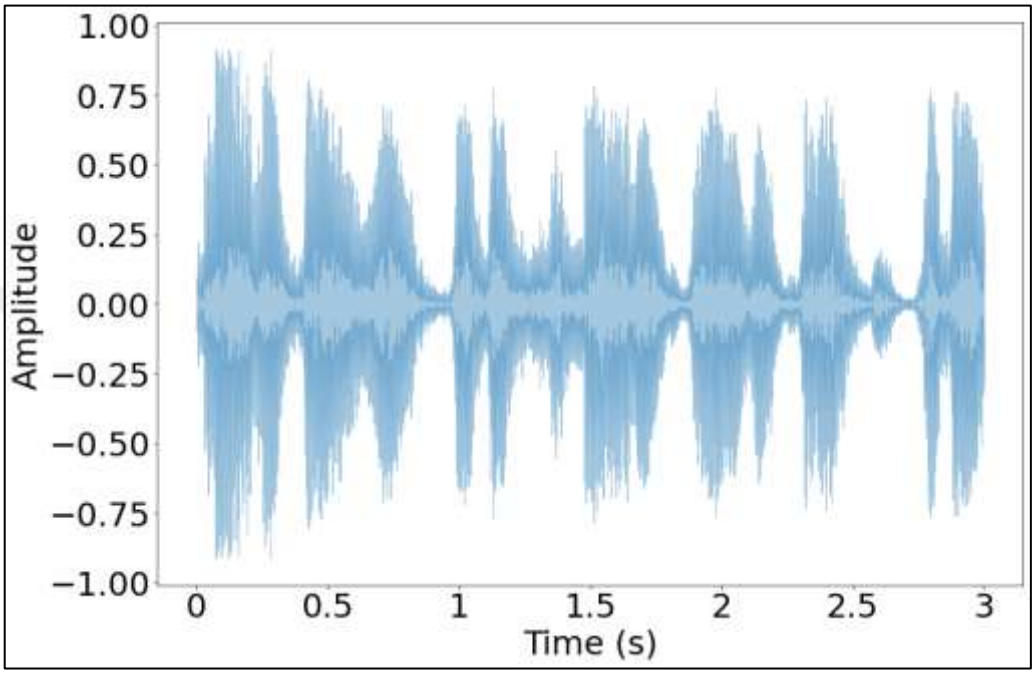

Figure 4.7: The Original Sample Type of Samples with Missing Information.

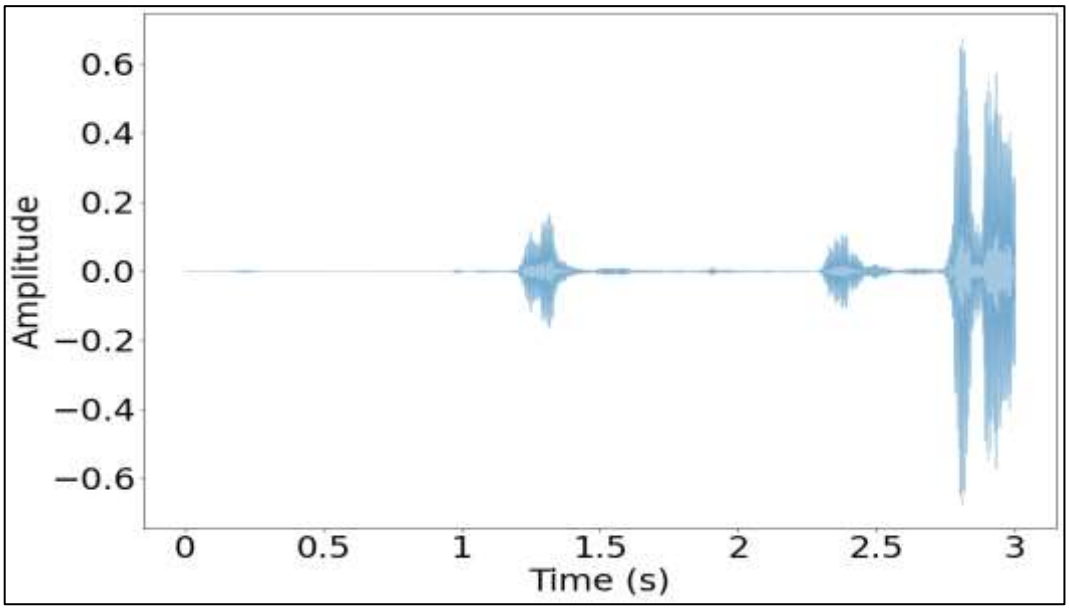

Figure 4.8: The Filtered Sample Type of Samples with Missing Information.

4.3 FEATURE EXTRACTION RESULTS

Practical experiments have been considered for setting MFCC parameters in order to acquire the best feature extraction. The adopted parameters in this work are as follows: Sample rate is 22050, Number of coefficients to extract is 40, Number of samples for FFT is 2048, Hop length (Sliding window for FFT) is 512, and number of segments that divide sample tracks into is 5. Each MFCC and its label has been saved in a file type JSON to be utilized later as input for a proposed model. Figure 4.9, figure 4.10, and Figure 4.11 show examples of MFCC feature extractions for APD, ED, and LD.

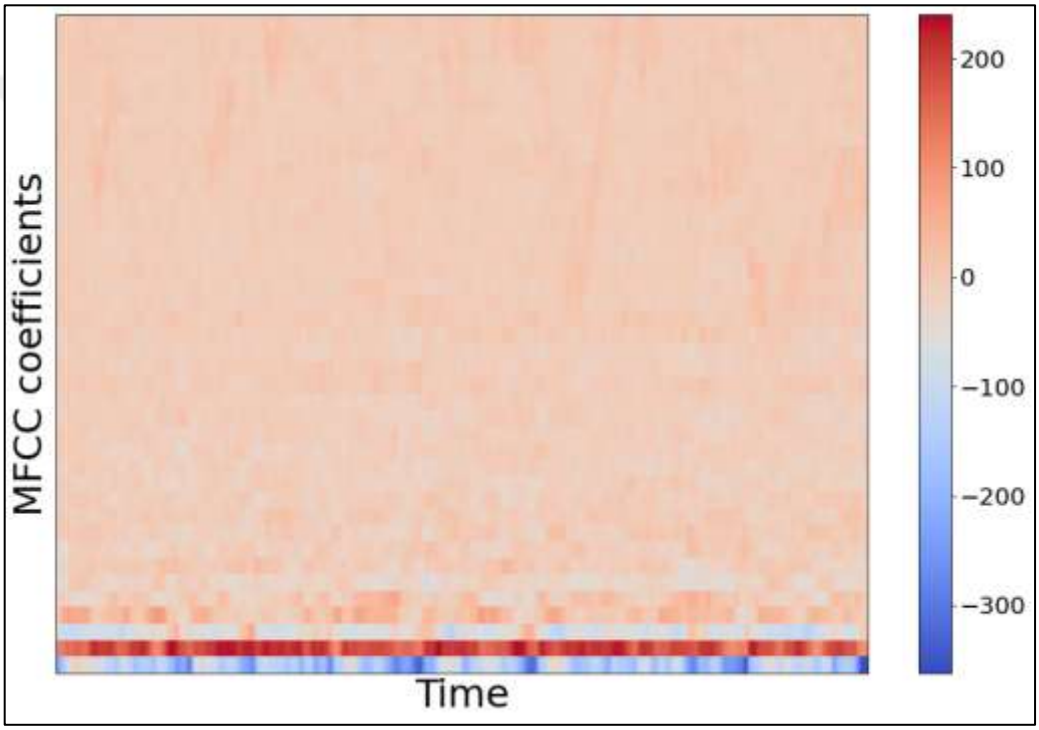

Figure 4.9: The Example of MFCC Feature Extractions for an APD Sample.

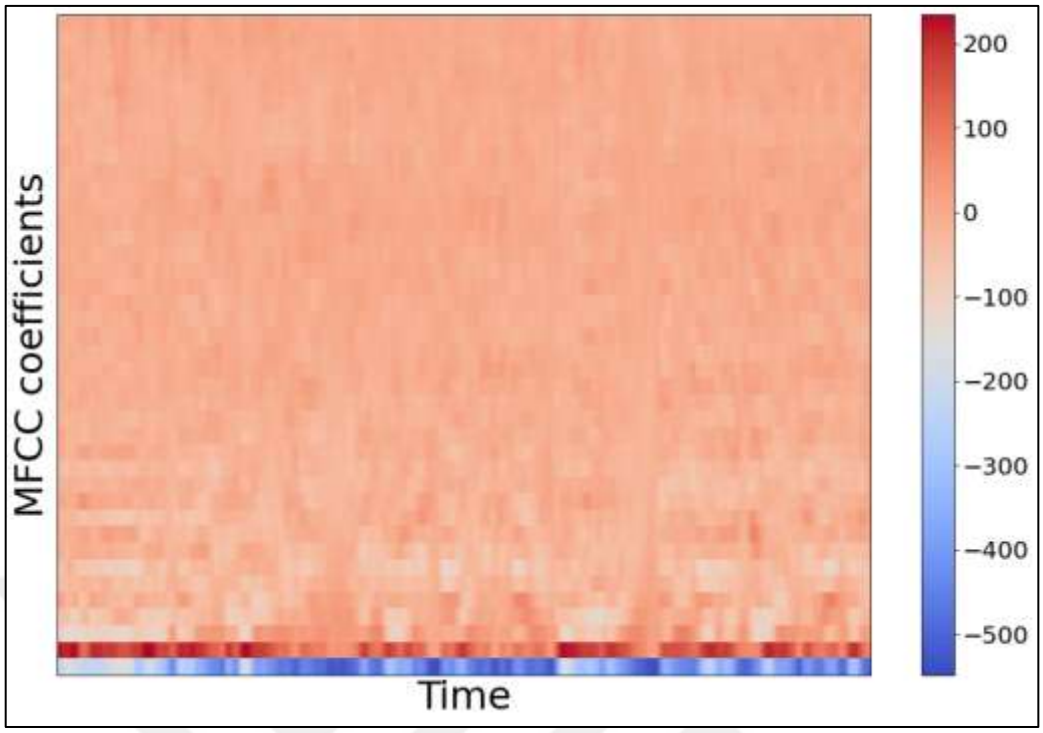

Figure 4.10: The Example of MFCC Feature Extractions for an ED Sample.

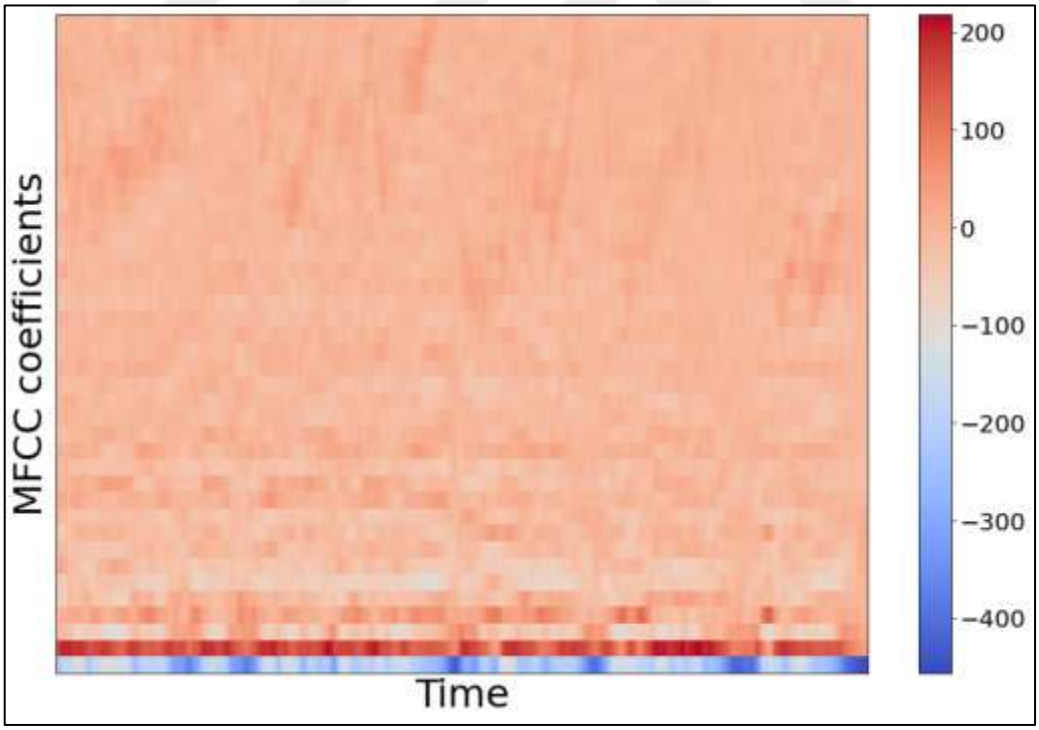

Figure 4.11: The Example of MFCC Feature Extractions for an LD Sample.

4.4 TRAINING AND TESTING RESULTS

As mentioned previously, three models are proposed in this work (MLP, CNN, and DRNN models). All experiments in this work adopt the training parameters as follows: optimizer of type Adam, the fixed learning rate of 0.0001, Loss function of type sparse categorical cross-entropy, and mini-batch size of 32. All dataset groups are randomly partitioned into 70% for training and 30% for testing in the MLP model, while in the CNN and DRNN models, all dataset groups (OSARA, FSARA, and MSARA) are randomly partitioned into 60% for training, 15% for validation and 25% for testing. So, the partitioned testing groups are approximately equivalent.

Vast training experiments have been taken into consideration for each proposed model and for all three groups of the utilized dataset. Moreover, investigate the five lengths of samples: 3, 4, 5, 6, and 7 seconds for all three groups (OSARA, FSARA, and MSARA).

4.4.1 MLP Training and Testing Results

Figure 4.12 to Figure 4.16 provides the training and testing curves in addition to the loss curves of the proposed MLP model for the OSARA dataset group and for all five lengths of samples (3, 4, 5, 6, and 7 seconds). Similarly, figure 4.17 to Figure 4.21 shows the training and testing curves in addition to the loss curves of the MLP model for the FSARA dataset group and all five lengths. Likewise, figure 4.22 to Figure 4.26 displays the training and testing curves in addition to the loss curves of the proposed MLP for the MSARA dataset group and all five lengths too.

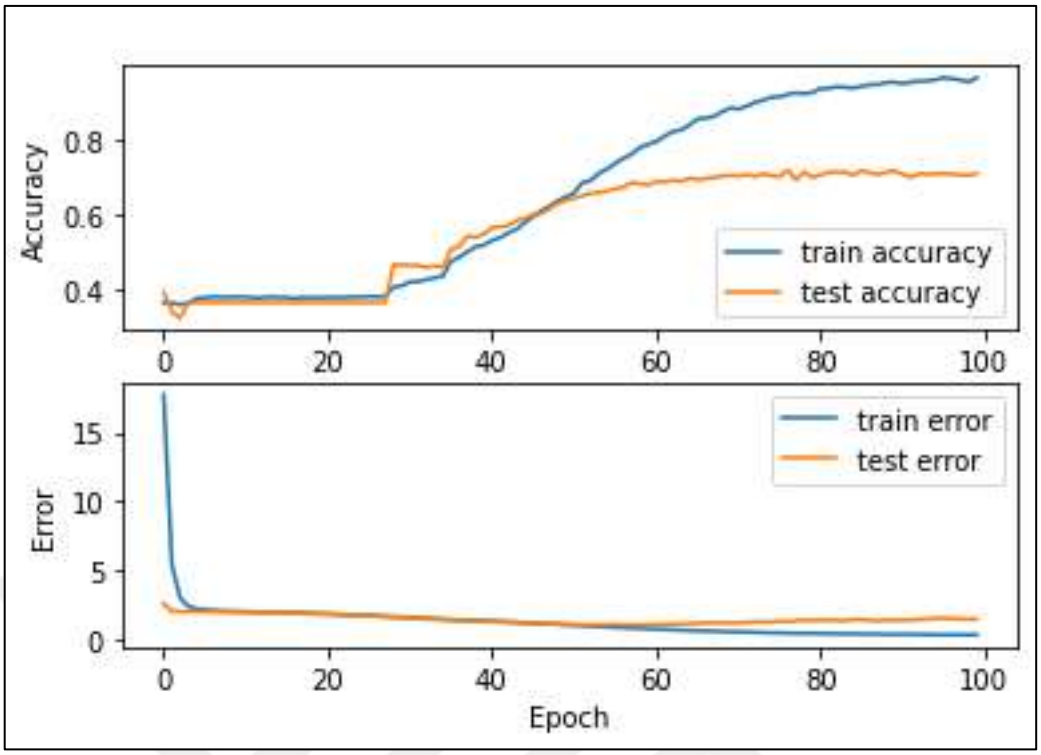

Figure 4.12: MLP Training and Testing Curves with Loss Curves for the OSARA Dataset Group and for the Length of 3 Seconds Samples.

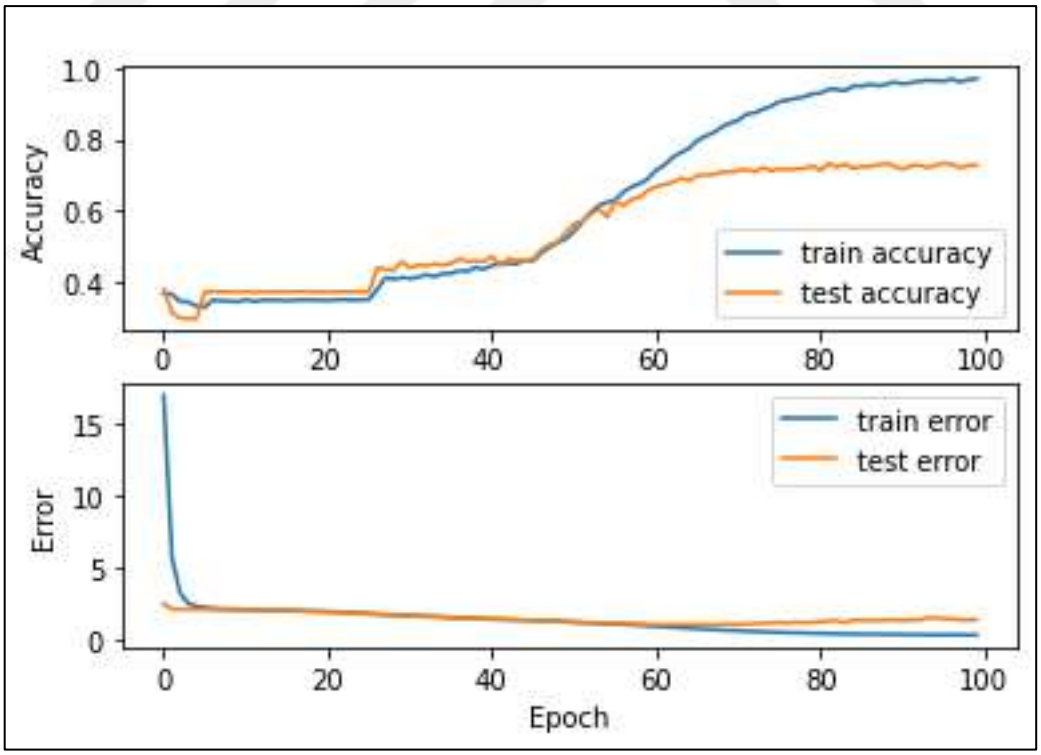

Figure 4.13: MLP Training and Testing Curves with Loss Curves for the OSARA Dataset Group and for the Length of 4 Seconds Samples.

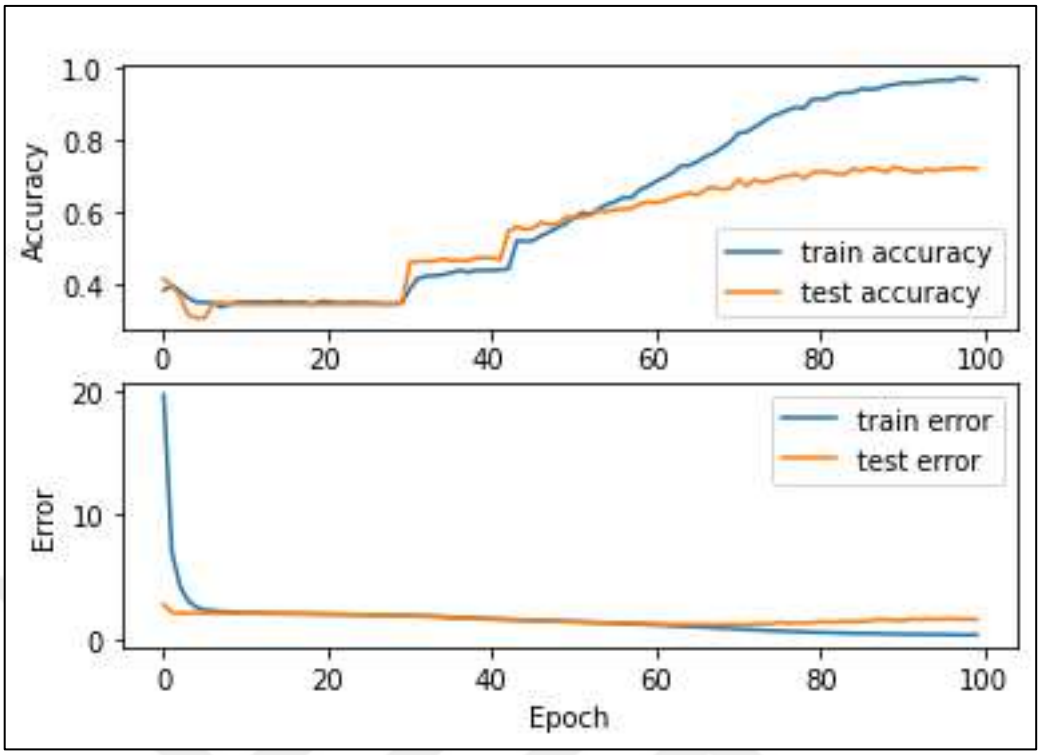

Figure 4.14: MLP Training and Testing Curves with Loss Curves for the OSARA Dataset Group and for the Length of 5 Seconds Samples.

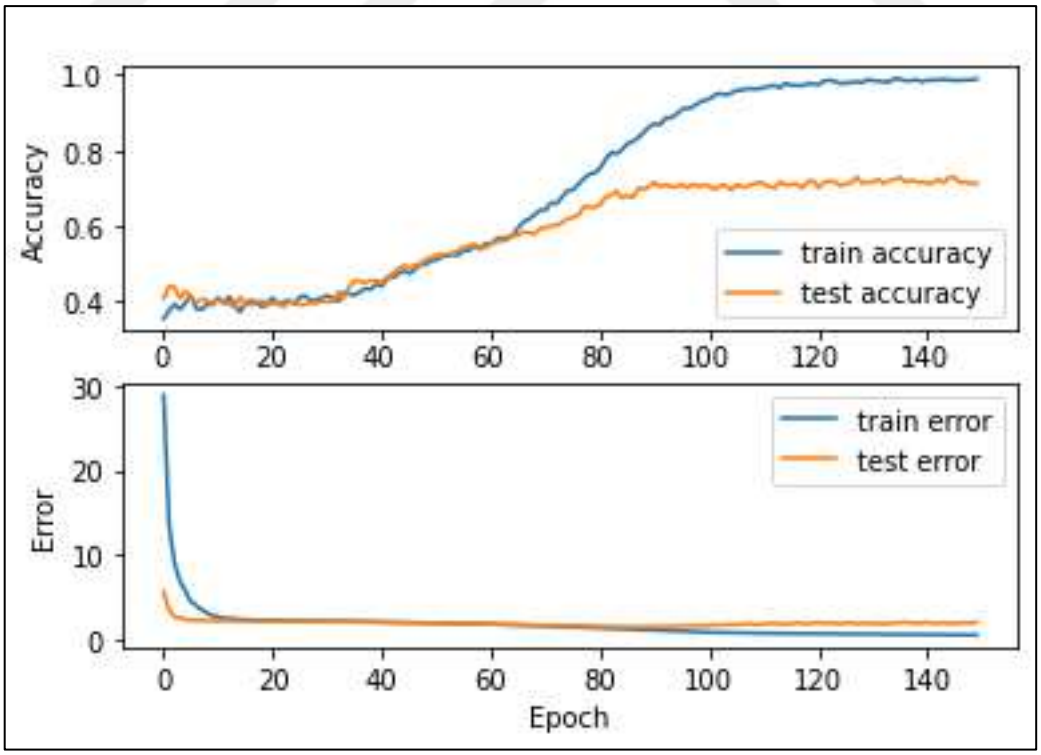

Figure 4.15: MLP Training and Testing Curves with Loss Curves for the OSARA Dataset Group and for the Length of 6 Seconds Samples.

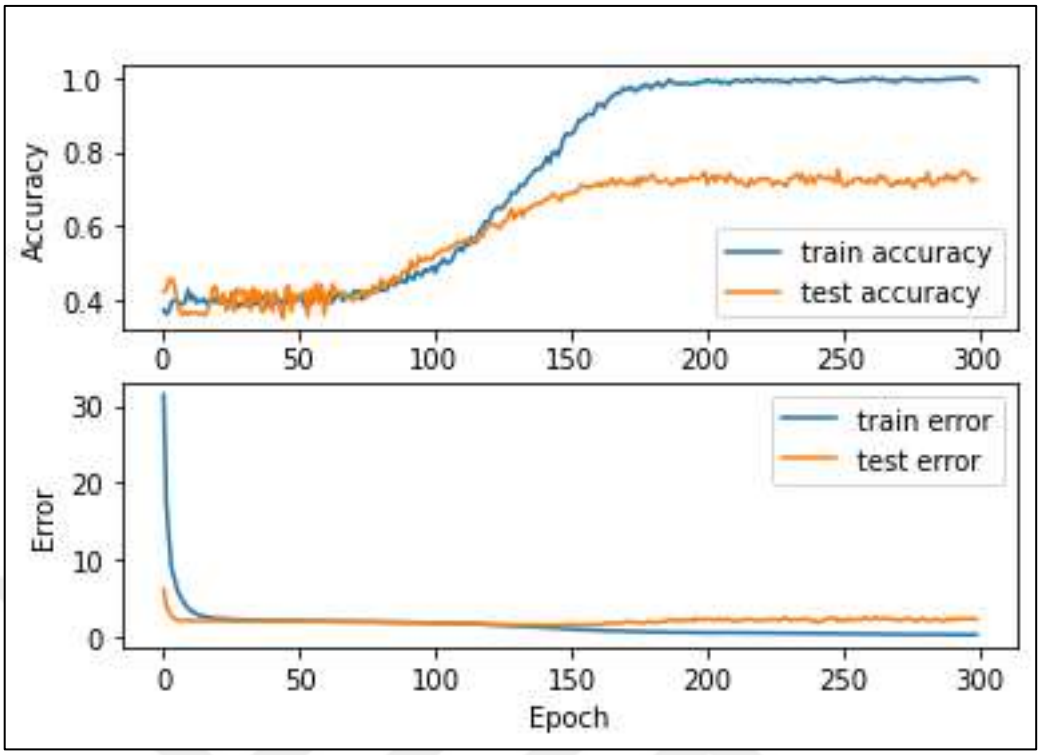

Figure 4.16: MLP Training and Testing Curves with Loss Curves for the OSARA Dataset Group and for the Length of 7 Seconds Samples.

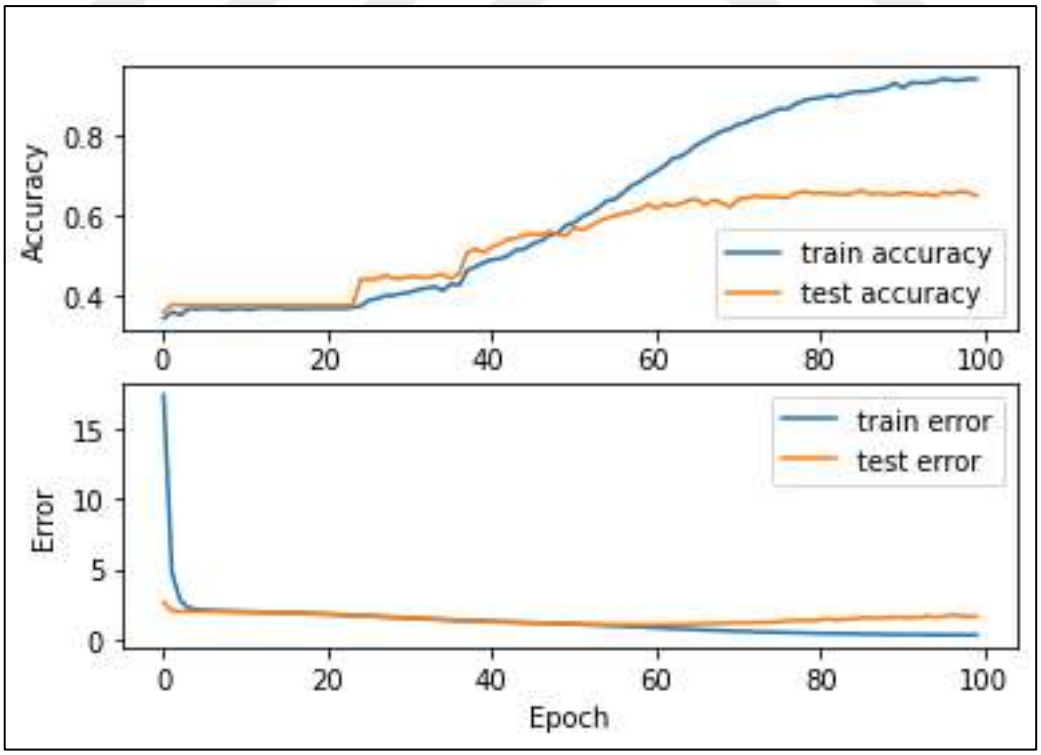

Figure 4.17: MLP Training and Testing Curves with Loss Curves for the FSARA Dataset Group and for the Length of 3 Seconds Samples.

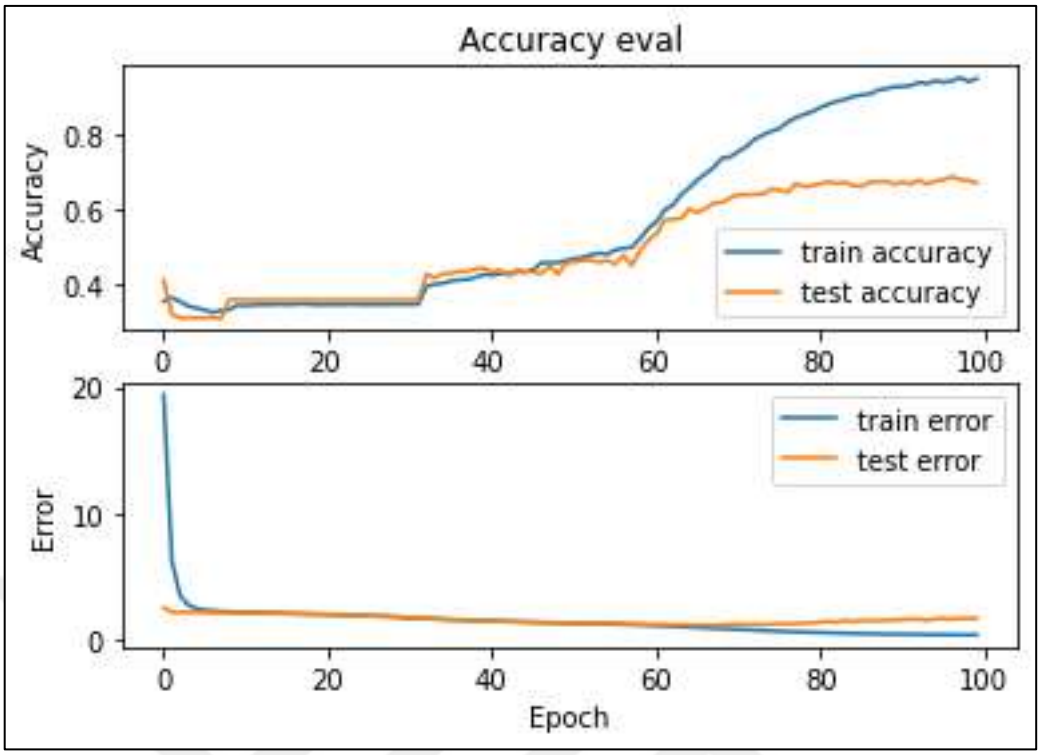

Figure 4.18: MLP Training and Testing Curves with Loss Curves for the FSARA Dataset Group and for the Length of 4 Seconds Samples.

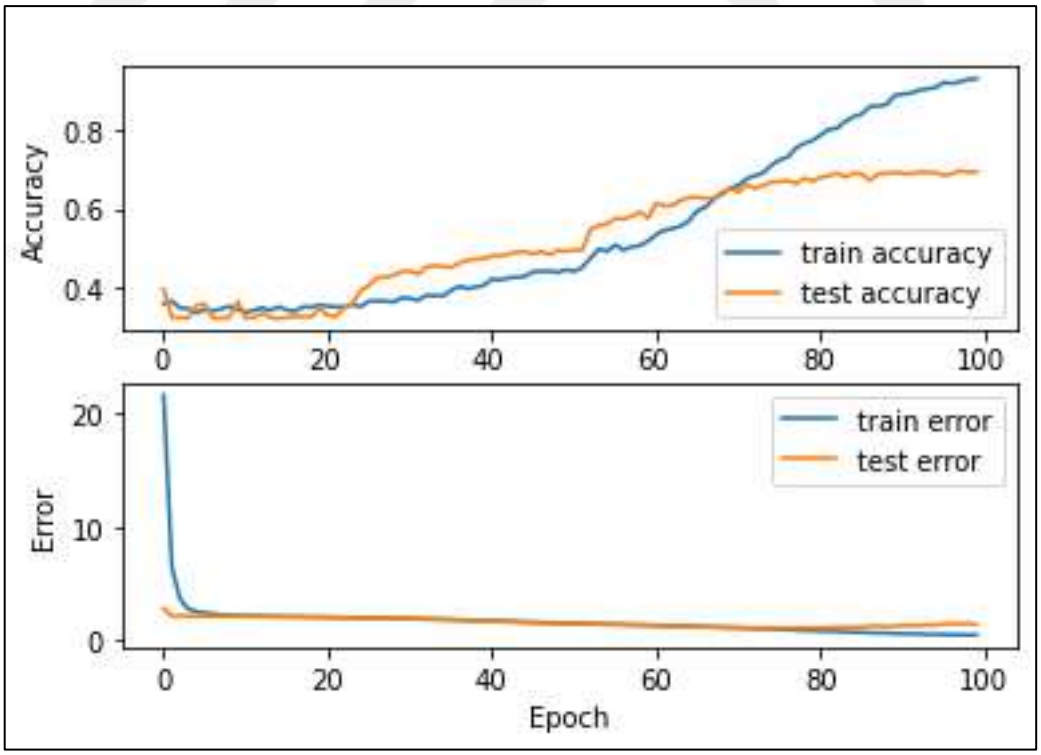

Figure 4.19: MLP Training and Testing Curves with Loss Curves for the FSARA Dataset Group and for the Length of 5 Seconds Samples.

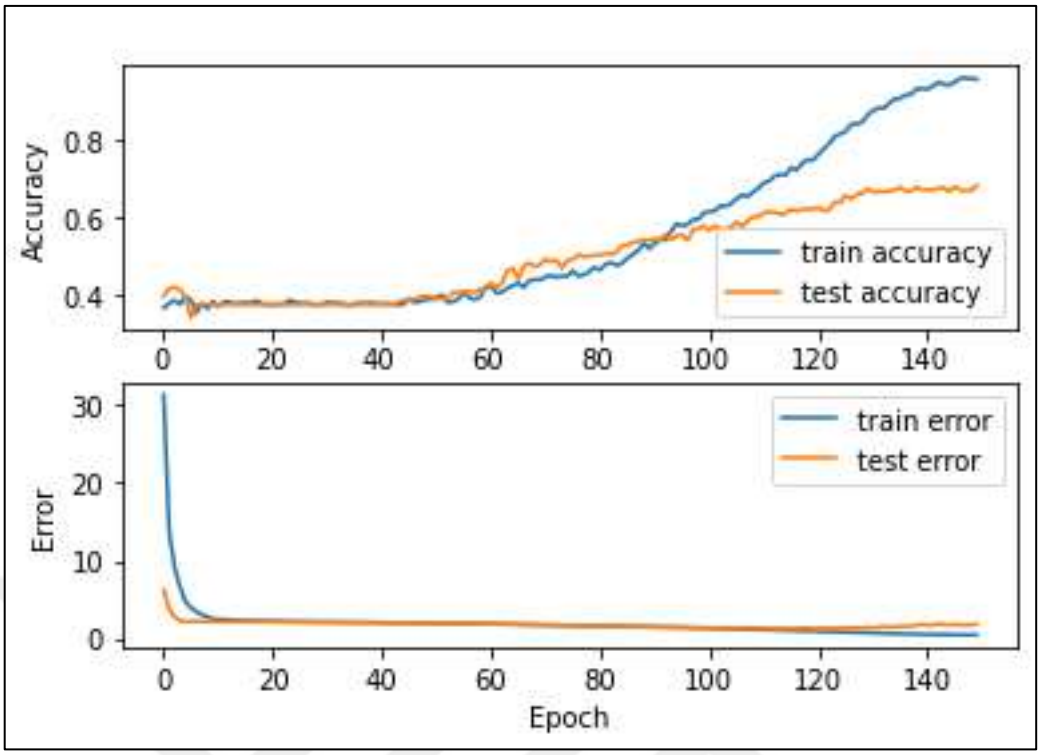

Figure 4.20: MLP Training and Testing Curves with Loss Curves for the FSARA Dataset Group and for the Length of 6 Seconds Samples.

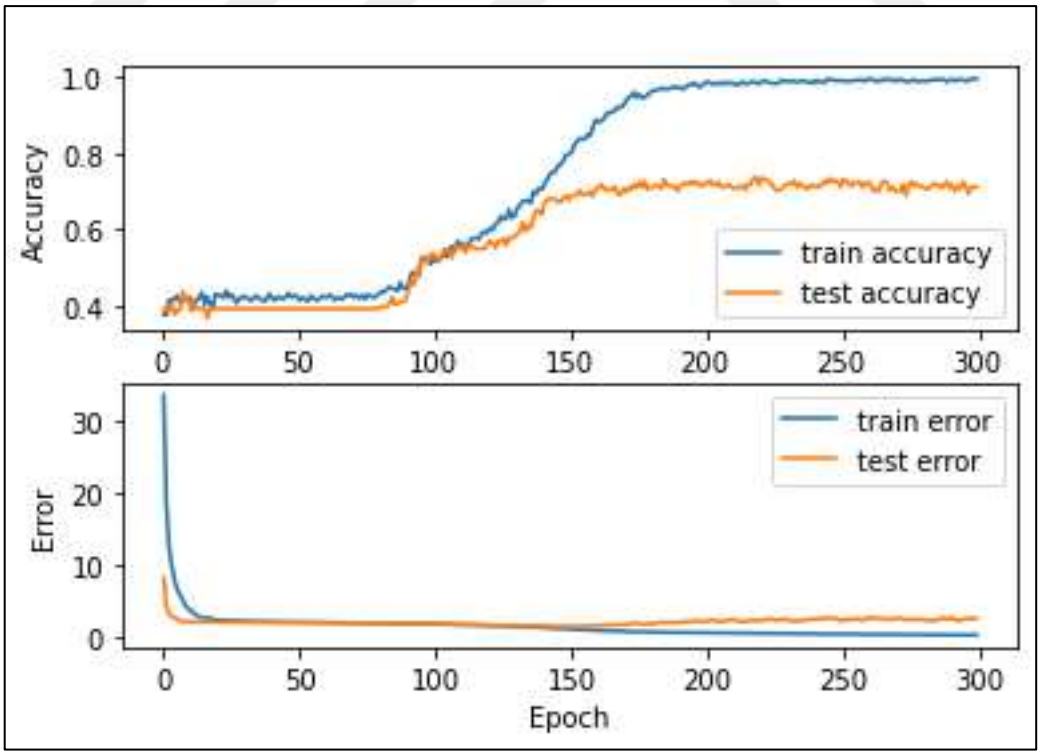

Figure 4.21: MLP Training and Testing Curves with Loss Curves for the FSARA Dataset Group and for the Length of 7 Seconds Samples.

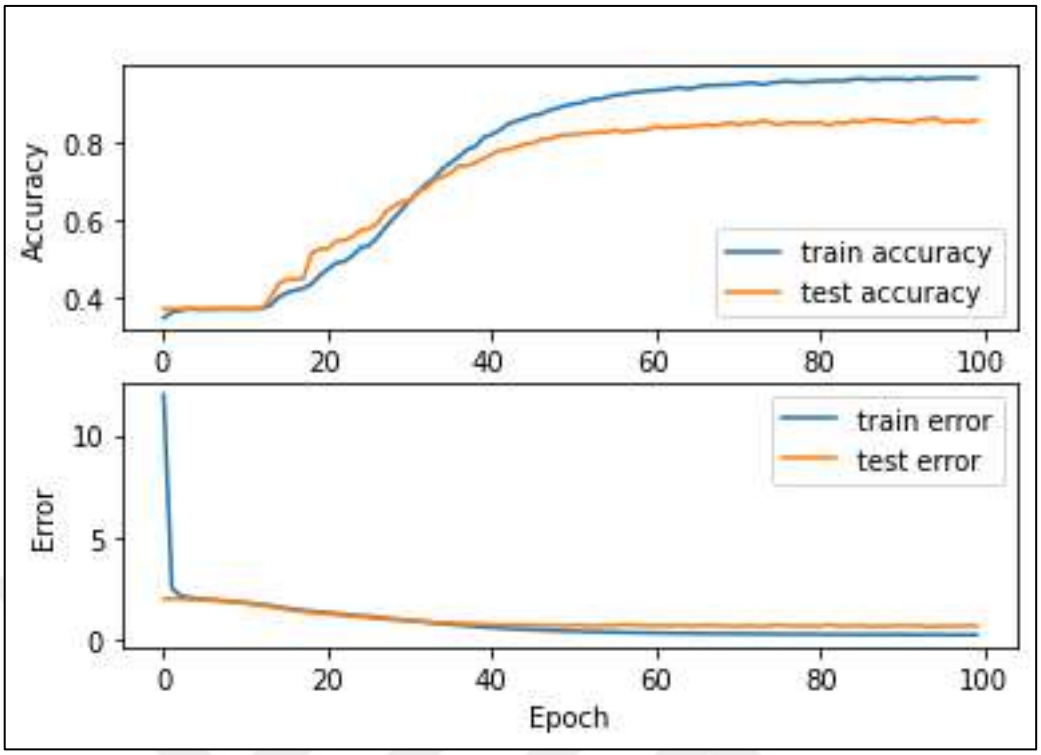

Figure 4.22: MLP Training and Testing Curves with Loss Curves for the MSARA Dataset Group and for the Length of 3 Seconds Samples.

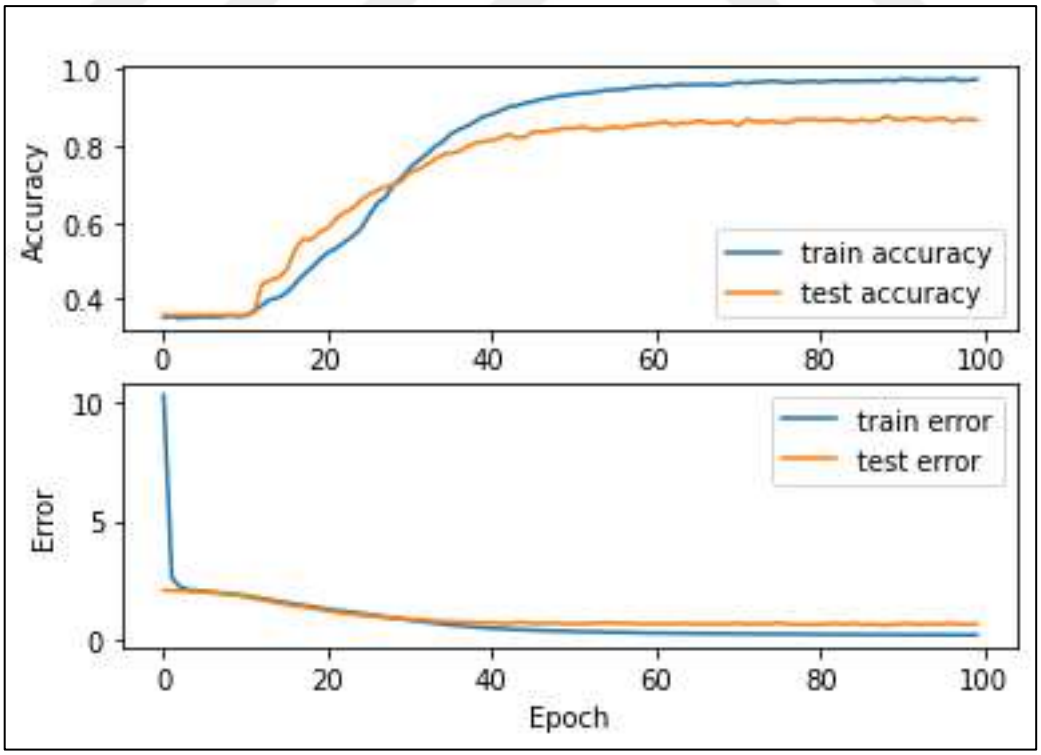

Figure 4.23: MLP Training and Testing Curves with Loss Curves for the MSARA Dataset Group and for the Length of 4 Seconds Samples.

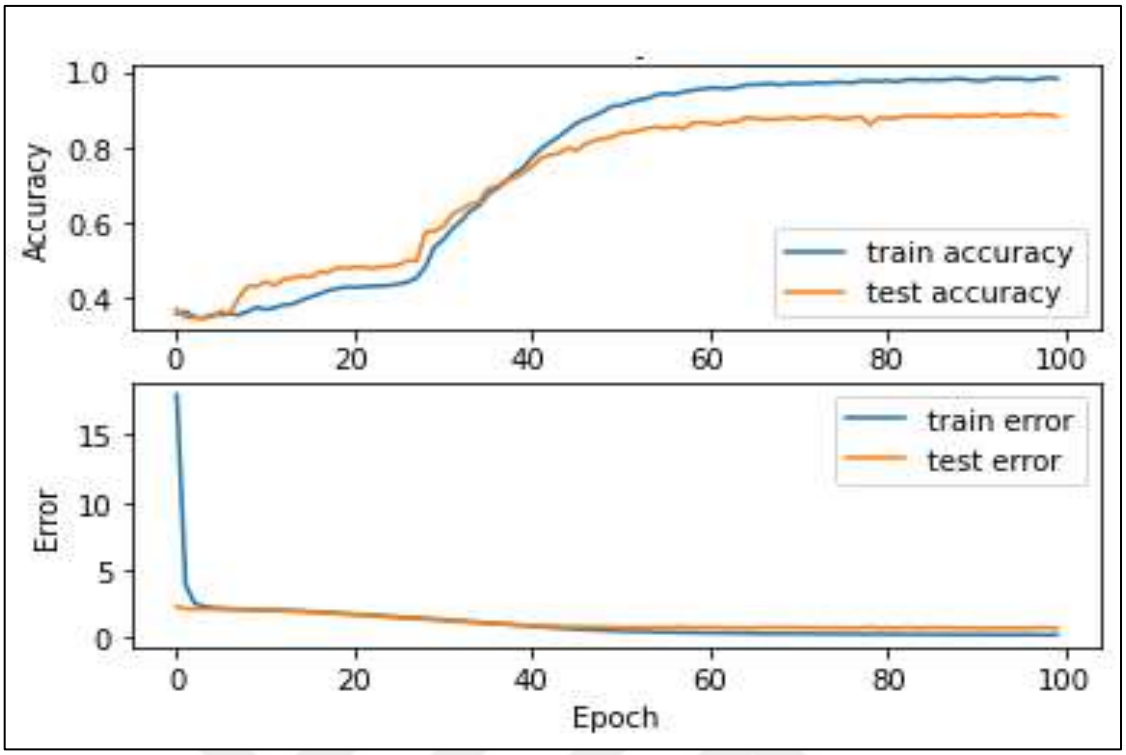

Figure 4.24: MLP Training and Testing Curves with Loss Curves for the MSARA Dataset Group and for the Length of 5 Seconds Samples.

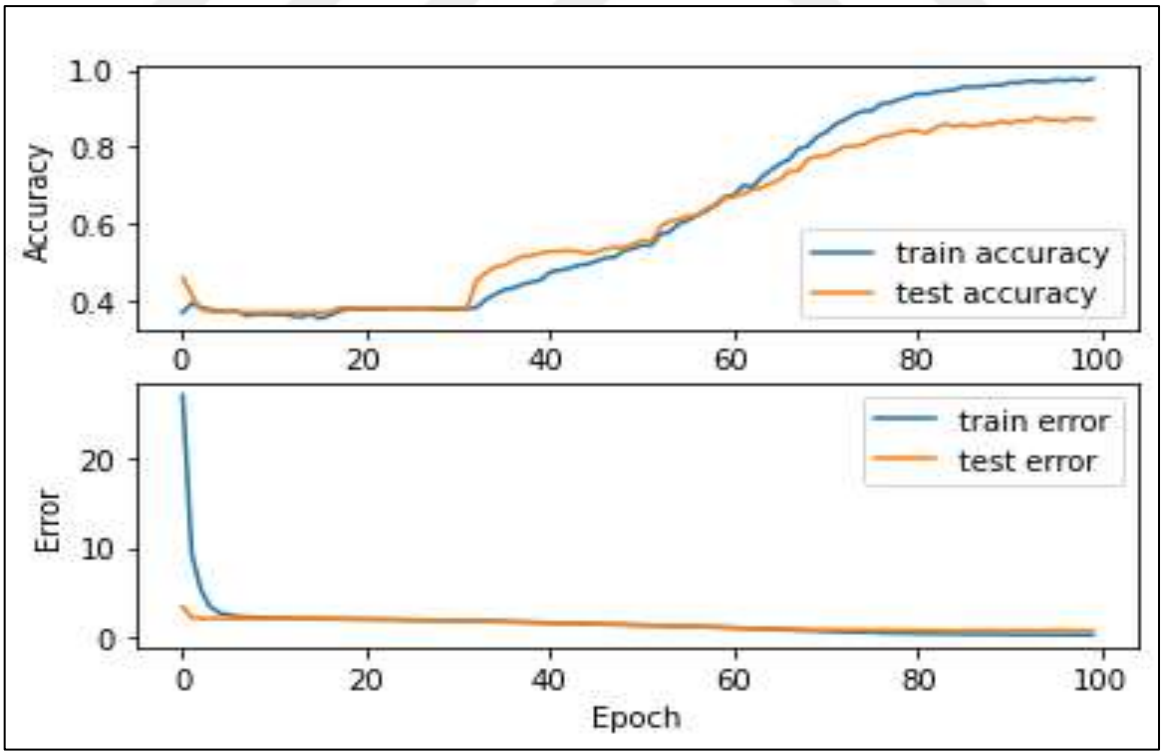

Figure 4.25: MLP Training and Testing Curves with Loss Curves for the MSARA Dataset Group and for the Length of 6 Seconds Samples.

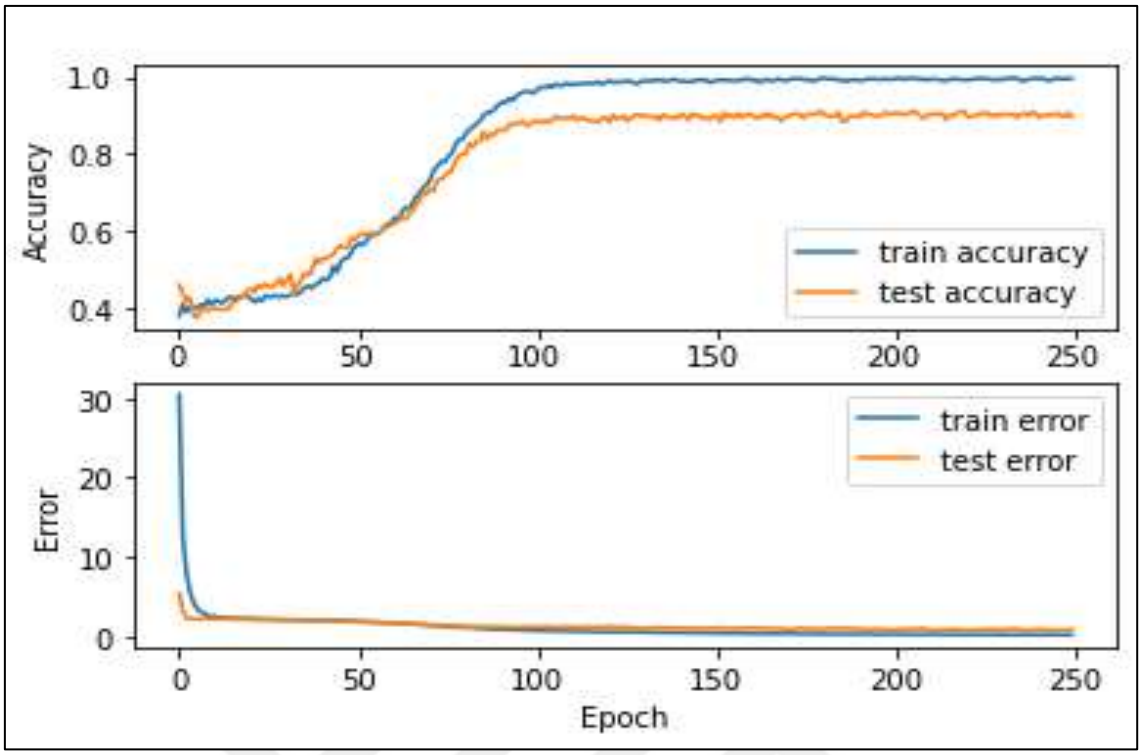

Figure 4.26: MLP Training and Testing Curves with Loss Curves for the MSARA Dataset Group and for the Length of 7 Seconds Samples.

From these figures, it can be observed that all trains of the proposed MLP model have been accomplished successfully. This means that in all cases the training accuracy curves are increased acceptably until accessing very high values. The testing accuracy performances are also taken care of during the training stage at each epoch as they are represented in the same figures. Their curves are increased along training epochs comparatively to the training curves. However, their performances differ among different cases. The MSARA dataset group has the best accuracies.

4.4.2 CNN Training and Testing Results

Figure 4.27 to Figure 4.31 provides the training and testing curves in addition to the loss curves of the proposed CNN model for the OSARA dataset group and for all five lengths of samples (3, 4, 5, 6, and 7 seconds). Similarly, figure 4.32 to Figure 4.36 shows the training and testing curves in addition to the loss curves of the CNN model for the FSARA dataset group and all five lengths. Likewise, Figure 4.37 to Figure 4.41 displays the training and testing curves in addition to the loss curves of the proposed CNN for the MSARA dataset group and all five lengths too.

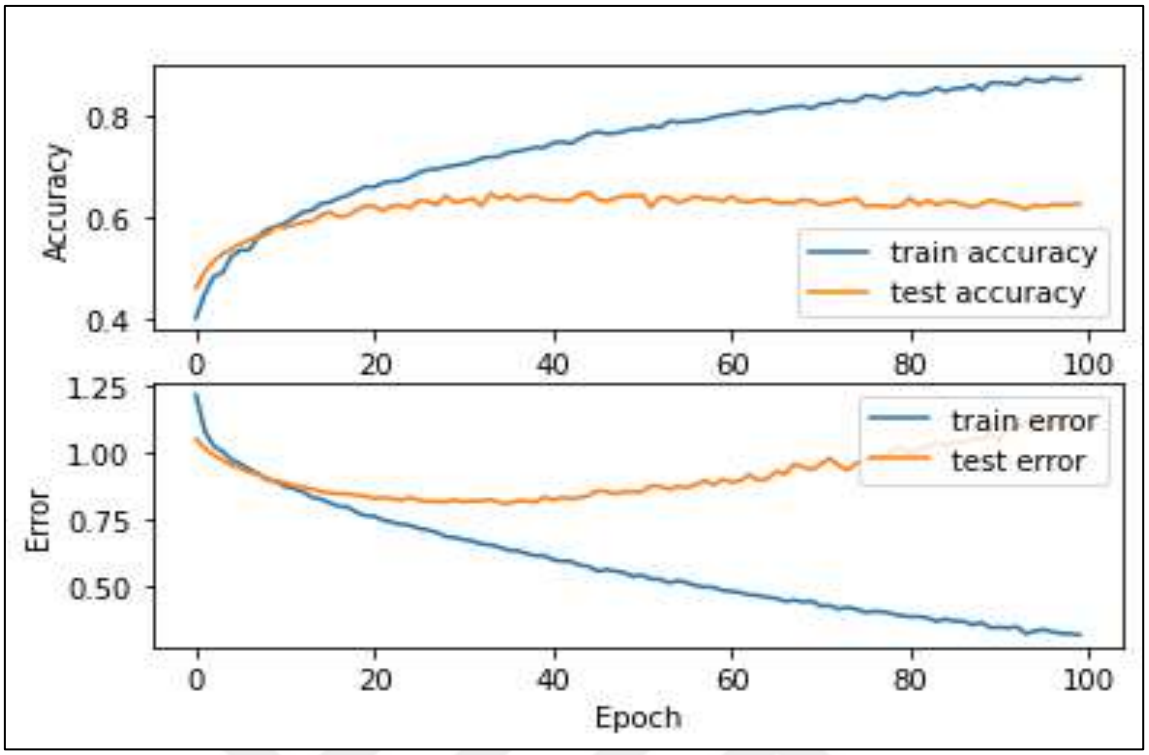

Figure 4.27: CNN Training and Testing Curves with Loss Curves for the OSARA Dataset Group and for the Length of 3 Seconds Samples.

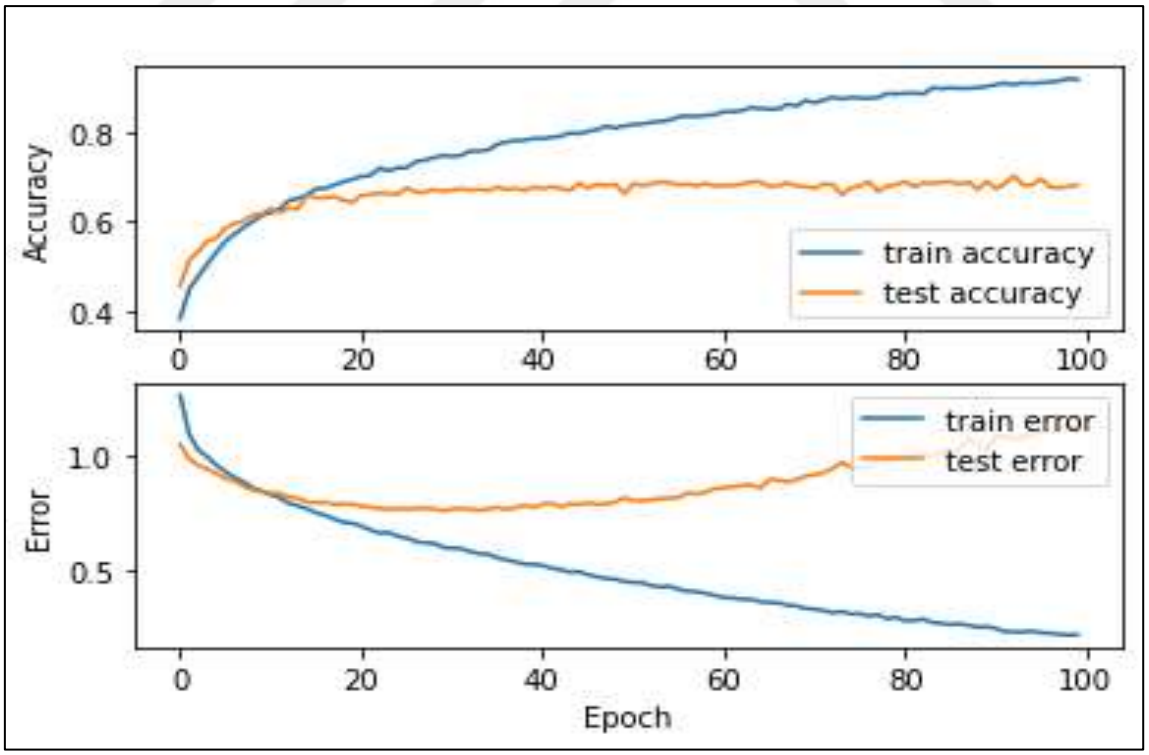

Figure 4.28: CNN Training and Testing Curves with Loss Curves for the OSARA Dataset Group and for the Length of 4 Seconds Samples.

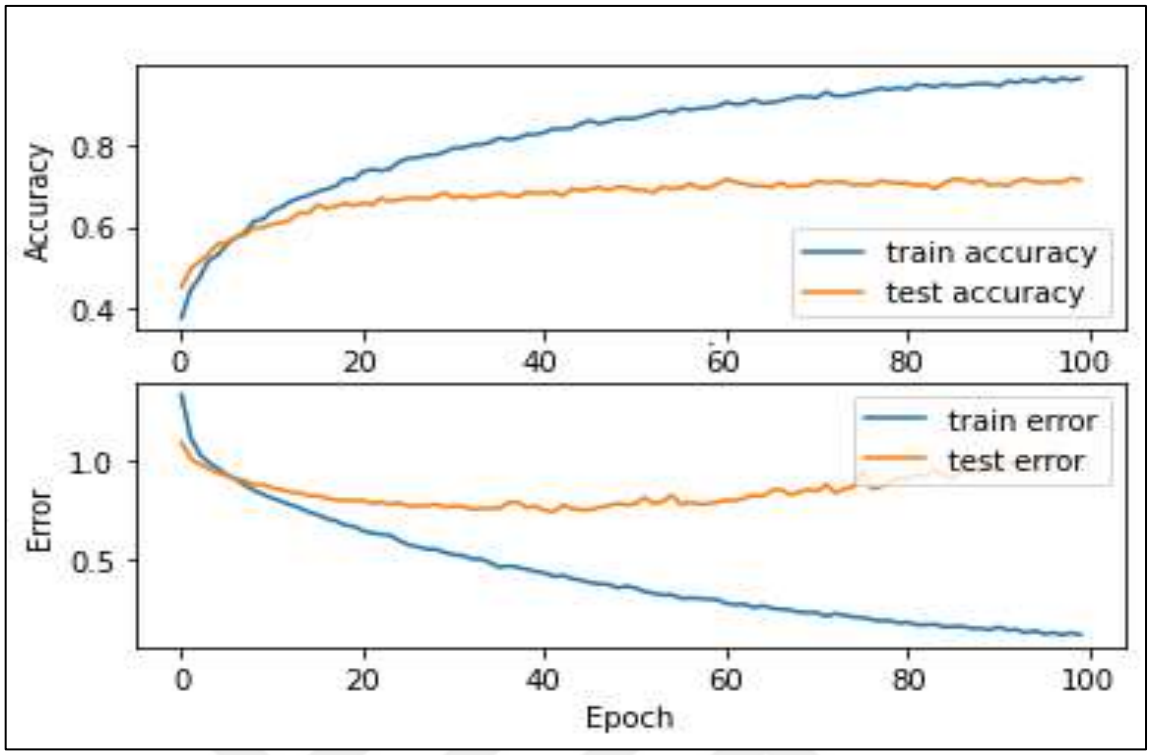

Figure 4.29: CNN Training and Testing Curves with Loss Curves for the OSARA Dataset Group and for the Length of 5 Seconds Samples.

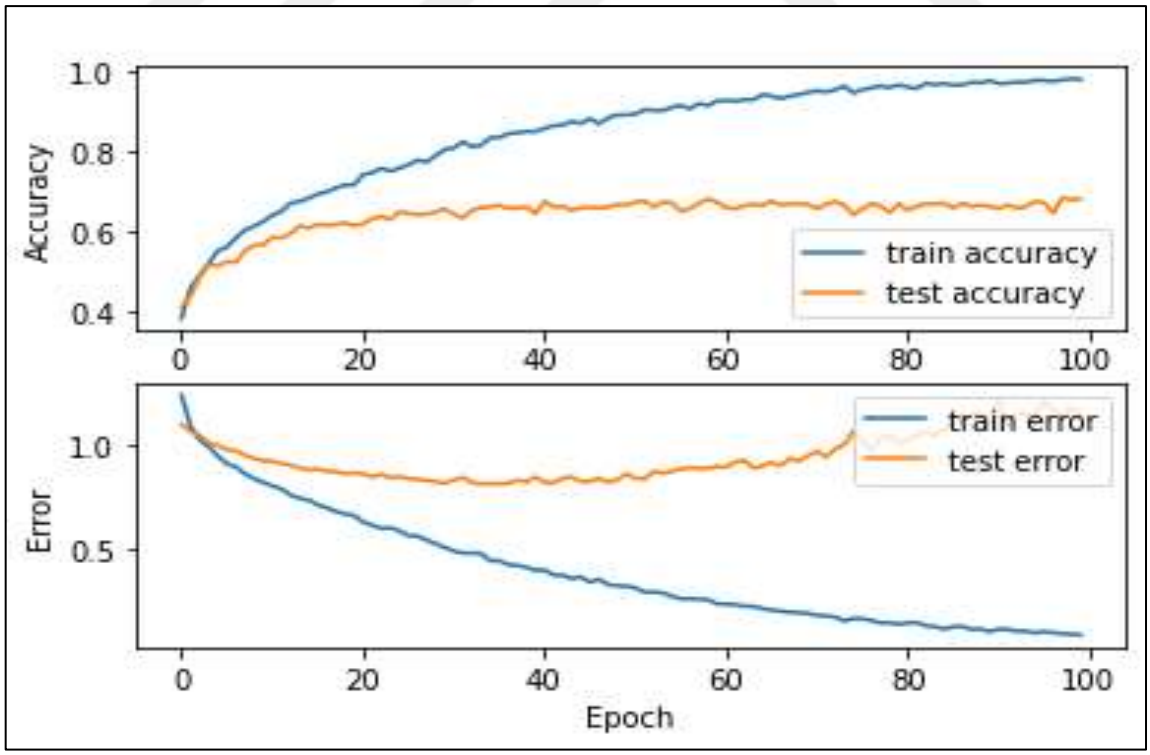

Figure 4.30: CNN Training and Testing Curves with Loss Curves for the OSARA Dataset Group and for the Length of 6 Seconds Samples.

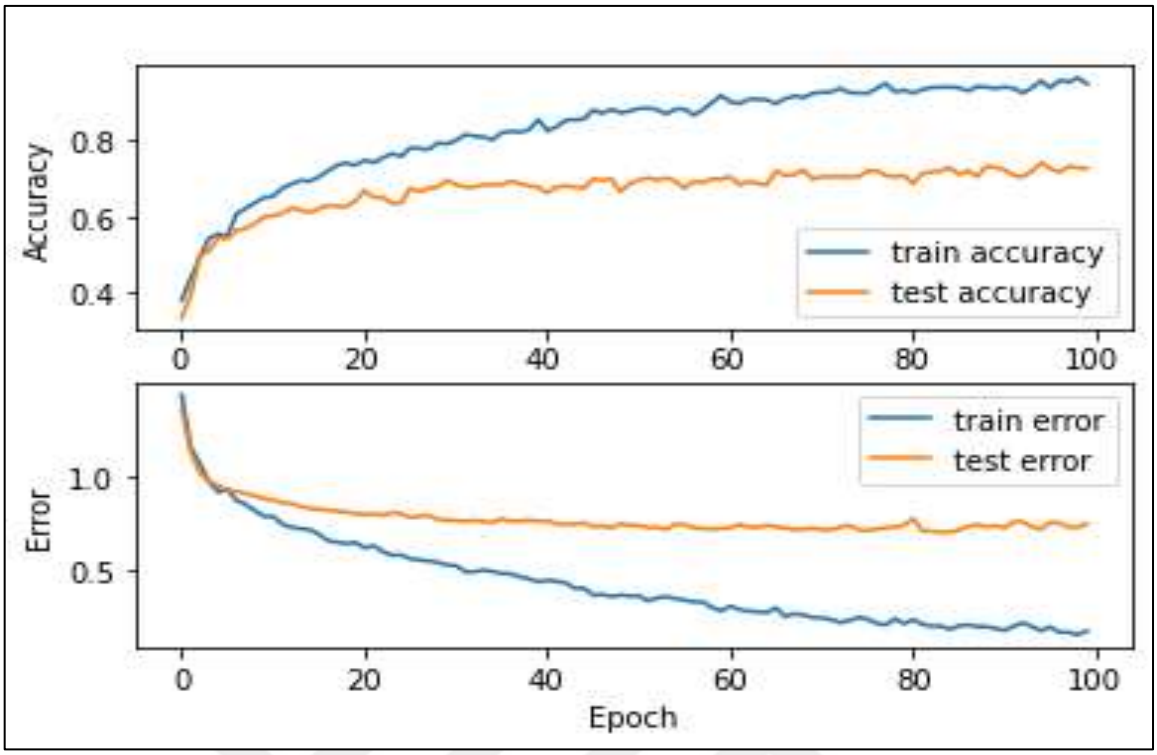

Figure 4.31: CNN Training and Testing Curves with Loss Curves for the OSARA Dataset Group and for the Length of 7 Seconds Samples.

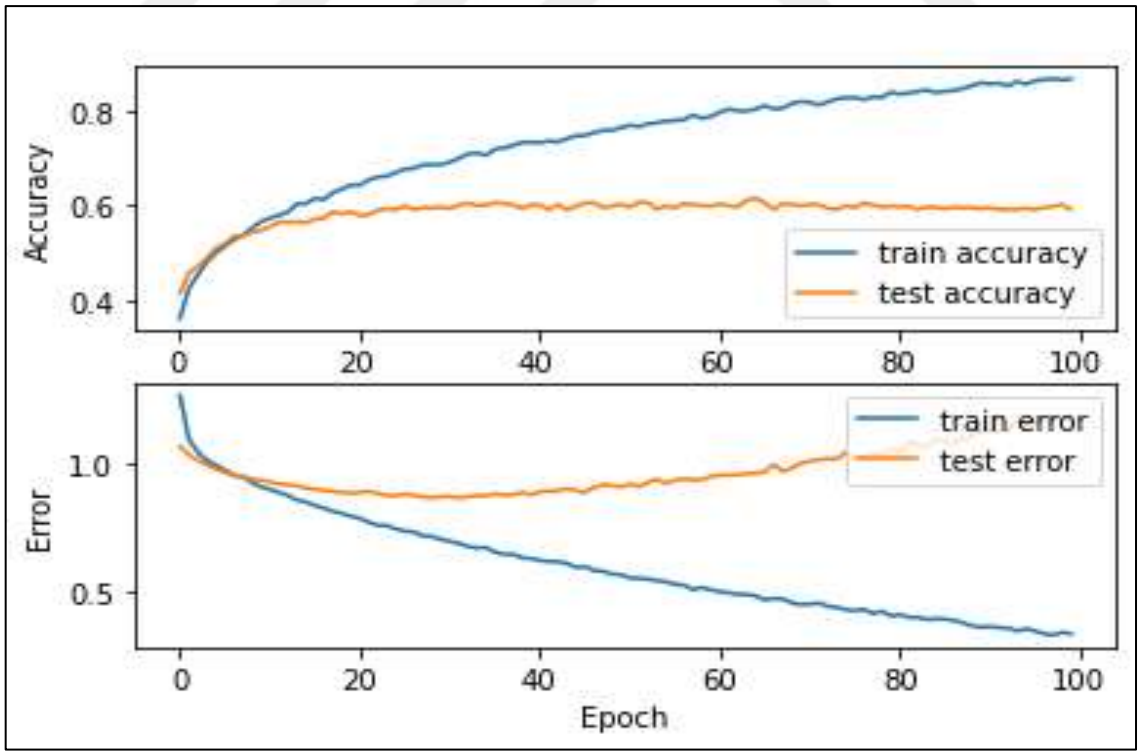

Figure 4.32: CNN Training and Testing Curves with Loss Curves for the FSARA Dataset Group and for the Length of 3 Seconds Samples.

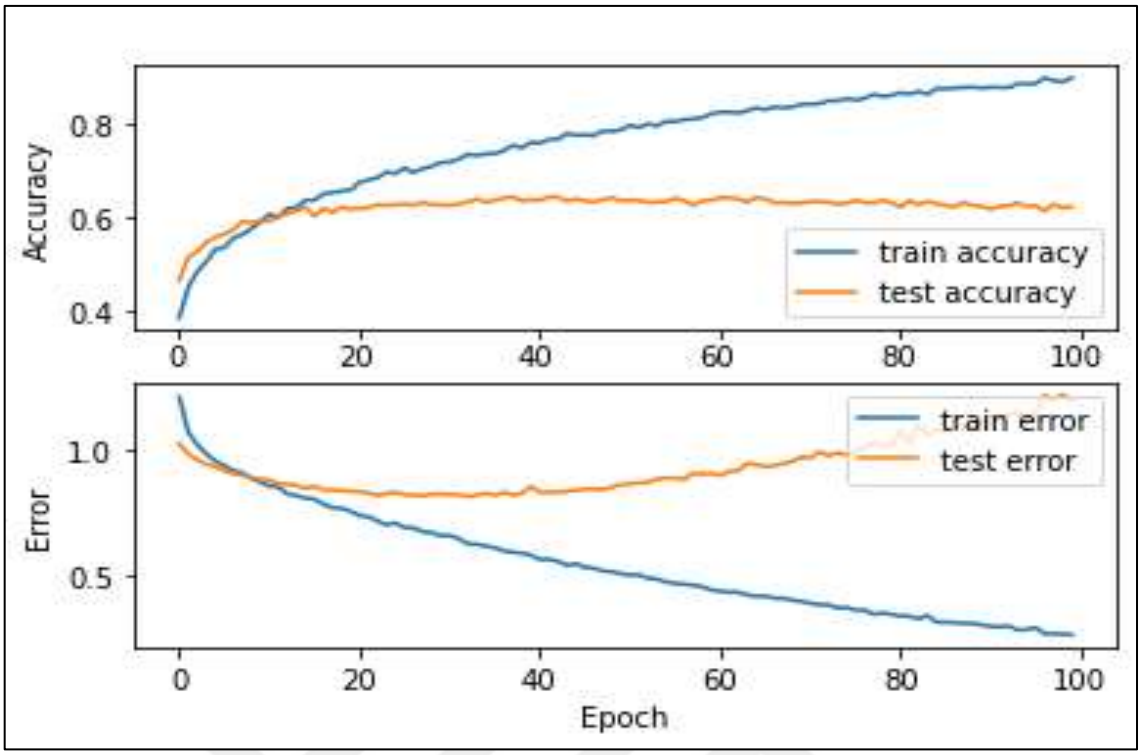

Figure 4.33: CNN Training and Testing Curves with Loss Curves for the FSARA Dataset Group and for the Length of 4 Seconds Samples.

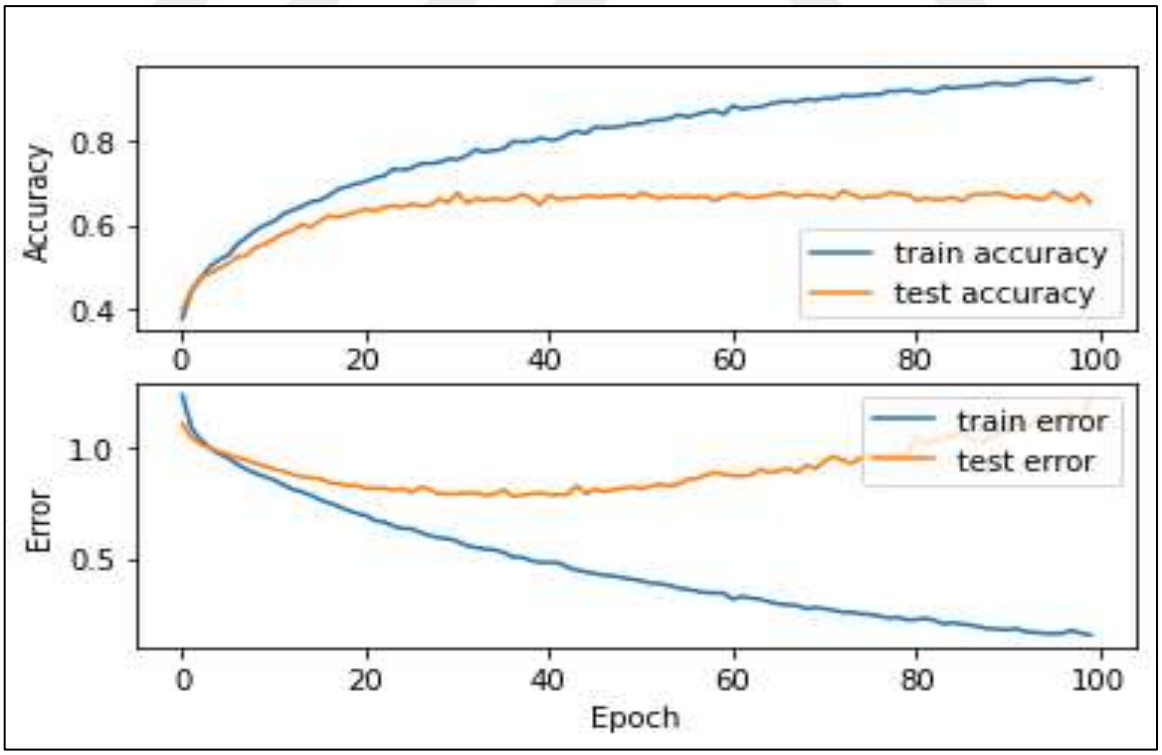

Figure 4.34: CNN Training and Testing Curves with Loss Curves for the FSARA Dataset Group and for the Length of 5 Seconds Samples.

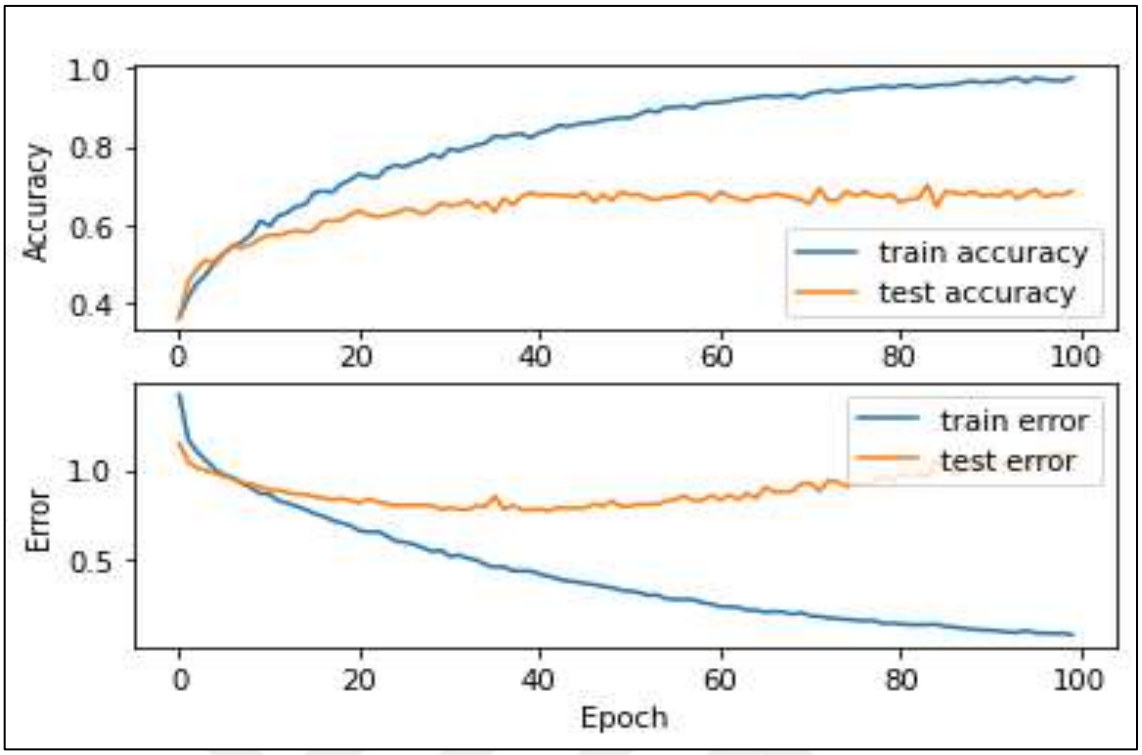

Figure 4.35: CNN Training and Testing Curves with Loss Curves for the FSARA Dataset Group and for the Length of 6 Seconds Samples.

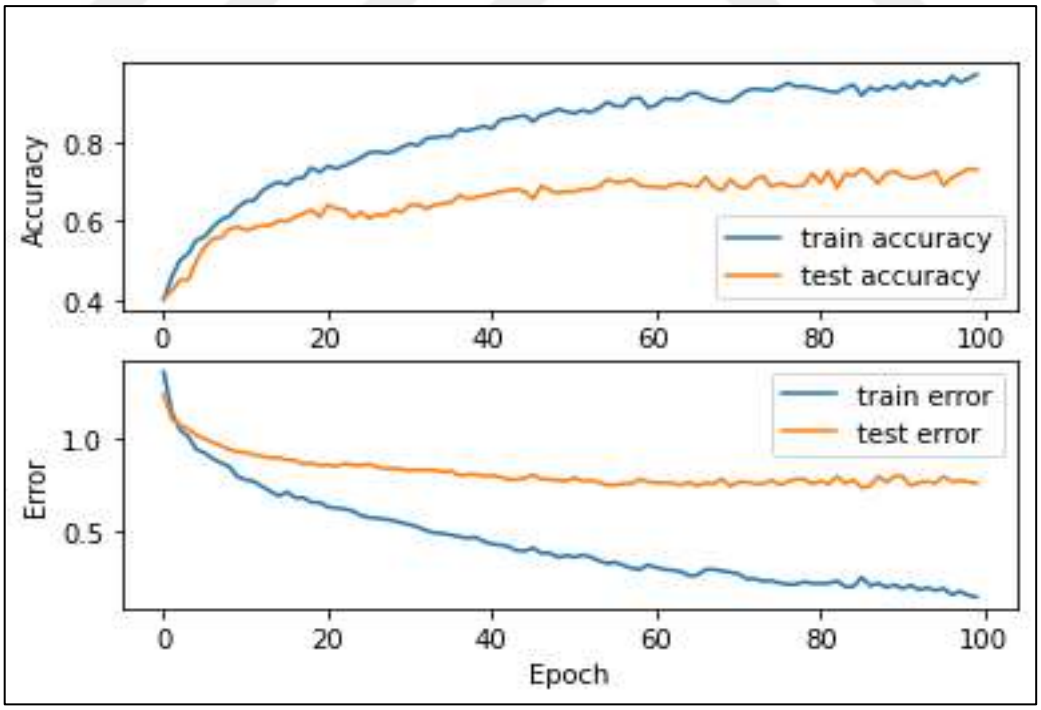

Figure 4.36: CNN Training and Testing Curves with Loss Curves for the FSARA Dataset Group and for the Length of 7 Seconds Samples.

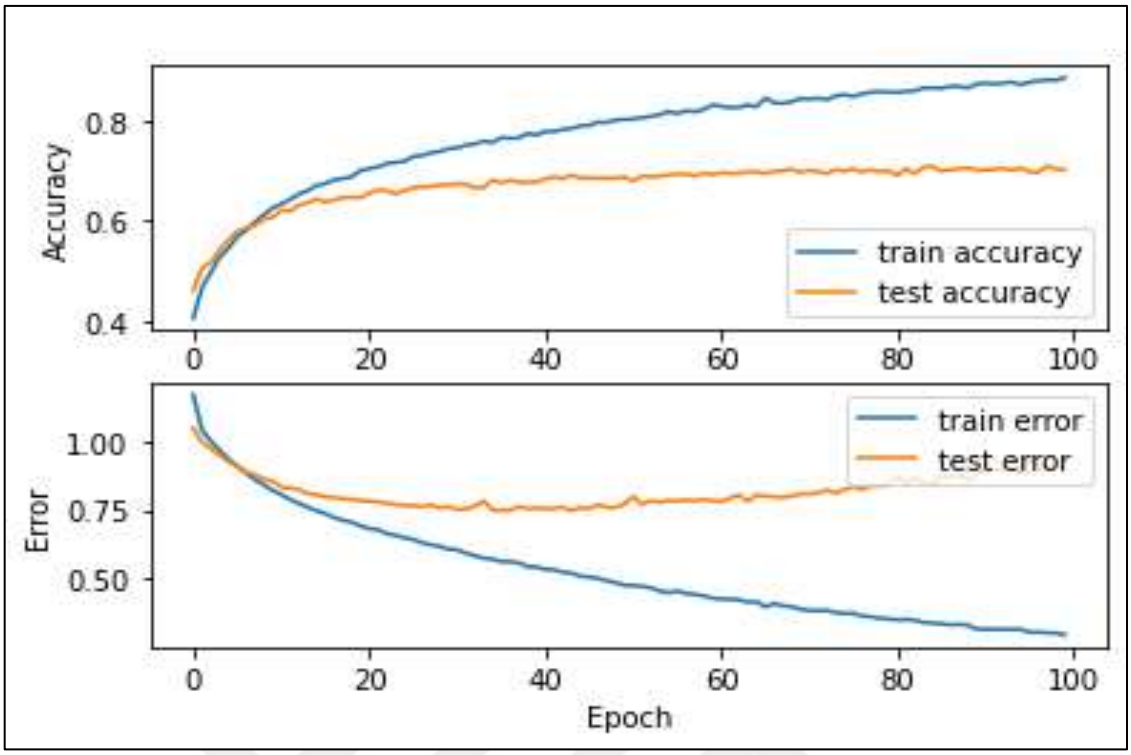

Figure 4.37: CNN Training and Testing Curves with Loss Curves for the MSARA Dataset Group and for the Length of 3 Seconds Samples.

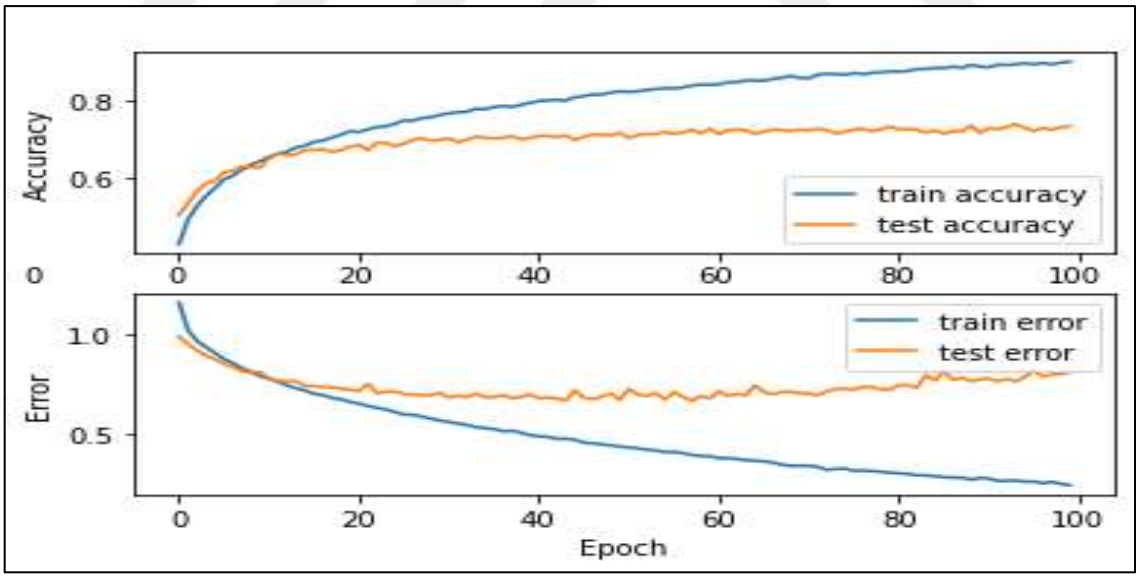

Figure 4.38: CNN Training and Testing Curves with Loss Curves for the MSARA Dataset Group and for the Length of 4 Seconds Samples.

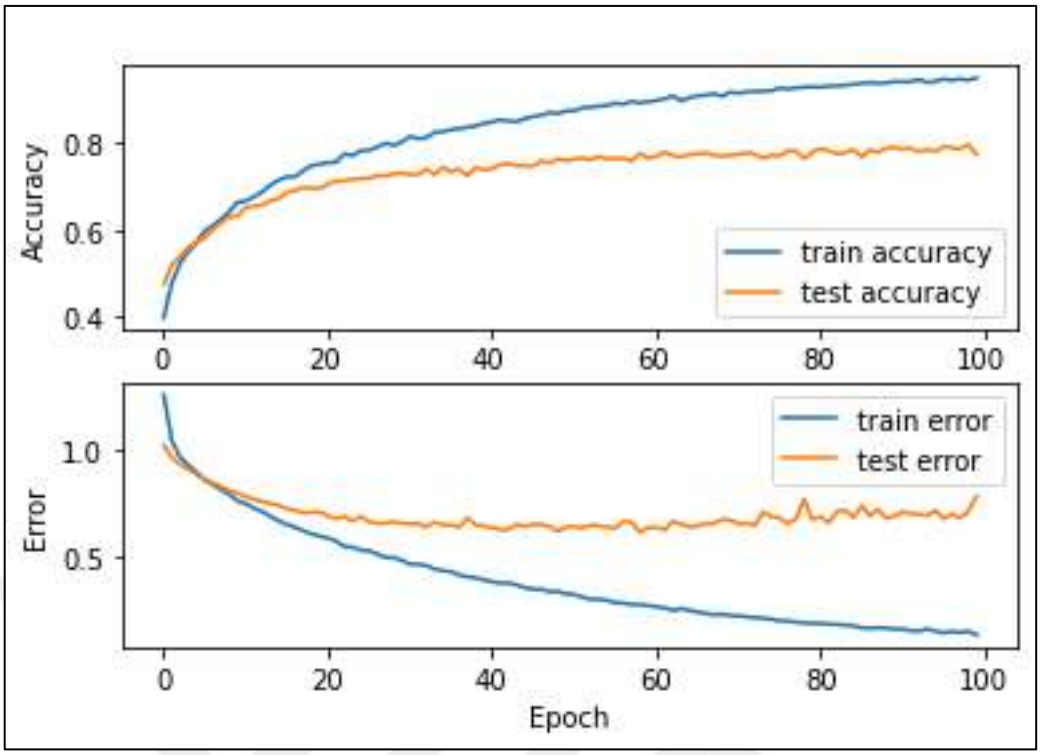

Figure 4.39: CNN Training and Testing Curves with Loss Curves for the MSARA Dataset Group and for the Length of 5 Seconds Samples.

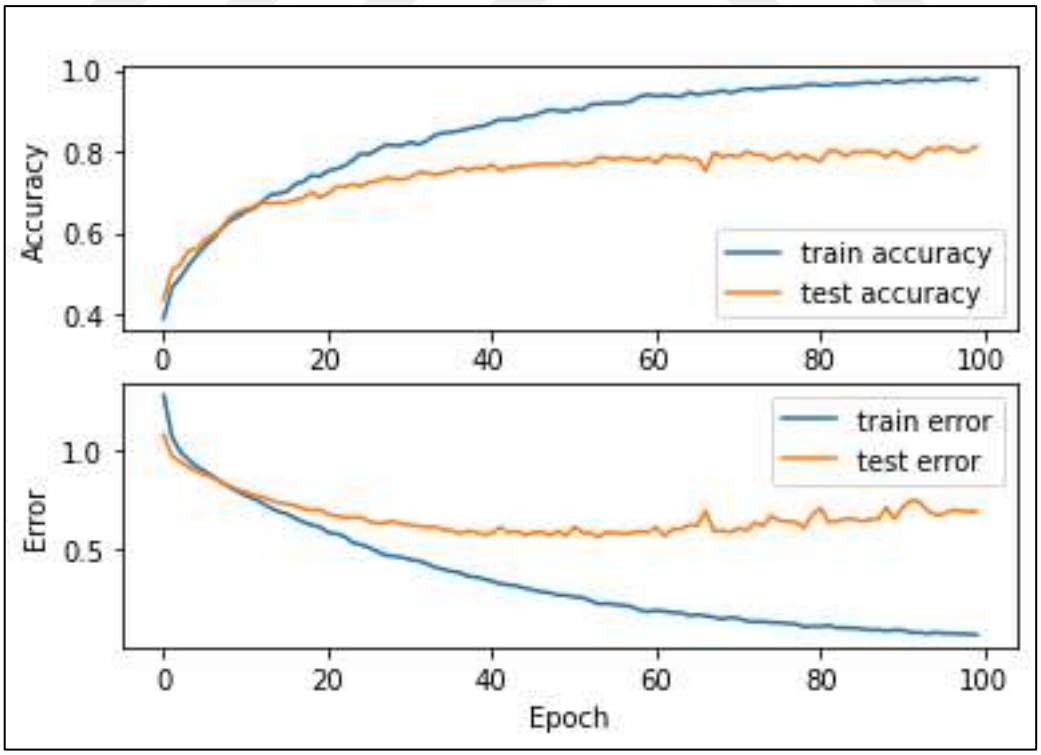

Figure 4.40: CNN Training and Testing Curves with Loss Curves for the MSARA Dataset Group and for the Length of 6 Seconds Samples.

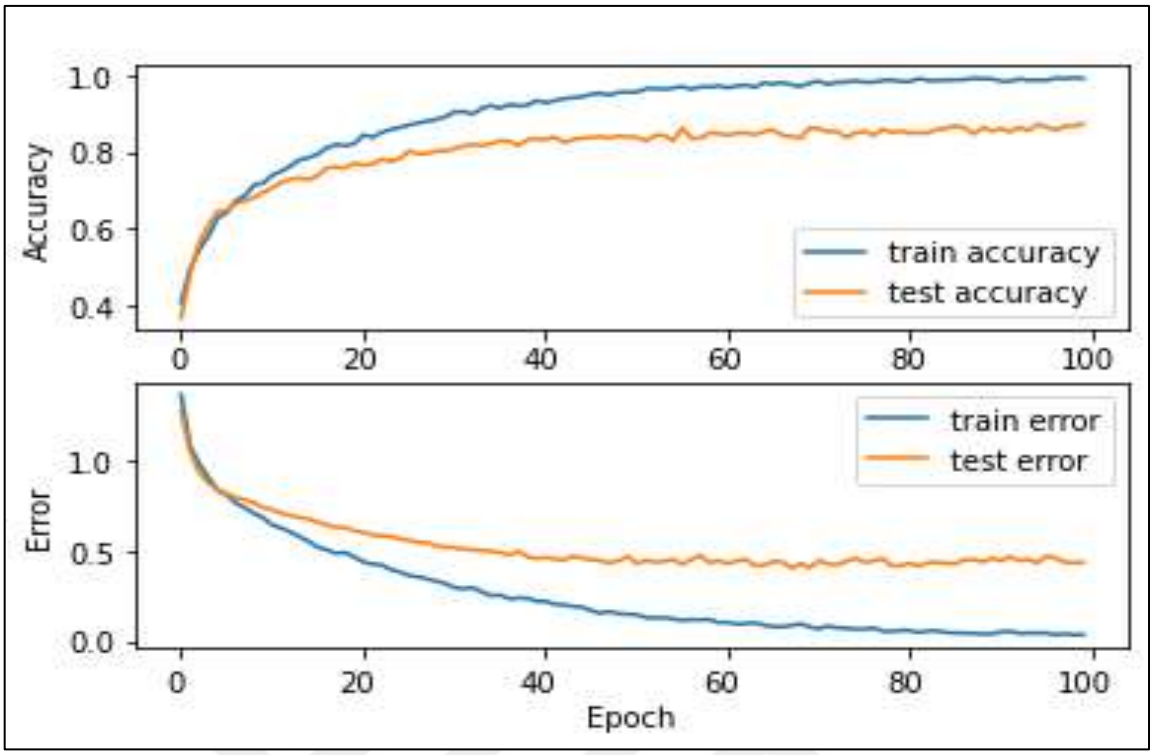

Figure 4.41: CNN Training and Testing Curves with Loss Curves for the MSARA Dataset Group and for the Length of 7 Seconds Samples.

From these figures, it can be investigated that all trains of the proposed CNN model have been reported positively too. This means that in all cases the training accuracy curves are reasonably improved until reach very high values. The testing accuracies are also taken care of during the training stage at each epoch as they are represented in the same figures. Their curves are increased along training epochs consistently to the training curves. Nevertheless, their performances vary among various cases. The MSARA dataset group again proves its ability to reach high accuracies.

4.4.3 DRNN Training and Testing Results

Figure 4.42 to Figure 4.46 provides the training and testing curves in addition to the loss curves of the proposed DRNN model for the OSARA dataset group and for all five lengths of samples (3, 4, 5, 6, and 7 seconds). Similarly, Figure 4.47 to Figure 4.51 show the training and testing curves in addition to the loss curves of the DRNN model for the FSARA dataset group and all five lengths. Likewise, Figure 4.52 to Figure 4.56 displays the training and testing curves in addition to the loss curves of the proposed DRNN for the MSARA dataset group and all the five lengths too.

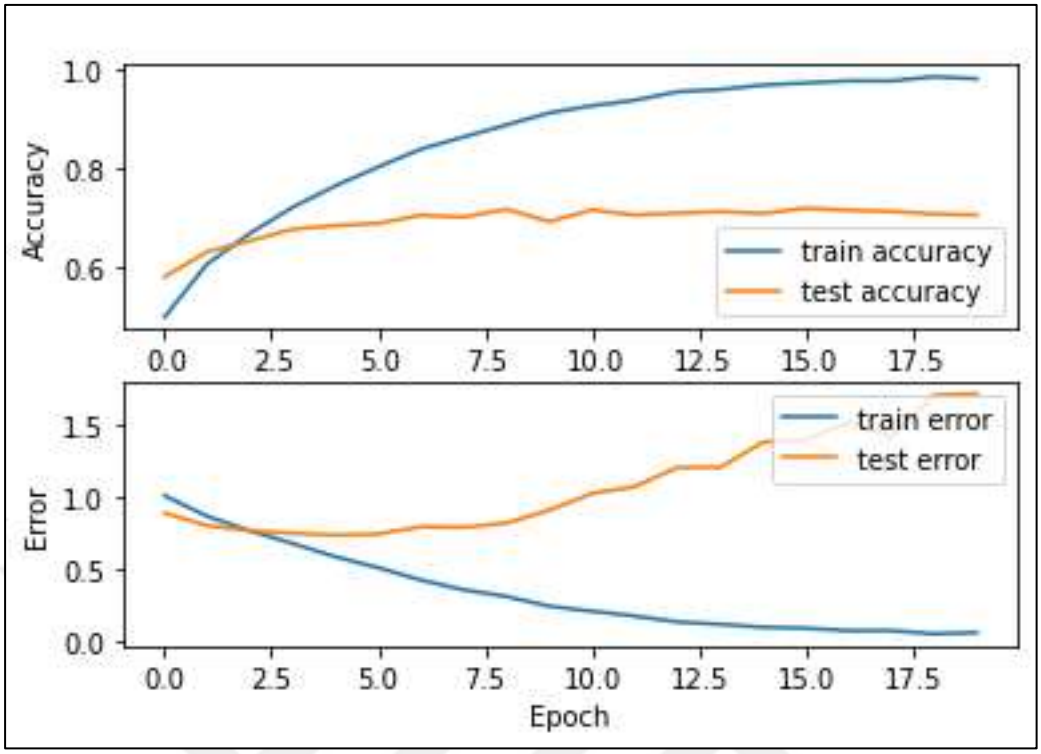

Figure 4.42: DRNN Training and Testing Curves with Loss Curves for the OSARA Dataset Group and for the Length of 3 Seconds Samples.

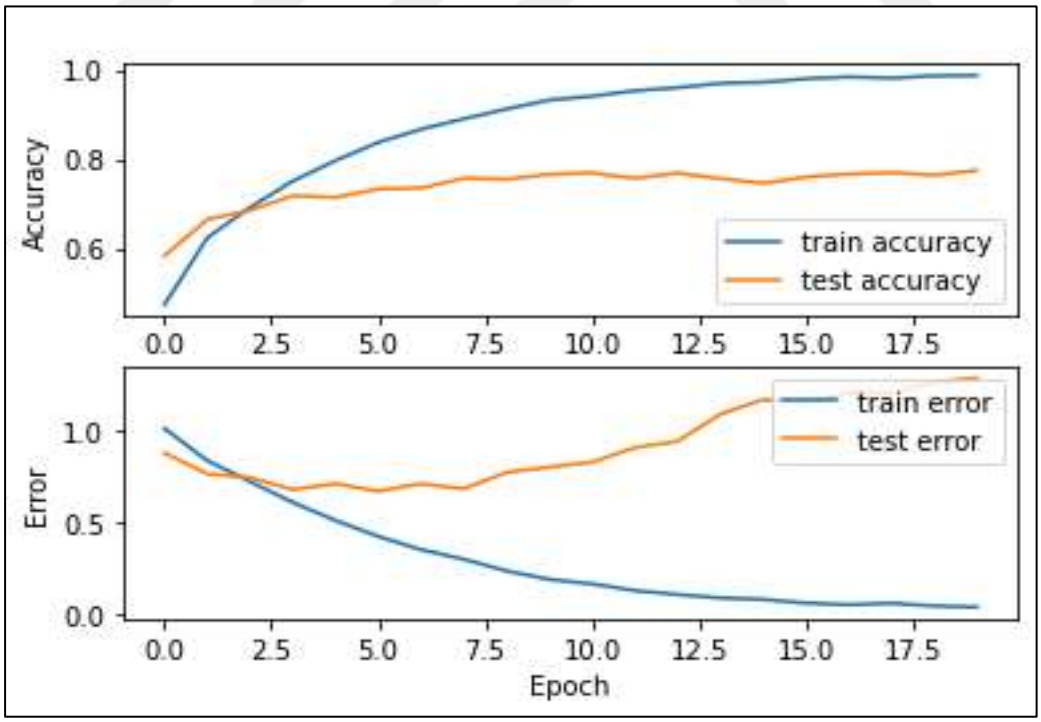

Figure 4.43: DRNN Training and Testing Curves with Loss Curves for the OSARA Dataset Group and for the Length of 4 Seconds Samples.

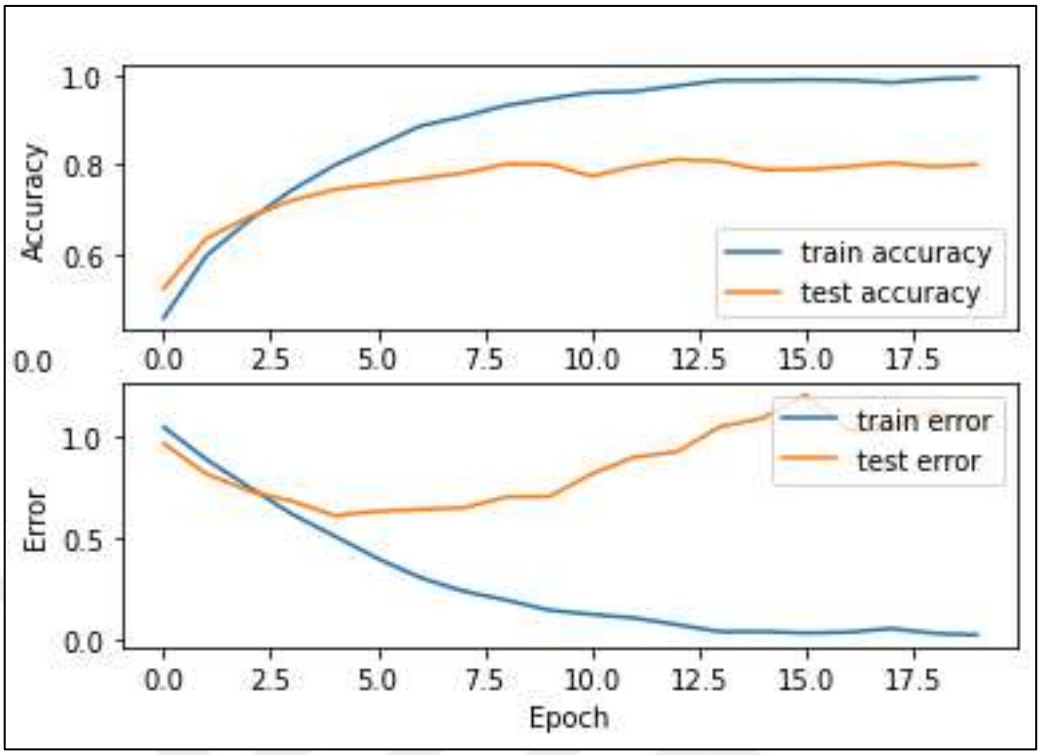

Figure 4.44: DRNN Training and Testing Curves with Loss Curves for the OSARA Dataset Group and for the Length of 5 Seconds Samples.

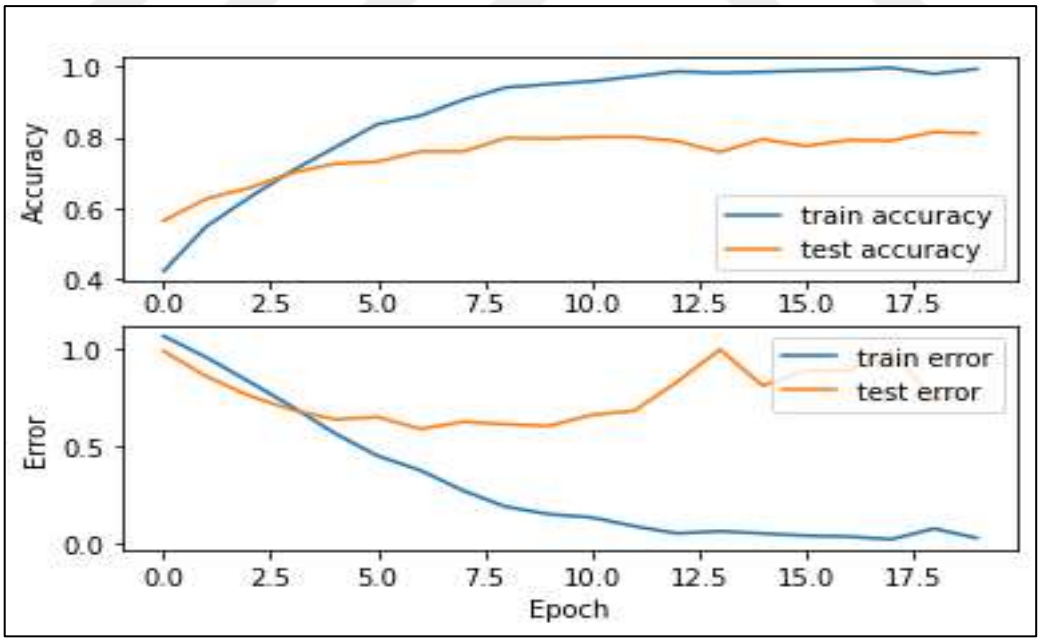

Figure 4.45: DRNN Training and Testing Curves with Loss Curves for the OSARA Dataset Group and for the Length of 6 Seconds Samples.

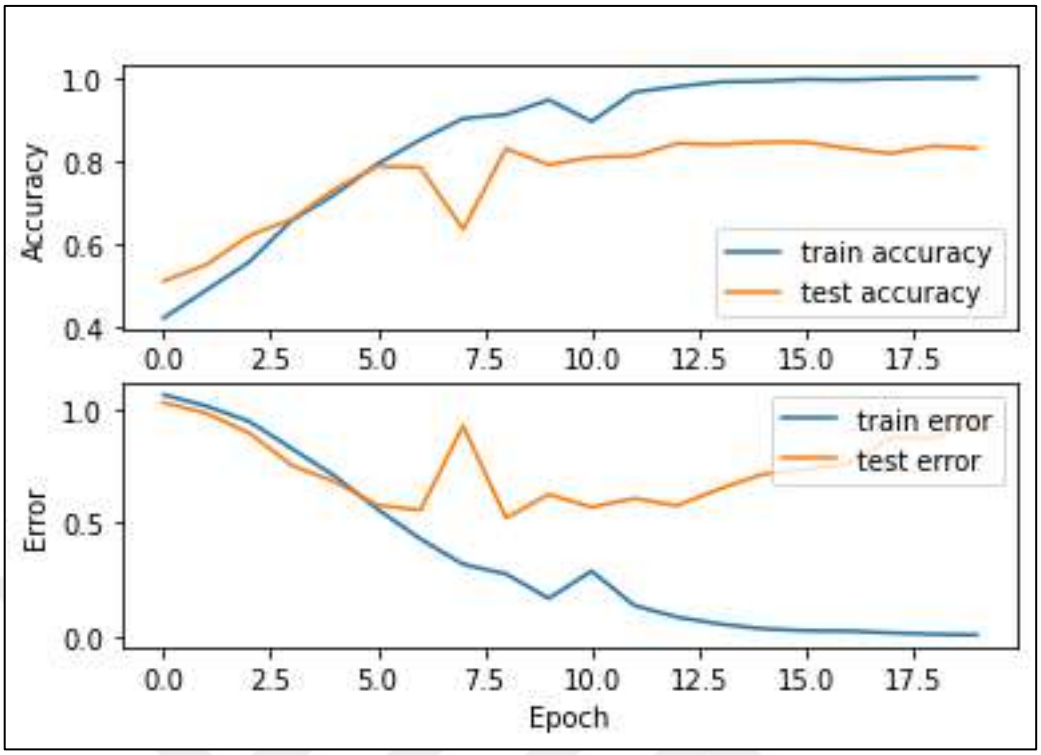

Figure 4.46: DRNN Training and Testing Curves with Loss Curves for the OSARA Dataset Group and for the Length of 7 Seconds Samples.

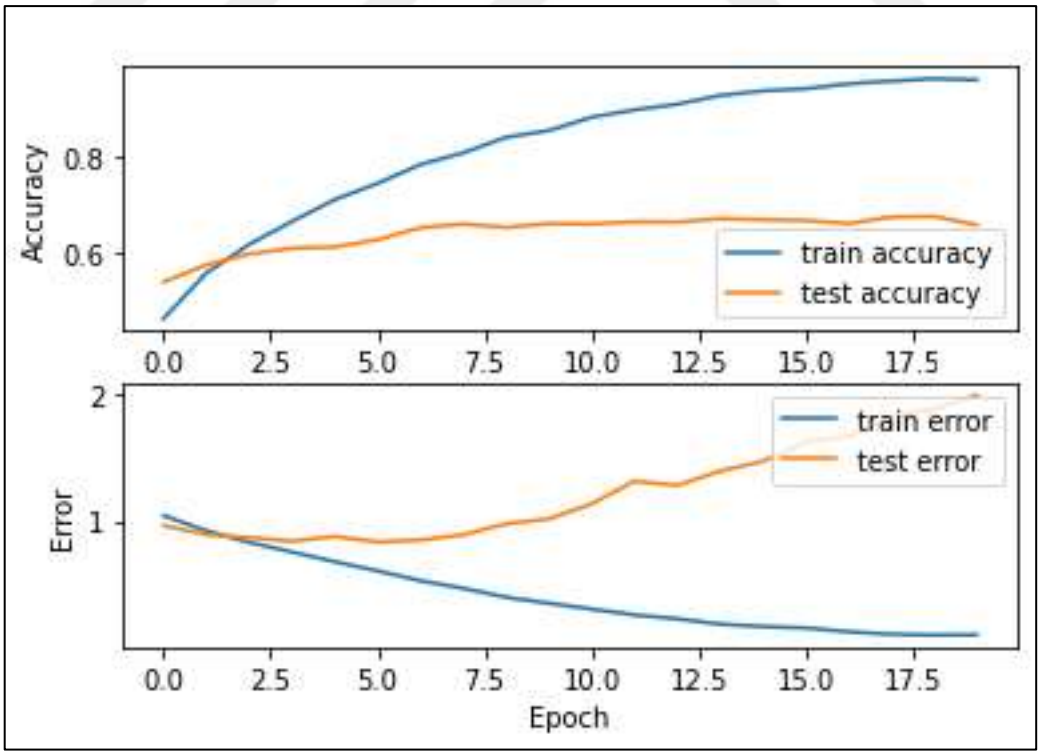

Figure 4.47: DRNN Training and Testing Curves with Loss Curves for the FSARA Dataset Group and for the Length of 3 Seconds Samples.

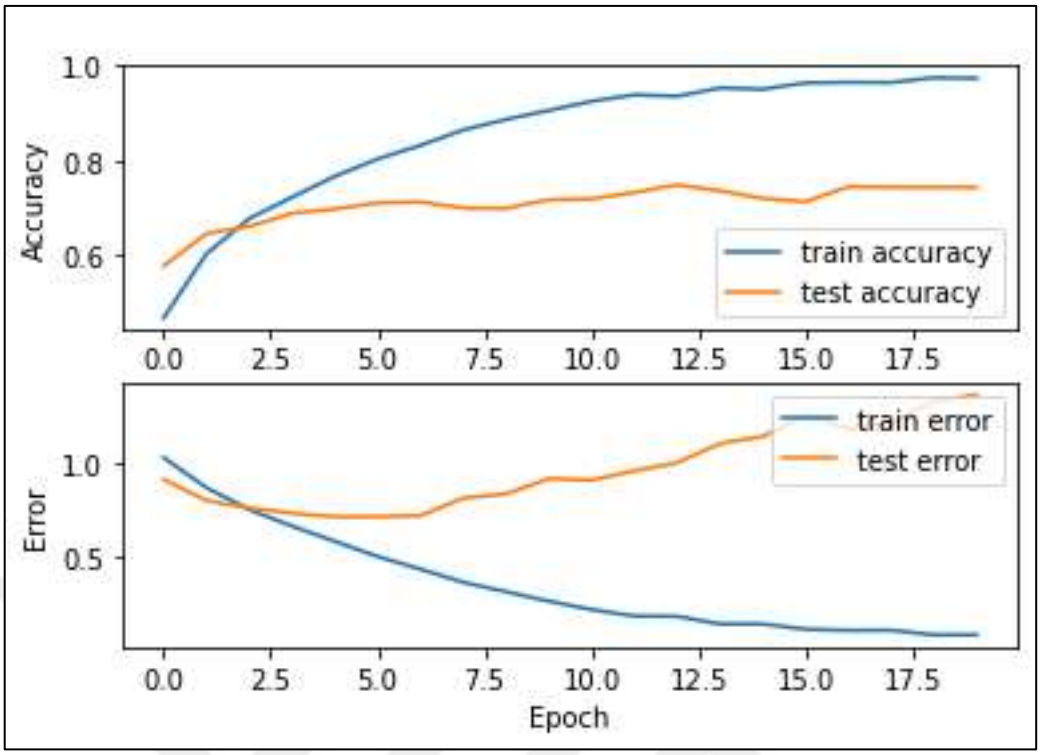

Figure 4.48: DRNN Training and Testing Curves with Loss Curves for the FSARA Dataset Group and for the Length of 4 Seconds Samples.

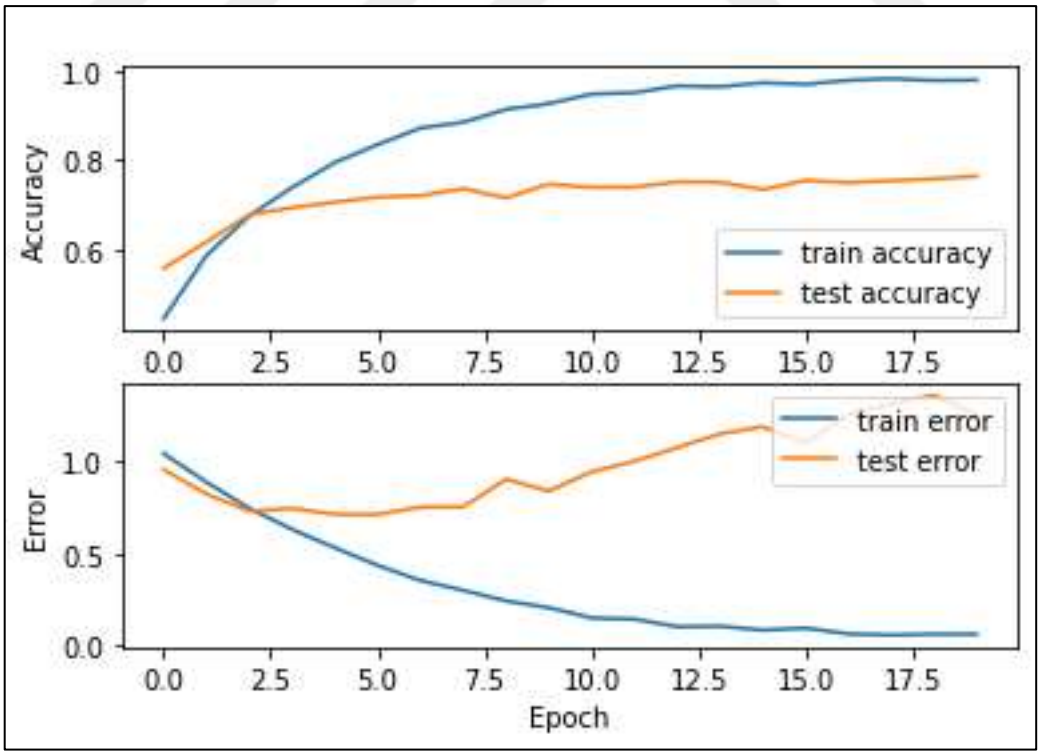

Figure 4.49: DRNN Training and Testing Curves with Loss Curves for the FSARA Dataset Group and for the Length of 5 Seconds Samples.

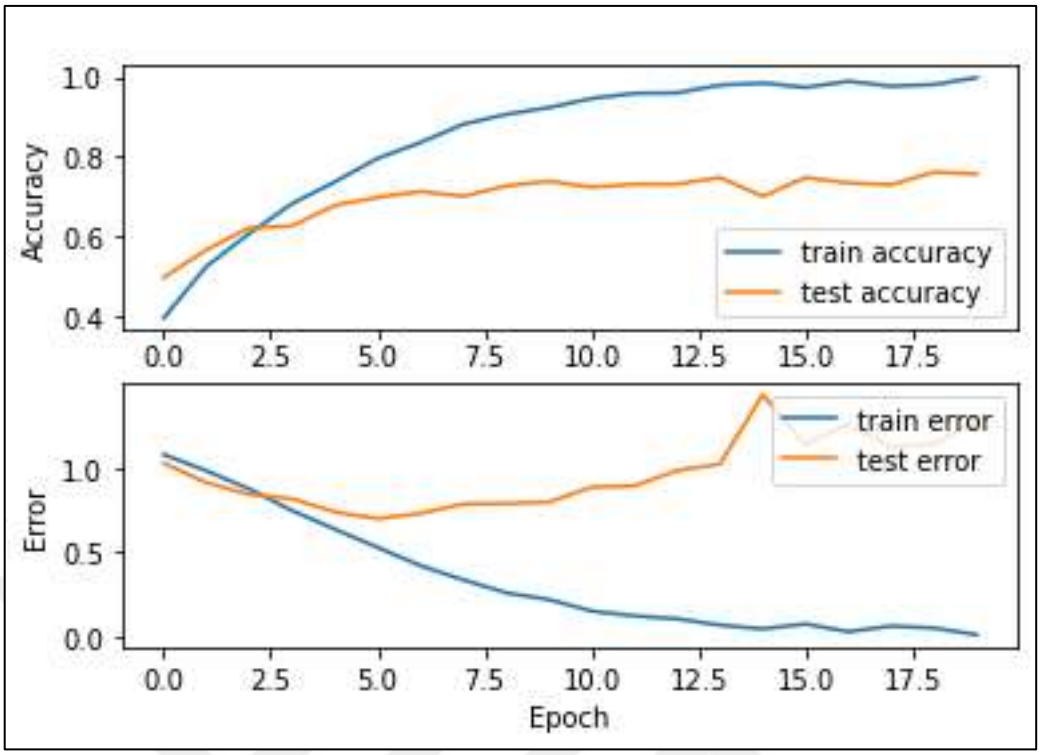

Figure 4.50: DRNN Training and Testing Curves with Loss Curves for the FSARA Dataset Group and for the Length of 6 Seconds Samples.

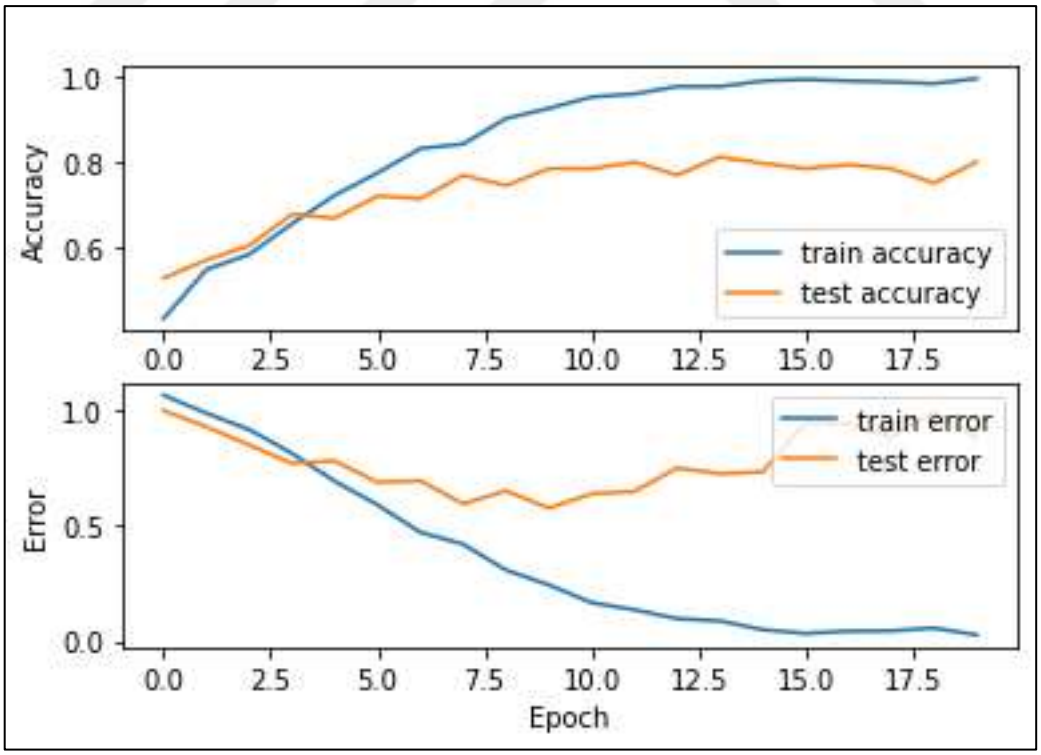

Figure 4.51: DRNN Training and Testing Curves with Loss Curves for the FSARA Dataset Group and for the Length of 7 Seconds Samples.

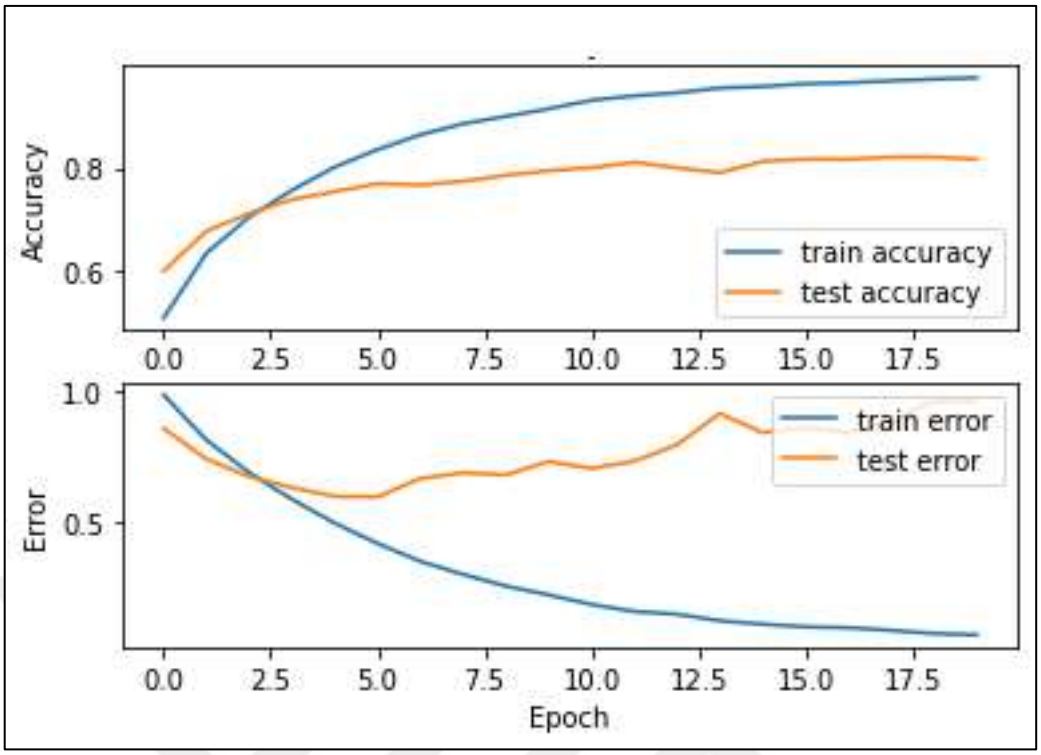

Figure 4.52: DRNN Training and Testing Curves with Loss Curves for the MSARA Dataset Group and for the Length of 3 Seconds Samples.

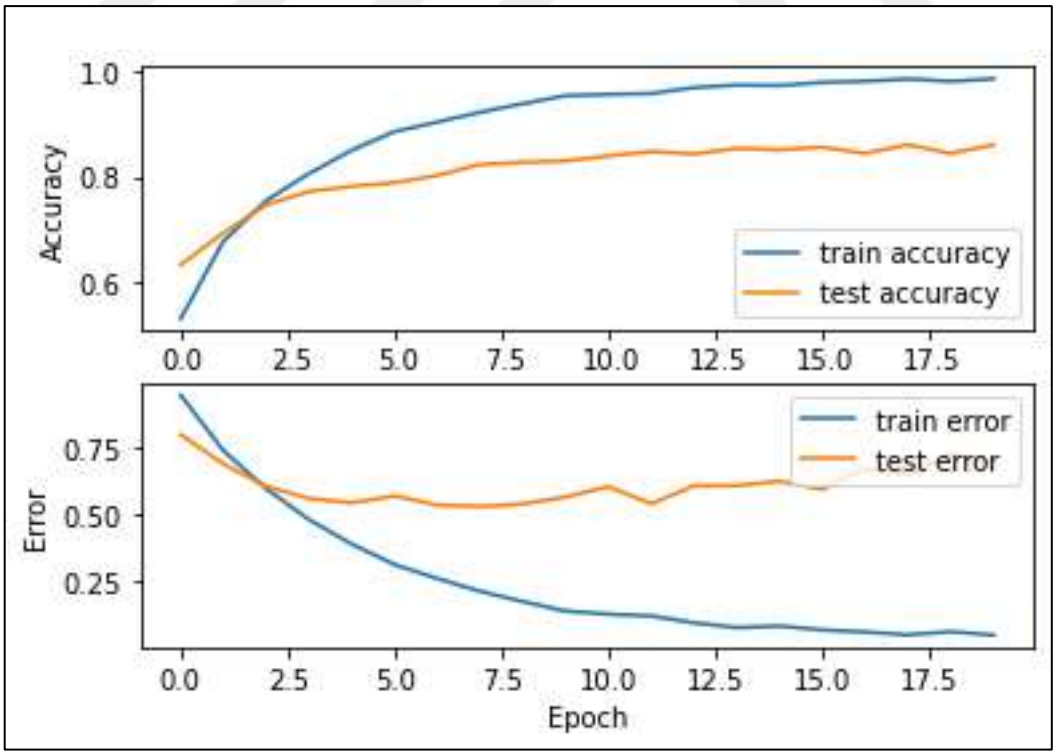

Figure 4.53: DRNN Training and Testing Curves with Loss Curves for the MSARA Dataset Group and for the Length of 4 Seconds Samples.

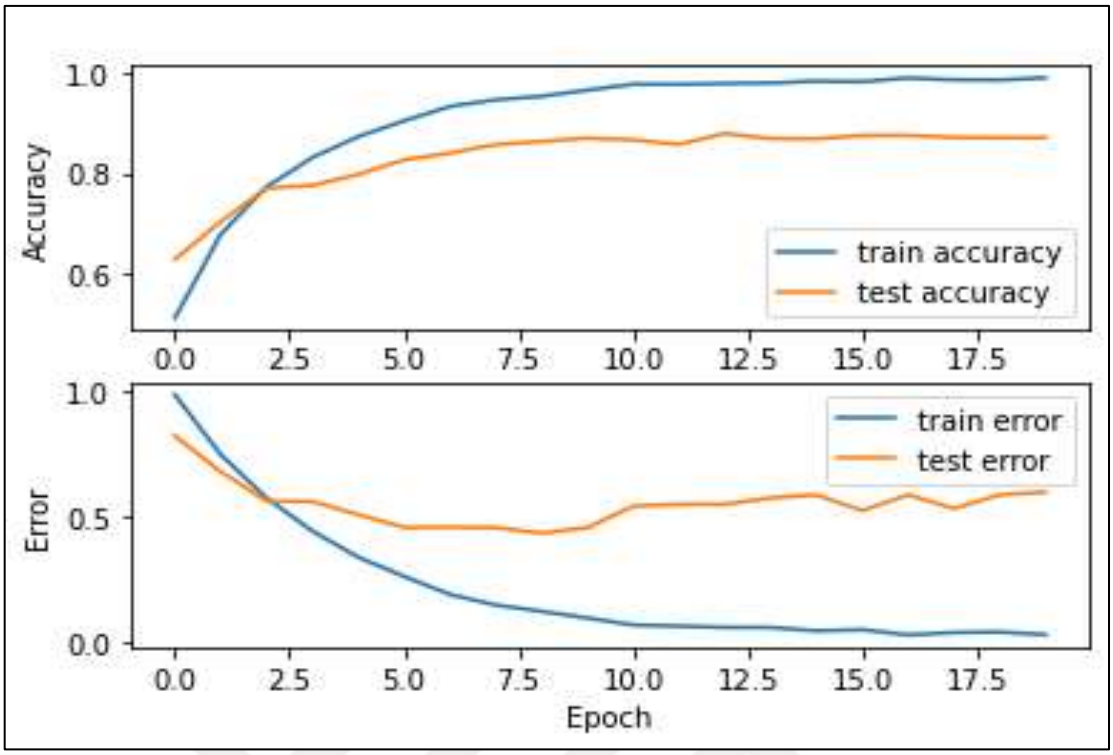

Figure 4.54: DRNN Training and Testing Curves with Loss Curves for the MSARA Dataset Group and for the Length of 5 Seconds Samples.

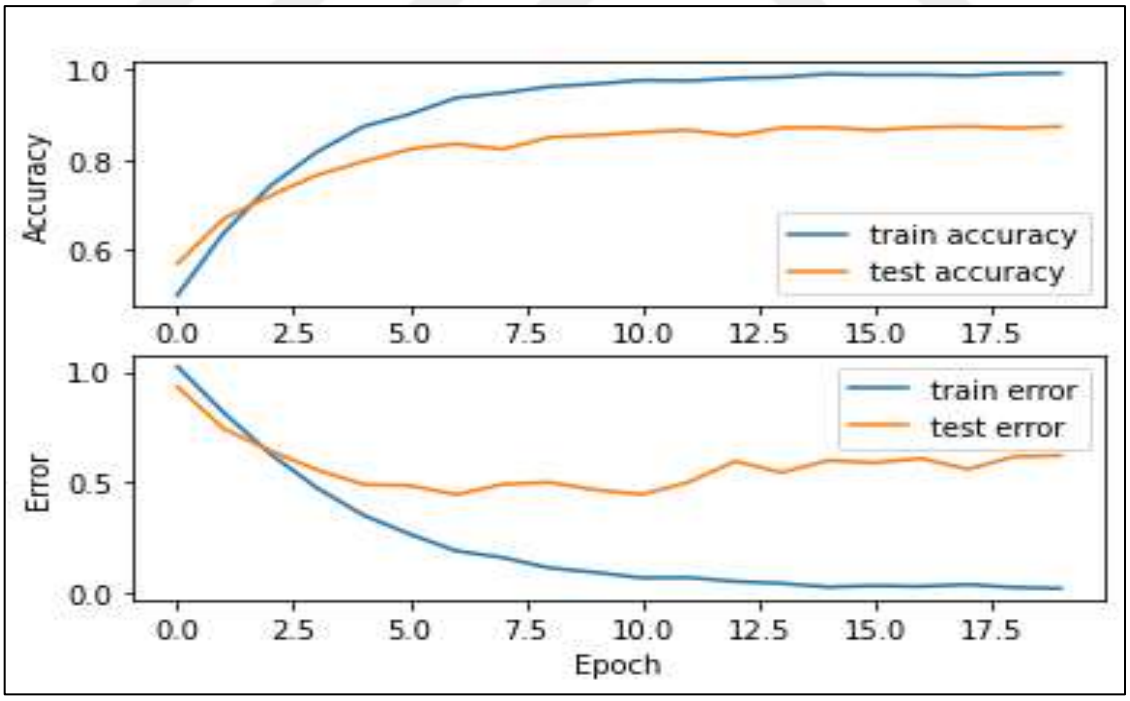

Figure 4.55: DRNN Training and Testing Curves with Loss Curves for the MSARA Dataset Group and for the Length of 6 Seconds Samples.

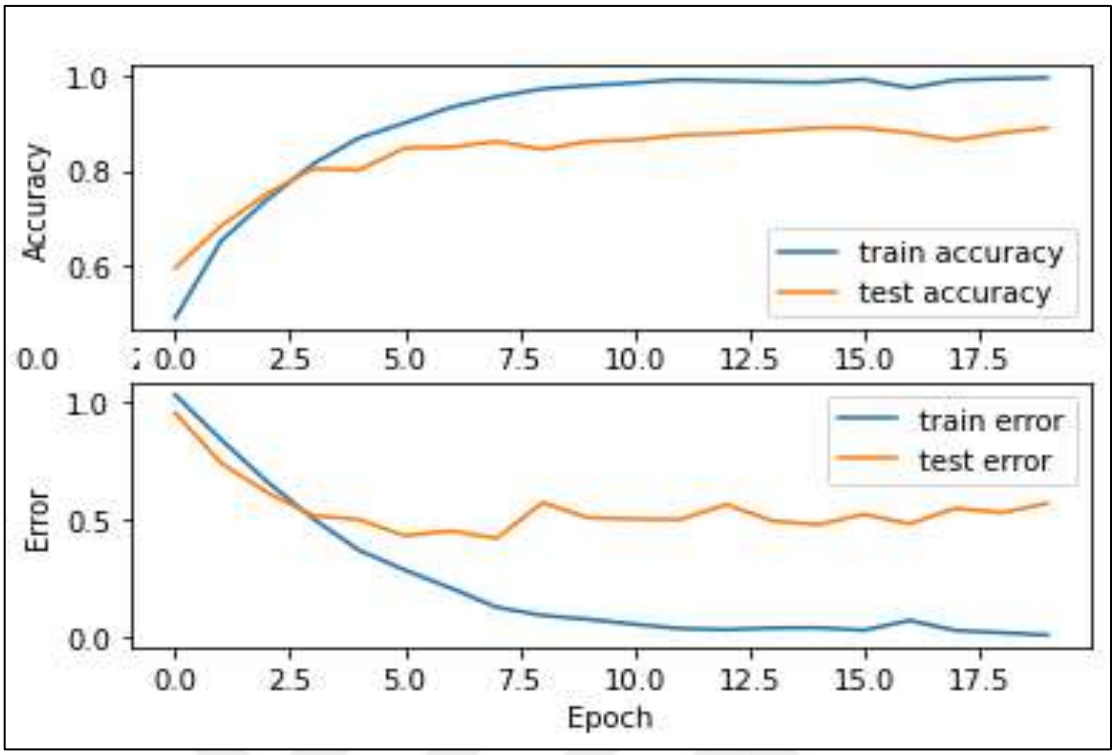

Figure 4.56: DRNN Training and Testing Curves with Loss Curves for the MSARA Dataset Group and for the Length of 7 Seconds Samples.

From these figures, it can be noticed that all trains of the proposed DRNN model have successfully been achieved too. This means that, in all cases, the training accuracy curves are increased successfully till getting very high values. The testing performances are also taken into consideration during the training stage at each epoch as they are represented in the same figures. Their curves are improved along training epochs according to the training curves. Nonetheless, their performances vary among various cases. Again, the MSARA dataset group has confirmed its capability of recording high accuracies.

In addition, the proposed DRNN model has generally reported the best performances. Its accuracies in both training and testing are worthy of attention compared to the proposed MLP and CNN models. This encourages us to select this model DRNN as the best proposed one in this work.

5. DISCUSSION AND CONCLUSIONS

5.1 MLP MODEL

The proposed MLP model has been tested for the three dataset groups of OSARA, FSARA, and MSARA. Moreover, all five lengths of samples: 3, 4, 5, 6, and 7 seconds have been taken into consideration. Table 5.1 displays the testing performances of the MLP model for all the cases.

Table 5.1: Testing Performances of the Proposed MLP Model for the Three Dataset Groups (OSARA, FSARA, and MSARA) and all the Lengths of Samples (3, 4, 5, 6, and 7)

Sample Length		**3 seconds**		**4 seconds**		**5 seconds**		**6 seconds**		**7 seconds**	
Accuracy & Loss		**Accuracy (%)**	**Loss**	**Accuracy (%)**	**Loss**	**Accuracy (%)**	**Loss**	**Accuracy (%)**	**Loss**	**Accuracy (%)**	**Loss**
Dataset Group	**OSARA**	**70.7**	**1.47**	**72.8**	**1.36**	**72.1**	**1.6**	**71.2**	**1.99**	**72.7**	**2.35**
	FSARA	**65**	**1.64**	**66.9**	**1.64**	**69.3**	**1.38**	**68.3**	**1.82**	**71.2**	**2.68**
	MSARA	**85.6**	**0.67**	**86.5**	**0.71**	**88**	**0.71**	**87.2**	**0.79**	**89.5**	**0.91**

From Table 5.1, it can be discovered that the OSARA group of the used dataset has acquired slight variance among almost all the performances of the five lengths of samples. This means that the range of variances for the accuracy is [70.7%, 72.8%] and the range of variances for the loss is [1.36, 2.35]. When looking closely at the table these ranges have gotten larger for the FSARA group where the accuracy range is becoming [65%, 71.2%] and the loss range has been reached [1.38, 2.68]. The MSARA dataset group has achieved the best performances in all five lengths of samples. High and slight values for the differences among the accuracy results have been achieved as the accuracy range values are increased to [85.6%, 89.5%]. Whilst low and slight values for the differences of almost all the loss results have been attained where the range values are decreased to [0.67, 0.91]. These outcomes are due to the specialty of each dataset group. To illustrate, the OSARA has the raw or original data that cannot be analyzed well., the FSARA has the filtered data that may show big variances in evaluations, and the MSARA consists combination of the OSARA and FSARA where this combination gives it more data with specialty to be considered efficiently.

It is worth noting that the achievement of the 7-second sample length for any dataset group are greater than the achievement of the 3-second sample length for any dataset group. To illustrate, the accuracy was 89.5% and the loss was 0.91 when the 7-second sample length for the MSARA dataset group was used, whereas the accuracy was 85.6% and the loss is 0.67 when the 3-second sample length for the same dataset group was used. The reason behind this is that the length of samples becomes more effective if it is long as it includes lots of information that can be worthy of analyzing and providing important results. Although, the specific length consists a smaller number of samples the 7 seconds involves 438 samples for the MSARA dataset group, whereas, the 3 seconds includes 2508 samples for the same dataset group. Therefore, it can be concluded that higher performances (accuracy) are achieved when a longer length of samples is used even if the number of samples is smaller. Figure 5.1 shows the confusion matrix for the MLP model with the MSARA 7-second dataset that represents the best case in this model.

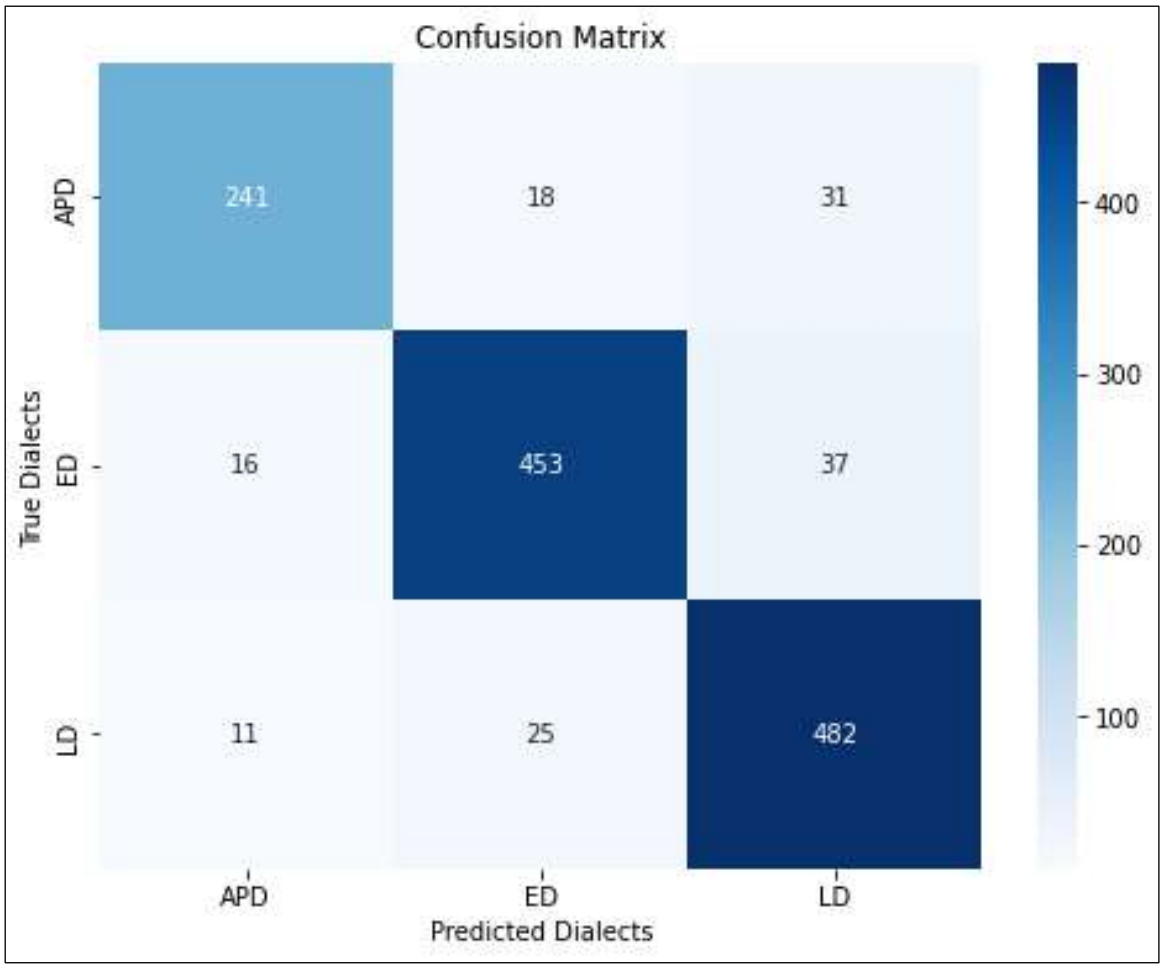

Figure 5.1: Confusion Matrix for the MLP Model with the MSARA 7-Second Dataset.

5.2 CNN MODEL

The proposed CNN model has been tested for the three dataset groups of OSARA, FSARA, and MSARA. Moreover, all five lengths of samples: 3, 4, 5, 6, and 7 seconds have been taken into consideration. Table 5.2 displays the testing performances of the CNN model for all the cases.

Table 5.2: Testing Performances of the Proposed CNN Model for the Three Dataset Groups (OSARA, FSARA and MSARA) and all the Lengths of Samples (3, 4, 5, 6 and 7)

Sample Length		3 seconds		4 seconds		5 seconds		6 seconds		7 seconds	
Accuracy & Loss		Accuracy (%)	Loss	Accuracy (%)	Loss	Accuracy (%)	Loss	Accuracy (%)	Loss	Accuracy (%)	Loss
Dataset Group	OSARA	62.6	1.11	68.1	1.09	71.2	0.99	68.1	1.15	72.6	0.74
	FSARA	59.1	1.17	62.3	1.19	65.2	1.22	68.4	1.09	72.9	0.75
	MSARA	70.1	0.92	73.6	0.81	79.8	0.71	81.2	0.69	87.3	0.43

From Table 5.2, it can be realized that the OSARA group of the used dataset has fluctuated differences among almost all the performances of the five lengths of samples. This means that the range of differences for the accuracy is [62.6%, 72.6%] and the range of variances for the loss is [0.74, 1.15]. Ranges are expansion for the FSARA group where the accuracy range is reported [59.1%, 72.9%] and the loss range is found [0.75, 1.22]. Once again, the best performances are registered for the MSARA dataset group in all five lengths of samples. Values have increased dramatically for the variances among the accuracy results achieved, the accuracy range values have been increased to [70.1%, 87.3%]. While low decreasing values have been obtained for the differences of all the loss results, the range values have been decreased to [0.43, 0.92]. These results are consistent with the achievements of the MLP model. Therefore, the same interpretations can be considered here. This means that the performance presented is because of the specifications of each dataset group, where the OSARA has the original samples that cannot be analyzed well, the FSARA has filtered samples that may demonstrate big variances in assessments, and the MSARA involves fusion between the OSARA and FSARA where this makes it includes more samples with specifications to analyzed efficiently.

Once again, the performance of the 7-second sample length for each dataset group is better than the performance of the 3-second sample length for each dataset group. To illustrate, the accuracy of the 7-second sample length for the MSARA dataset group is 87.3% and the loss is 0.43, while the accuracy of the 3-second sample length for the same dataset group is 70.1% and the loss is 0.92. This can also be explained by the length of samples is becomes more valuable if it is high as it consists big set of information that can be effective in getting desired results. Therefore, it can also be discovered that the longer the length of samples, the better the results, although the number of samples is smaller. Figure 5.2 shows the confusion matrix for the CNN model with the MSARA 7-second dataset that represents the best case in this model.

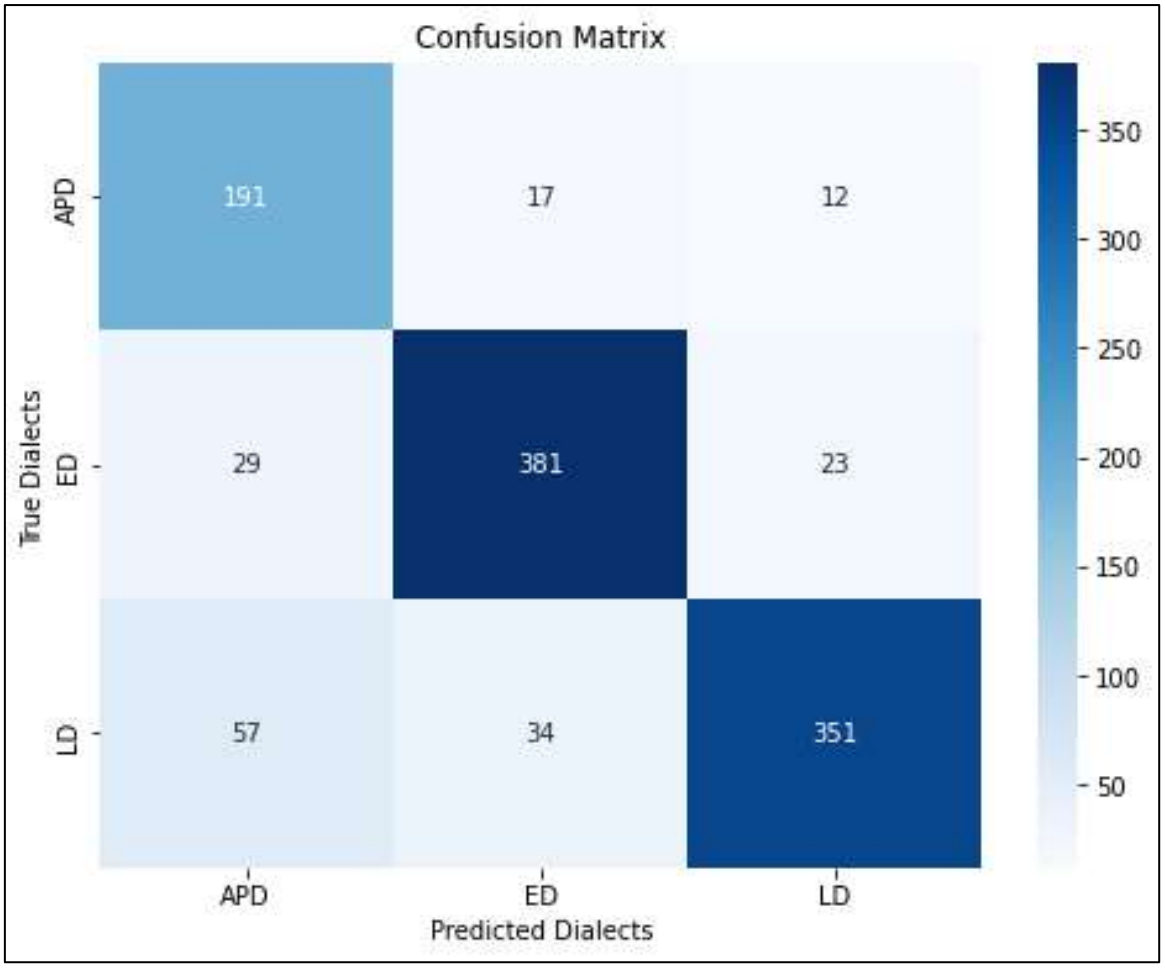

Figure 5.2: Confusion Matrix for the CNN Model with the MSARA 7-Second Dataset.

5.3 DRNN MODEL

The proposed DRNN model has been tested for the three dataset groups of OSARA, FSARA, and MSARA. Moreover, all five lengths of samples: 3, 4, 5, 6, and 7 seconds have been taken into consideration. Table 5.3 displays the testing performances of the DRNN model for all the cases.

Table 5.3: Testing Performances of the Proposed DRNN Model for the Three Dataset Groups (OSARA, FSARA and MSARA) and all the Lengths of Samples (3, 4, 5, 6 and 7).

Sample Length		3 seconds		4 seconds		5 seconds		6 seconds		7 seconds	
Accuracy & Loss		Accuracy (%)	Loss	Accuracy (%)	Loss	Accuracy (%)	Loss	Accuracy (%)	Loss	Accuracy (%)	Loss
Dataset Group	OSARA	70.3	1.71	77.4	1.34	79.9	1.05	81.1	0.78	82.9	0.95
	FSARA	65.9	1.98	74.3	1.37	76.3	1.25	75.7	1.27	79.9	0.87
	MSARA	81.6	0.95	85.9	0.68	87.1	0.59	88.7	0.60	90.7	0.46

From Table 5.3, it can be discovered that the OSARA group of the used dataset has made considerable differences amongst almost all the performances of the five lengths of samples. This means that variances of the accuracies have increased dramatically between the range of [70.3%, and 82.9%] and variances of the loss have fluctuated significantly between the range of [1.36, 2.35]. These ranges and differences have also been expanded for the FSARA group, the accuracy range is attained [65.9%, 79.9%] and the loss range is become [0.87, 1.98]. Once again, the best performances can also be noticed for the MSARA dataset group in all five lengths of samples. Significant values for the results have been benchmarked as the accuracy range values have been raised dramatically between the range of [81.6%, and 90.7%], and the loss range values have fluctuated between the range of [0.46, 0.95]. These outcomes have been also consistent with accomplishments in the MLP and CNN models. Once again, the same illustrations can be made here. To illustrate, the specifications of each dataset group are determined by the obtained performances, where the OSARA has the raw data that cannot be considered well, the FSARA has filtered data that may give big differences in testing, and the MSARA has composited information between the OSARA and FSARA, it includes more data with specifications to consider effectively.

Moreover, the performances of the 7-second sample length for all dataset groups are also recorded as better outcomes than the performances of the 3-second sample length for all dataset groups. For instance, the accuracy of the 7 seconds sample length for the MSARA dataset group is 90.7% and the loss is 0.46, while the accuracy of the 3 seconds sample length for the same dataset group (MSARA) is 81.6% and the loss is 0.95. This is also because of the length of samples that becomes more affected if it is high as it involves a big set of information which can be significant in yielding and analyzing valuable results. Therefore, it can be confirmed that the bigger the length of samples, the better the performances, although the number of samples is smaller.

Several factors, including architecture and training duration, have been taken into consideration to support the selection of a successful model. In terms of architecture, all three models are made up of unique and practical architecture. As an example, the CNN employs a similar architecture specifically made for image inputs but with a large number of hidden layers, the MLP uses a feedforward architecture that takes into account appropriate adaptation for its inputs and adopts a small number of hidden layers, and the DRNN takes advantage of a more efficient architecture that incorporates useful feedback added to multiple hidden layers. Additional calculations are made for the MLP, CNN, and DRNN models concerning the training time. According to their results, the MLP takes 605.7 seconds, the CNN 404.5 seconds, and the DRNN 485.8 seconds. These findings indicate that training phases take a long period, particularly for CNN and DRNN. Since simple analyses only take a very short amount of time, it makes sense that testing takes even less time than training. Therefore, the best results from the three suggested machine learning models are benchmarked for the DRNN, and accuracy is employed here as the most acceptable rationale for selecting the best model. In particular, the DRNN with the MSARA dataset group and 7-second sample duration performs the best, with the maximum accuracy of 90.7% and the lowest loss of 0.46 reported. Figure 5.3 shows the confusion matrix for the DRNN model with the MSARA 7-second dataset that represents the best case among all the models.

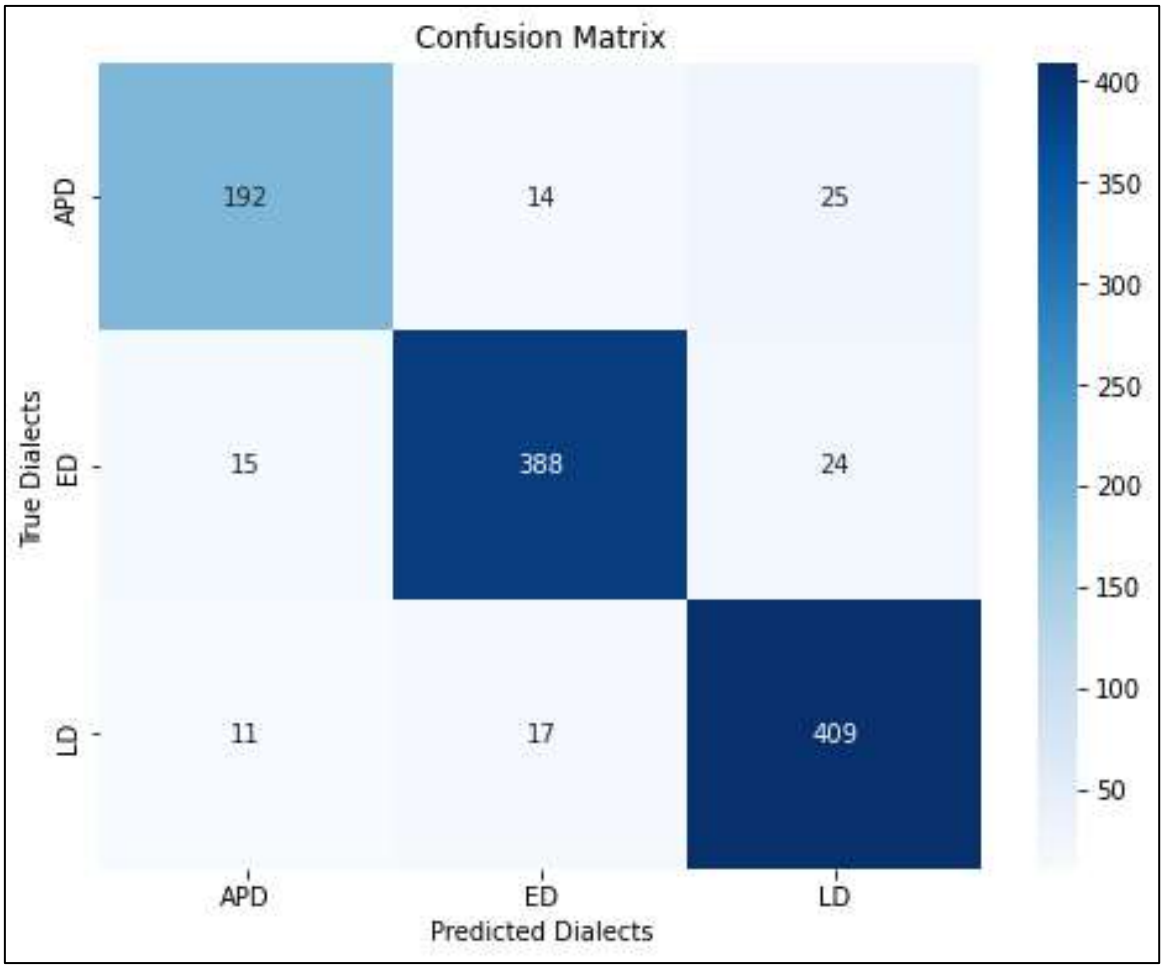

Figure 5.3: Confusion Matrix for the CNN Model with the MSARA 7-Second Dataset.

5.4 CONCLUSION

A technique for differentiating between the three primary Arabic dialects APD, ED, and LD was created in this work. The first step in the system's development was gathering the SARA dataset. The SARA was then prepared and divided into three groups (OSARA, FSARA, and MSARA) during the preparation stage. The MFCC was taken into consideration in the feature extraction step that followed. Following that, three models of artificial neural networks MLP, CNN, and DRNN were put out. The three dataset groupings, various sample durations, and various evaluations were all considered during the lengthy experiments used to train and evaluate the three models.

In conclusion, it has been discovered that the suggested DRNN model with the 7-second sample duration and MSARA group achieves the maximum accuracy of 90.70% and the lowest loss of 0.46. The ability of the MSARA group to achieve the best testing results for

each of the three suggested models has been verified. Additionally, it might be examined whether, even with fewer samples, better results are obtained with longer sample lengths. This study could be used in future research to identify speakers' nationalities by analyzing a portion of their everyday speech. in addition to obtaining a database containing the languages of every Arab nation.

REFERENCES

[1] R. J. Baken, "Clinical measurement of speech and voice," *(No Title)*, 1987.

[2] P. Ladefoged and K. Johnson, "A course in phonetics, 7th edn. Cengage Learning," *Inc., Boston*, 2014.

[3] P. Boersma, "Praat, a system for doing phonetics by computer," *Glot. Int.*, vol. 5, no. 9, pp. 341–345, 2001.

[4] K. C. Ryding, *A reference grammar of Modern Standard Arabic*. Cambridge University Press, 2005.

[5] J. Owens, "Arabic sociolinguistics," *Arabica*, vol. 48, no. Fasc. 4, pp. 419–469, 2001.

[6] K. Versteegh, *Arabic language*. Edinburgh University Press, 2014.

[7] C. A. Ferguson, "Diglossia." Word, 1959.

[8] C. Holes, *Modern Arabic: Structures, functions, and varieties*. Georgetown University Press, 2004.

[9] Y. Suleiman, *A war of words: Language and conflict in the Middle East*, vol. 19. Cambridge University Press, 2004.

[10] J. Owens, *A linguistic history of Arabic*. Oxford University Press, 2006.

[11] P. Behnstedt and M. Woidich, *Arabische Dialektgeographie: Eine Einführung*, vol. 78. Brill, 2005.

[12] N. Haeri, *Sociolinguistic Market of Cairo*. Routledge, 2021.

[13] A. G. Chejne, *The Arabic language: Its role in history*. U of Minnesota Press, 1968.

[14] R. Bassiouney, "A new direction for Arabic sociolinguistics," in *Perspectives on Arabic Linguistics XXIX: Papers from the Annual Symposium on Arabic*, John Benjamins Amsterdam and Philadelphia, 2017, pp. 7–30.

[15] N. Y. Habash, *Introduction to Arabic natural language processing*. Morgan & Claypool Publishers, 2010.

[16] K. Shaalan, S. Siddiqui, M. Alkhatib, and A. Abdel Monem, “Challenges in Arabic natural language processing,” in *Computational linguistics, speech and image processing for Arabic language*, World Scientific, 2019, pp. 59–83.

[17] O. F. Zaidan and C. Callison-Burch, “Arabic dialect identification,” *Computational Linguistics*, vol. 40, no. 1, pp. 171–202, 2014.

[18] W. Salloum and N. Habash, “Dialectal to standard Arabic paraphrasing to improve Arabic-English statistical machine translation,” in *Proceedings of the first workshop on algorithms and resources for modelling of dialects and language varieties*, 2011, pp. 10–21.

[19] H. Bouamor, N. Habash, and K. Oflazer, “A Multidialectal Parallel Corpus of Arabic.,” in *LREC*, 2014, pp. 1240–1245.

[20] A. Albirini, “The acquisition of Arabic as a first language,” in *The Routledge handbook of Arabic linguistics*, Routledge, 2017, pp. 227–248.

[21] H. S. Al-Khalifa, “Introduction to the special issue on Arabic NLP: Current state and future challenges,” *Journal of King Saud University-Computer and Information Sciences*, vol. 26, no. 4. Elsevier, pp. 355–356, 2014.

[22] R. R. Ziedan, M. N. Micheal, A. K. Alsammak, M. F. M. Mursi, and A. S. Elmaghraby, “A unified approach for Arabic language dialect detection,” *29th International Conference on Computer Applications in Industry and Engineering, CAINE 2016*, no. March 2017, pp. 165–170, 2016.

[23] M. W. Cowell, “A REFERENCE GRAMMAR OF SYRIAN ARABIC (BASED ON THE DIALECT OF DAMASCUS). ARABIC SERIES, NUMBER SEVEN.,” 1964.

[24] Y. Abou Taha, “Contact-induced change in the Levantine: Evidence from Lebanese and Palestinian Arabic.” Université d’Ottawa/University of Ottawa, 2022.

[25] C. Holes, *Arabic historical dialectology: Linguistic and sociolinguistic approaches*, vol. 30. Oxford University Press, 2018.

[26] B. Ingham, *North East Arabian Dialects: Monograph 3*. Routledge, 2013.

[27] O. Hafez, “Phonological and morphological integration of loanwords into Egyptian Arabic,” *Égypte/Monde Arabe*, no. 27–28, pp. 383–410, 1996.

[28] C. Miller, “Arabic urban vernaculars: Development and change,” in *Arabic in the City*, Routledge, 2007, pp. 15–46.

[29] C. Miller, E. Al-Wer, D. Caubet, and J. C. E. Watson, *Arabic in the city: Issues in dialect contact and language variation*. Routledge, 2007.

[30] A. Hachimi, “Shifting sands: Language and gender in Moroccan Arabic,” *Gender across languages: the linguistic representation of women and men*, vol. 1, pp. 27–52, 2001.

[31] T. F. Milchell, “Pronouncing Arabic, Vol. I Clarendon Press. Oxford 1990 XII+ 167 pp. ISBN 0-19-815151-9”.

[32] W. Diem, “Some glimpses at the rise and early development of the Arabic orthography,” *Orientalia*, vol. 45, pp. 251–261, 1976.

[33] M. Younes, “Emphasis spread in two Arabic dialects,” *AMSTERDAM STUDIES IN THE THEORY AND HISTORY OF LINGUISTIC SCIENCE SERIES 4*, p. 119, 1993.

[34] M. Najafian, S. Khurana, S. Shan, A. Ali, and J. Glass, “Exploiting Convolutional Neural Networks for Phonotactic Based Dialect Identification,” *ICASSP, IEEE International Conference on Acoustics, Speech and Signal Processing - Proceedings*, vol. 2018-April, pp. 5174–5178, 2018, doi: 10.1109/ICASSP.2018.8461486.

[35] S. Shon, A. Ali, and J. Glass, “Convolutional Neural Networks and Language Embeddings for End-to-End Dialect Recognition,” *Speaker and Language Recognition Workshop, ODYSSEY 2018*, no. March, pp. 98–104, 2018, doi: 10.21437/Odyssey.2018-14.

[36] R. Al-Sabbagh and R. Girju, “Mining the Web for the Induction of a Dialectical Arabic Lexicon.,” in *LREC*, 2010.

[37] W. Salloum and N. Habash, “Dialectal Arabic to English machine translation: Pivoting through modern standard Arabic,” in *Proceedings of the 2013 Conference of*

the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, 2013, pp. 348–358.

[38] A. Abdelali, H. Mubarak, Y. Samih, S. Hassan, and K. Darwish, "Arabic dialect identification in the wild," *arXiv preprint arXiv:2005.06557*, 2020.

[39] A. Ali *et al.*, "Automatic dialect detection in arabic broadcast speech," *arXiv preprint arXiv:1509.06928*, 2015.

[40] A. Hanani and R. Naser, "Spoken Arabic dialect recognition using X-vectors," *Natural Language Engineering*, vol. 26, no. 6, pp. 691–700, 2020, doi: 10.1017/S1351324920000091.

[41] S. Aich, S. Chakraborty, and H.-C. Kim, "Convolutional neural network-based model for web-based text classification.," *International Journal of Electrical & Computer Engineering (2088-8708)*, vol. 9, no. 6, 2019.

[42] X. Zhang, J. Zhao, and Y. LeCun, "Character-level convolutional networks for text classification," *Advances in neural information processing systems*, vol. 28, 2015.

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

[44] K. Cho, B. Van Merriënboer, D. Bahdanau, and Y. Bengio, "On the properties of neural machine translation: Encoder-decoder approaches," *arXiv preprint arXiv:1409.1259*, 2014.

[45] A. Vaswani, "Attention is all you need," *arXiv preprint arXiv:1706.03762*, 2017.

[46] J. Lee and K. Toutanova, "Pre-training of deep bidirectional transformers for language understanding," *arXiv preprint arXiv:1810.04805*, vol. 3, no. 8, 2018.

[47] H. A. Alsayadi, S. Al-Hagree, F. A. Alqasemi, and A. A. Abdelhamid, "Dialectal Arabic Speech Recognition using CNN-LSTM Based on End-to-End Deep Learning," *2022 2nd International Conference on Emerging Smart Technologies and Applications, eSmarTA 2022*, 2022, doi: 10.1109/eSmarTA56775.2022.9935427.

[48] S. Nasr, R. Duwairi, and M. Quwaider, "End-to-End Speech Recognition For Arabic

Dialects," *Arabian Journal for Science and Engineering*, vol. 48, no. 8, pp. 10617–10633, 2023, doi: 10.1007/s13369-023-07670-7.

[49] M. El-Haj, P. Rayson, and M. Aboelezz, "Arabic dialect identification in the context of bivalency and code-switching," in *Proceedings of the 11th International Conference on Language Resources and Evaluation, Miyazaki, Japan.*, European Language Resources Association, 2018, pp. 3622–3627.

[50] S. Shon, A. Ali, and J. Glass, "MIT-QCRI Arabic dialect identification system for the 2017 multi-genre broadcast challenge," in *2017 ieee automatic speech recognition and understanding workshop (asru)*, IEEE, 2017, pp. 374–380.

[51] O. Zaidan and C. Callison-Burch, "The arabic online commentary dataset: an annotated dataset of informal arabic with high dialectal content," in *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, 2011, pp. 37–41.

[52] H. Elfardy and M. Diab, "Sentence level dialect identification in Arabic," in *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, 2013, pp. 456–461.

[53] N. Habash and F. Sadat, "Arabic preprocessing schemes for statistical machine translation," in *Proceedings of the human language technology conference of the naacl, companion volume: Short papers*, 2006, pp. 49–52.

[54] M. Elmahdy, M. Hasegawa-Johnson, and E. Mustafawi, "A transfer learning approach for under-resourced Arabic dialects speech recognition," in *Workshop on Less Resourced Languages, new technologies, new challenges and opportunities (LTC 2013)*, Citeseer, 2013, pp. 60–64.

[55] A. R. Ali, "Multi-dialect Arabic speech recognition," in *2020 International Joint Conference on Neural Networks (IJCNN)*, IEEE, 2020, pp. 1–7.

[56] M. Al-Badrashiny, R. Eskander, N. Habash, and O. Rambow, "Automatic transliteration of romanized dialectal Arabic," in *Proceedings of the eighteenth conference on computational natural language learning*, 2014, pp. 30–38.

[57] A. Poliak, Y. Belinkov, J. Glass, and B. Van Durme, “On the evaluation of semantic phenomena in neural machine translation using natural language inference,” *arXiv preprint arXiv:1804.09779*, 2018.

[58] A. Shoukry and A. Rafea, “SATALex: Telecom Domain-specific Sentiment Lexicons for Egyptian and Gulf Arabic Dialects.,” in *WEBIST*, 2019, pp. 169–176.

[59] S. Rosenthal, N. Farra, and P. Nakov, “SemEval-2017 task 4: Sentiment analysis in Twitter,” *arXiv preprint arXiv:1912.00741*, 2019.

[60] N. Habash and O. Rambow, “MAGEAD: A morphological analyzer and generator for the Arabic dialects,” in *Proceedings of the 21st International Conference on Computational Linguistics and 44th Annual Meeting of the Association for Computational Linguistics*, 2006, pp. 681–688.

[61] R. Cotterell and K. Duh, “Low-resource named entity recognition with cross-lingual, character-level neural conditional random fields,” *arXiv preprint arXiv:2404.09383*, 2024.

[62] C. Miller, A. Barontini, M.-A. Germanos, J. Guerrero, and C. Pereira, “Studies on Arabic Dialectology and Sociolinguistics,” 2019.

[63] D. S. Williamson, Y. Wang, and D. Wang, “Complex ratio masking for monaural speech separation,” *IEEE/ACM transactions on audio, speech, and language processing*, vol. 24, no. 3, pp. 483–492, 2015.

[64] D. Yu, M. Kolbæk, Z.-H. Tan, and J. Jensen, “Permutation invariant training of deep models for speaker-independent multi-talker speech separation,” in *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2017, pp. 241–245. doi: 10.1109/ICASSP.2017.7952154.

[65] W. Mack and E. A. P. Habets, “Deep filtering: Signal extraction and reconstruction using complex time-frequency filters,” *IEEE Signal Processing Letters*, vol. 27, pp. 61–65, 2019.

[66] H. Schröter, T. Rosenkranz, A. N. Escalante-B, M. Aubreville, and A. Maier, “CLCNET: Deep Learning-Based Noise Reduction for Hearing aids using Complex

Linear Coding," in *ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2020, pp. 6949–6953. doi: 10.1109/ICASSP40776.2020.9053563.

[67] H. Schroter, A. N. Escalante-B, T. Rosenkranz, and A. Maier, "Deepfilternet: A low complexity speech enhancement framework for full-band audio based on deep filtering," in *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, IEEE, 2022, pp. 7407–7411.

[68] H. Schröter, A. Maier, A. N. Escalante-B, and T. Rosenkranz, "Deepfilternet2: Towards real-time speech enhancement on embedded devices for full-band audio," in *2022 International Workshop on Acoustic Signal Enhancement (IWAENC)*, IEEE, 2022, pp. 1–5.

[69] M. H. Bahari, N. Dehak, H. Van Hamme, L. Burget, A. M. Ali, and J. Glass, "Non-negative factor analysis of Gaussian mixture model weight adaptation for language and dialect recognition," *IEEE Transactions on Audio, Speech and Language Processing*, vol. 22, no. 7, pp. 1117–1129, 2014, doi: 10.1109/TASLP.2014.2319159.

[70] M. A. Hossan, S. Memon, and M. A. Gregory, "A novel approach for MFCC feature extraction," in *2010 4Th international conference on signal processing and communication systems*, IEEE, 2010, pp. 1–5.

[71] M. T. S. Al-Kaltakchi, R. R. O. Al-Nima, and M. A. M. Abdullah, "Comparisons of extreme learning machine and backpropagation-based i-vector approach for speaker identification," *Turkish Journal of Electrical Engineering and Computer Sciences*, vol. 28, no. 3, pp. 1236–1245, 2020.

[72] A. S. Haq, M. Nasrun, C. Setianingsih, and M. A. Murti, "Speech recognition implementation using MFCC and DTW algorithm for home automation," *Proceeding of the electrical engineering computer science and informatics*, vol. 7, no. 2, pp. 78–85, 2020.

[73] R. Ranjan and A. Thakur, "Analysis of feature extraction techniques for speech recognition system," *International Journal of Innovative Technology and Exploring Engineering*, vol. 8, no. 7C2, pp. 197–200, 2019.

[74] M. Limkar, R. Rao, and V. Sagvekar, “Isolated digit recognition using MFCC and DTW,” *International Journal on Advanced Electrical and Electronics Engineering,(IJAEEE)*, vol. 1, no. 1, 2012.

[75] M. E. Rahaman, S. M. S. Alam, H. S. Mondal, A. S. Muntaseer, R. Mandal, and M. Raihan, “Performance Analysis of Isolated Speech Recognition Technique Using MFCC and Cross-Correlation,” *2019 10th International Conference on Computing, Communication and Networking Technologies, ICCCNT 2019*, pp. 1–4, 2019, doi: 10.1109/ICCCNT45670.2019.8944534.

[76] N. Chauhan, T. Isshiki, and D. Li, “Speaker recognition using LPC, MFCC, ZCR features with ANN and SVM classifier for large input database,” in *2019 IEEE 4th international conference on computer and communication systems (ICCCS)*, IEEE, 2019, pp. 130–133.