OPTIMISING SWAHILI DIGRAPH RECOGNITION USING A FEATURE VECTOR REDUCTION MODEL

Tirus Muya Maina

A Thesis Submitted in Partial Fulfilment of the Requirements for the Degree of Doctor of Philosophy in Computer Science of Murang'a University of Technology

August, 2025

DECLARATION

Student's Declaration

I hereby declare that this Thesis is my original work and to the best of my knowledge has not been presented for a degree award in this or any other university. I understand what plagiarism is, and I am aware of the University's policy in this regard. I understand that any false claim in respect of this work shall result in disciplinary action in accordance with the University Anti-Plagiarism Policy.

_____ _____

Tirus Muya Maina Date

Reg No.: SC501/5213/2019

Supervisors' Approval

The undersigned certify that they have read and hereby recommend for acceptance of Murang'a University of Technology a thesis entitled **"Optimising Swahili Digraph Recognition Using a Feature Vector Reduction Model".**

_____ _____

Dr Aaron Mogeni Oirere, Date

Department of Computer Science,

Murang'a University of Technology

_____ _____

Dr. Stephen Kahara Wanjau, Date

Department of Computer Science,

Murang'a University of Technology

DEDICATION

This thesis is dedicated to my lovely wife, Pauline Nyambura Mwangi, my children, Genevieve Waigumo Muya, David Rufas Maina Muya and Charlotte Wanja Muya; my parents, the Late Dr. David M. Kahando and Salome Waigumo; my siblings, Charity Mweru, Nancy Wagi and Douglas Mari. I am sincerely grateful for their steadfast support, constant prayers, and words of encouragement throughout my research journey. Their presence and belief in me have been instrumental to the success of this academic endeavour.

ACKNOWLEDGEMENT

With utmost humility, I first thank the Almighty God for His enduring grace, granting me good health, longevity, financial provision, and the privilege to pursue advanced studies. I owe immense gratitude to my supervisors, Dr Aaron Mogeni Oirere, PhD and Dr. Stephen Kahara Wanjau, PhD, whose dedicated mentorship and expert guidance at every stage of this research journey were invaluable. Their deep understanding of the research domain profoundly influenced the trajectory and refinement of my work.

I also wish to convey my sincere appreciation to Prof. Geoffrey Muchiri Muketha, Professor of the School of Computing and Information Technology; Dr. John Ndia, Dean of the School of Computing and Information Technology; Dr Aaron Mogeni Oirere, Chairman of the Computer Science Department. My heartfelt thanks extend to all faculty and staff members of the School of Computing and Information Technology at Murang'a University of Technology for their consistent encouragement, insightful feedback, and unwavering support throughout this endeavour.

I am deeply thankful to my wife Pauline Nyambura, my children, Genevieve Waigumo, David Rufas Maina and Charlotte Muya; my parents, the Late Dr. David M. Kahando and Salome Waigumo, for their steadfast love, prayers, and continuous emotional support that kept me grounded and motivated along this path. May you all my family be abundantly blessed.

Finally, I express sincere gratitude to my friends and research collaborators, Dr Peter Mwangi, and Mr. Jackson Kamiri, along with all those named or unnamed whose encouragement and assistance were instrumental in the success of this journey. Your support is deeply appreciated and will always be remembered.

ABSTRACT

Recognising and classifying digraph letter pairs representing single phonemes remains a key linguistic and computational challenge in low-resource languages. Digraphs are essential for accurate pronunciation and comprehension in Natural Language Processing, speech and text recognition, transcription, and machine translation. Swahili, a widely spoken East African lingua franca, is a low-resource language with limited linguistic datasets and computational tools, featuring a rich phonetic and syllabic structure. Current models focus on high-resource languages and word-level processing, often neglecting the complexities of individual phonemes and digraphs. This study proposed an optimised model for Swahili digraph recognition by introducing a Feature Vector Reduction Model using Multinomial Logistic Regression, aimed at enhancing both computational efficiency and linguistic precision. The research pursued four key objectives: (1) phonetic and linguistic analysis of Swahili digraphs, (2) development of an annotated digraph corpus using a CNN-based model, (3) optimisation of digraph recognition using FRMLR, and (4) empirical validation of the FRMLR model. Anchored in positivist epistemology and employing the Design Science Research Methodology, the study integrates empirical validation, reproducibility, and measurable outcomes to advance NLP for low-resource languages. A significant contribution of this study is the creation of the Annotated Swahili Digraph Corpus, comprising 31,197 lexicon entries annotated with digraph-vowel interactions and phonotactic rules. This multimodal corpus built using CNN-based extraction, incorporates both text and synthetic speech data and achieved a recognition accuracy of 92.5% (MAE = 0.112, $R^2 = 0.89$), thus filling a critical gap in Swahili linguistic resources. The FRMLR framework effectively reduced 39-dimensional Mel-Frequency Cepstral Coefficients into compact 5-dimensional probability vectors, resulting in 70% memory savings (from 1500 MB to 450 MB) and a 77.8% decrease in training times (from 45 minutes to 10 minutes). It consistently outperformed baseline methods such as Principal Component Analysis with Support Vector Machines, achieving a precision of 91.8% and a recall of 90.3%, with minimal misclassifications. Statistical validation confirmed the effectiveness of FRMLR. A one-way ANOVA test ($p = 0.003$) showed a statistically significant improvement in FRMLR's performance compared to baseline models, Post hoc analysis using Tukey's HSD test further validated that FRMLR significantly outperformed Logistic Regression ($p = 0.002$), SVM ($p = 0.015$), and KNN ($p = 0.001$), while performing comparably to CNN ($p = 0.670$). The AUC-ROC analysis consistently produced values exceeding 0.91 across the training, validation, and test datasets. This reinforces the model's strong ability to distinguish between phonetic patterns. The low false positive rates (FPR: 0.07–0.10) further confirm FRMLR's precision in minimising misclassifications. This quality is crucial for applications that require high-confidence predictions, such as educational tools and speech recognition systems. This study enhances computational linguistics by integrating machine learning with natural language processing for ASR in low-resource contexts. The FRMLR model effectively reduces phonetic feature dimensions, and CNN aids in corpus construction and pattern recognition. These approaches establish scalable benchmarks for transcription accuracy and voice systems, promoting digital inclusivity while striking a balance between linguistic rigour and computational efficiency. Future research should extend the FRMLR model to dialectal variants, multimodal datasets, and other agglutinative and low-resource languages.

TABLE OF CONTENTS

DECLARATION ..... **ii**
DEDICATION ..... **iii**
ACKNOWLEDGEMENT ..... **iv**
ABSTRACT ..... **v**
LIST OF TABLES ..... **ix**
LIST OF FIGURES ..... **x**
CHAPTER ONE: INTRODUCTION **1**
1.1 Background of the Study 1
1.2 Problem Statement 3
1.3 Objectives of the Study 5
1.3.1 General Objective 5
1.3.2 Specific Objectives 5
1.4 Research Questions 5
1.5 Significance of the Study 6
1.6 Scope of the Study 7
1.7 Limitations of The Study 8
1.8 Contributions of the Thesis 9
1.9 Organisation of the Thesis 11
CHAPTER TWO: LITERATURE REVIEW **14**
2.1 Introduction 14
2.2 Phonetic and Linguistic Principles 18
2.3 Frameworks 21
2.3.1 Distinctive Feature Theory 21
2.3.2 Formal Language Theory (FLT) 23
2.3.3 Optimisation Theory 25
2.3.4 Information Theory 26
2.3.5 Statistical Learning Theory 29
2.3.6 Swahili's Lingua Franca Status 31
2.4 Machine Learning Paradigms 36
2.4.1 Supervised Learning 37
2.4.2 Unsupervised Learning 39
2.4.3 Semi-Supervised Learning 40
2.4.4 Reinforcement Learning (RL) 42
2.5 Feature Extraction in Language Processing 46
2.5.1 Linear prediction coding (LPC) 46
2.5.2 Mel-Frequency Cepstrum Coefficients 47

2.5.3 Perceptual Linear Prediction 50
2.5.4 Relative Spectral-Perceptual Linear Prediction (RASTA-PLP) 51
2.6 Feature Dimensionality Reduction Techniques 53
2.6.1 Principal Component Analysis 53
2.6.2 Linear Discriminant Analysis (LDA) 55
2.7 Optimisation Techniques 55
2.8 Multinomial Logistic Regression in Language Processing 57
2.9 Validation strategies 59
2.9.1 Empirical Validation 60
2.9.2 Comparison with existing Models 66
2.10 Conceptual Framework 67
2.11 Summary 69
CHAPTER THREE: METHODOLOGY 70
3.1 Introduction 70
3.2 Research Philosophy 70
3.3 Research Design 71
3.4 Research Process 75
3.5 Research Population 77
3.6 Data Sources and Collection 81
3.7 Data Analysis 85
3.8 Model Feature Utilisation Architecture for Swahili Corpus 88
3.9 Training and Evaluation 93
3.10 Ethical Considerations 94
3.11 Summary 95
CHAPTER FOUR: DEVELOPMENT OF SWAHILI DIGRAPH CORPUS ... 97
4.1 Introduction 97
4.2 Results 97
4.2.1 Acoustic and Phonetic Features of Swahili Digraphs 97
4.2.2 The Swahili Digraph Extraction Model 98
4.2.3 The Model Performance Evaluation 99
4.2.4 The Annotated Swahili Digraph Corpus 101
4.2.5 Synthetic Multimodal Digraph corpus 102
4.3 Discussion 104
4.4 Summary 107
CHAPTER FIVE: FEATURE REDUCTION USING MLR 108
5.1 Introduction 108
5.2 Comparative Evaluation of Feature Extraction Techniques 108
5.3 Preprocessing 109

5.3.1 Mel-Frequency Cepstral Coefficients 109
5.3.2 39-Dimensional Feature Vectors 111
5.4 Feature Reduction Using FRMLR 112
5.4.1 Generating the Probability Vector 117
5.4.2 Dimensionality Reduction Analysis 120
5.4.3 Confusion Matrix of FRMLR Model 122
5.5 Results on Model Performance 123
5.5.1 Model Performance 123
5.5.2 Comparison with Conventional Methods 124
5.6 Discussion 126
5.7 Summary 128
CHAPTER SIX: EMPIRICAL VALIDATION OF FRMLR MODEL 130
6.1 Introduction 130
6.2 Results 131
6.2.1 FRMLR Cross-validation and Hyperparameter Tuning 131
6.2.2 FRMLR Model Evaluation Metrics 134
6.2.3 FRMLR Hyperparameter Tuning 136
6.2.4 Final FRMLR Model Testing 137
6.2.5 FRMLR Performance on Benchmark Datasets 137
6.2.6 FRMLR vs. Baseline Models Comparative Analysis 139
6.2.7 AUC-ROC Empirical Validation of the FRMLR Model 142
6.2.8 Statistical Significance 145
6.3 Discussion 148
6.4 Summary 150
CHAPTER SEVEN: SUMMARY, CONCLUSION AND RECOMMENDATIONS 151
7.1 Summary 151
7.2 Conclusion 155
7.3 Recommendations 156
7.3.1 Policy Recommendations 156
7.3.2 Recommendations for Practice 157
7.4 Future Research 158
REFERENCES 159
APPENDICES 190

LIST OF TABLES

Table 2.1: ML Paradigms and Their Applications 44
Table 2.2 Optimisation Techniques 57
Table 3.1: Strategies for Objective Implementation 74
Table 3.2 Swahili Digraph Frequency by Vowel 80
Table 3.2:Swahili Datasets 81
Table 4.1: Evaluated Performance Metrics 100
Table 4.2: The Annotated Swahili Digraph Corpus 101
Table 4.3: Synthetic vs. Human MFCC Metrics 104
Table 5.1: Comparative Evaluation of Feature Extraction Techniques 108
Table 5.2: Categorisation of Swahili Digraphs by Phonetic Properties 115
Table 5.3: Performance Metrics for Vowels and Model Evaluation 116
Table 5.4: Classification of New MFCC Vector with Predicted Vowel Class 117
Table 5.5: Analysis of Digraphs and Probability Vectors 118
Table 5.6: Confusion Matrix of FRMLR Model 122
Table 5.7: Comparison: FRMLR Model vs. 39-Dimensional MFCC 124
Table 5.8: Conventional Methods vs. FRMLR 125
Table 6.1:10-fold cross-validation 132
Table 6.2: FRMLR Model Evaluation Metrics 135
Table 6.3: Hyperparameter Tuning Summary 136
Table 6.4: Final FRMLR Model Testing 137
Table 6.5: FRMLR Model Performance on Benchmark Datasets 138
Table 6.6: FRMLR vs. Baseline Models Comparative Analysis 139
Table 6.7: AUC-ROC Empirical Validation of the FRMLR Model 143
Table 6.8 Performance Metrics Analysis 146
Table 6.9 ANOVA Results 147
Table 6.10 Post Hoc Analysis (Tukey's HSD Test) 147

LIST OF FIGURES

Figure 2.1 Machine learning categories and algorithms 44
Figure 2.2 MFCC feature extraction process 49
Figure 2.3 The process of PLP 51
Figure 2.4: RASTA-PLP Feature Extraction Process 53
Figure 3.1: Model Architecture 91
Figure 4.1: CNN-based digraph extraction model 99
Figure 5.1: Dimensionality reduction and computational complexity 121
Figure 6.1: 10-fold cross-validation accuracy of FRMLR model 133
Figure 6.2: Heatmap: FRMLR vs. Baseline Model Performance 141
Figure 6.3: Empirical Validation of AUC-ROC Curve for FRMLR Model 144

CHAPTER ONE

INTRODUCTION

1.1 Background of the Study

Speech recognition software is employed in automatic speech recognition (ASR) systems to convert spoken words into text [1]. As stated by [2] Technology plays a pivotal role in facilitating and enhancing interactions among humans as well as between humans and computers. This entails grouping spoken words into distinct digraph, consonant, and vowel sound categories. Speech-enabled applications that depend on correctly identifying phonetic units of speech need ASR [3] [4].

However, speech recognition encounters difficulties due to the unpredictable attributes of speech signals and the extensive array of phonetic units. Therefore, the accuracy with which ASR systems can discern phonemes holds paramount importance for applications relying on speech, such as speech recognition and synthesis [5] [2].

Every language incorporates consonants, vowels, and digraphs; however, the collection of distinct phonemes is unique to each language. Phonemes are the voiced components of any language's sound and are an important part of speech that is typically recognised with ease; as a result, they serve as important cues that improve speech recognition performance for both humans and machines. As a result, improved performance of ASR depends on the ability to recognise vowel sounds, consonants, and digraphs [6].

Swahili, a prominent language in East Africa, is notable for its phonetic and syllabic structure. It utilises the international Latin alphabet, with letters specifically chosen to represent Swahili phonemes. Swahili incorporates digraphs, pairs of letters representing individual sounds. As [7]outlines, the Swahili alphabet comprises five vowels: *a, e, i, o, u*, and nineteen consonants: *b, c, d, f, g, h, j, k, l, m, n, p, r, s, t, v, w, y, z* (excluding *x* and *q*). Furthermore, it includes nine digraphs: *ch, dh, gh, kh, ng', ny, sh, th, ng.*

Swahili poses special linguistic difficulties, particularly when it comes to digraph identification and recognition. The adoption of ASR has been hindered by

performance issues, specifically related to low recognition accuracy [8]. This is particularly true for low-resource languages (LRLs) like Swahili [4].

The terminology "high resource languages" and "low resource languages" originates from the recognition, as observed by [9] , Natural Language Processing (NLP) systems and resources pose significant obstacles in the field of NLP presently. However, there is a notable scarcity of systems or resources catering to these languages [10] [11]. Consequently, the language community faces the challenge of developing tools and resources, as emphasised in [12] [13].

High-resource languages are those that benefit from abundant digital resources, such as large text corpora, annotated datasets, pre-trained models (like BERT and GPT), and robust NLP tools, including tokenisers and parsers. These languages typically enjoy widespread usage, significant economic influence, or strong institutional support. Some prominent examples of high-resource languages include English, Mandarin, Spanish, and French. The key features of these languages include extensive datasets that facilitate the training of AI models, mature natural language processing (NLP) pipelines that support various applications such as machine translation and speech recognition, and substantial investments from both commercial and academic sectors [9] [12] [13].

High-resource languages like English and Mandarin benefit from abundant datasets and support, while Swahili, despite a large speaker base, is considered low-resource in technology. It struggles with data availability and tooling, with scarce annotated datasets for tasks like named entity recognition. Swahili Wikipedia has only about 60,000 articles compared to English's 6.7 million, and basic natural language processing tools are often underdeveloped. Machine translation systems like Google Translate yield unreliable results. Economic and institutional barriers further this issue, as Swahili lacks the global economic influence of languages like Mandarin, resulting in underinvestment. While not endangered, the limited digital resources for Swahili risk marginalising its speakers and limiting access to AI-driven services and educational tools [9] [12] [13].

Studies [14] [15] have concentrated on enhancing ASR performance, which can be done at the recognition or feature extraction (FE) levels. The present study aimed to concentrate on the feature extraction (FE) level due to its significant influence on recognition accuracy.

Researchers [14] and [15] have proposed several methods to enhance the performance of Automatic Speech Recognition (ASR) systems, including feature transformation, the development of multiple accent models, and pronunciation modelling alongside acoustic model adaptation. The inherent characteristics of all languages, combined with the dependence of ASR performance on speech databases, have posed significant challenges to the cross-language portability of ASR systems. Given that Swahili is a syllable-based language and that language acquisition begins with the learning of phoneme pronunciation, this study focused on phoneme recognition, particularly emphasising Swahili digraphs.

A specialised approach to solving probabilistic classification problems is Multinomial Logistic Regression (MLR), which estimates the probability of each value of the dependent variable using a linear combination of the observed characteristics and a few problem-specific factors [14] [15]. The Multiple Logistic Regression model forecasts event outcomes based on designated training datasets by gauging the likelihood of their incidence. It typically computes MLR output by dividing the likelihood of success by the probability of failure. These probability ratios are derived from regression analyses. MLR primarily serves as a classifier for identifying vowels, consonants, and digraphs, as evidenced in studies conducted by [16] [17].

1.2 Problem Statement

Swahili's complex phonetic structure, including consonant-vowel interplay, tone, stress, and vowel harmony compounded by digraph usage, poses significant challenges for accurate word recognition. While existing methodologies can detect such elements, their computational inefficiency limits their suitability for real-time automatic speech recognition in low-resource settings [18] [19].

Swahili's agglutinative structure and tonal variations result in complex word formations and pitch-dependent meanings, posing significant challenges for

morphological analysis and part-of-speech tagging in computational linguistics [6] [20] [21]. Swahili digraphs are inherently complex, and while conventional models can recognise them manually, achieving both accuracy and computational efficiency remains difficult, limiting their viability for real-time language processing [22] [23] [24] [25].

Inefficient digraph recognition hampers translation accuracy and speech command interpretation in Swahili NLP systems, despite ongoing advances in computational linguistics. Existing models struggle with the language's phonetic intricacies as nasalised digraphs, vowel harmony, and agglutinative structures, resulting in poor scalability and latency for real-time applications. Addressing this requires a tailored, optimised recognition model that balances linguistic precision with computational speed [26] [27] [28].

Despite major advances in ASR for high-resource languages like English, where error rates dropped from 25% to below 5% due to transformer-based and self-supervised models, Swahili ASR still suffers from error rates above 20–30%, driven by limited phonetic corpora, inadequate acoustic modelling, and insufficient integration of language-specific features like digraphs and vowel harmony [22] [23] [24] [25].

Studies have shown that linguistic complexity varies across languages, contributing to elevated ASR errors for low-resource systems [29] [30] [31].However, most literature prioritises vowel and consonant recognition, often neglecting the role of digraphs despite their phonemic significance and impact on ASR performance [32] [33] .This oversight undermines key NLP tasks such as transcription, translation, and speech-to-text generation. In morphologically rich languages like Swahili, where meaning is encoded through agglutinative and phonological rules, the lack of phonology-aware and morphologically sensitive models continues to limit ASR scalability and linguistic precision.

1.3 Objectives of the Study

1.3.1 General Objective

The main objective of this study is to develop a feature vector reduction model that optimises Swahili digraph recognition by utilising Multinomial Logistic Regression while maintaining computational efficiency and linguistic accuracy.

1.3.2 Specific Objectives

a) To analyse the Swahili digraph leveraging on phonetic and linguistic principles to identify and categorise digraphs.
b) To develop a CNN-based model to create an accurate and annotated Swahili Digraph Corpus for enhanced linguistic analysis and natural language processing.
c) To develop and optimise a Feature Reduction Model using Multinomial Logistic Regression for phonetic accuracy and computational efficiency in Swahili digraph recognition.
d) To empirically validate the Feature Reduction Model using Multinomial Logistic Regression (FRMLR) for Optimising Swahili digraph recognition.

1.4 Research Questions

a) How can phonetic and linguistic principles be leveraged to accurately identify and categorise Swahili digraphs?
b) How can a CNN-based digraph extraction model be developed to create an Annotated Swahili Digraph Corpus?
c) How can the Feature Reduction Model using Multinomial Logistic Regression (FRMLR) be optimised to improve phonetic accuracy in Swahili digraph recognition?
d) How can the Feature Reduction Model using Multinomial Logistic Regression (FRMLR) be empirically validated to ensure its generalisation capability across diverse datasets?

1.5 Significance of the Study

a) Linguistic Significance

Swahili was recognised as a major African language with a growing digital presence. Accurate digraph recognition was identified as a critical need for Natural Language Processing (NLP) applications to ensure seamless integration into digital technologies. Existing systems lacked the precision required to handle the unique phonetic and linguistic features of Swahili. This research addressed these shortcomings by contributing to the understanding and computational representation of Swahili phonology, creating a foundation for improved linguistic analysis and greater adoption in digital tools.

b) Technological Advancement

The study addressed key technical challenges by focusing on feature vector reduction in large linguistic datasets, which was essential for improving computational efficiency. The use of Multinomial Logistic Regression (MLR) provided a robust method for classification and dimensionality reduction. The proposed approach optimised processing speed and accuracy, enabling advancements in real-time Swahili NLP applications, such as speech recognition systems, text-to-speech synthesis, and transcription tools.

c) Corpus Development

The study created an annotated Swahili Digraph Corpus, filling a critical gap in available linguistic resources. This annotated corpus served as a long-term asset for Swahili NLP research. The research utilised Convolutional Neural Networks (CNNs), known for their ability to process images, to analyse written Swahili digraphs. This methodology enhanced the flexibility and applicability of the developed model.

d) Practical Applications

The proposed model enhanced the accuracy of Swahili digraph recognition, which had broad implications for practical applications. It supported speech recognition systems, text-to-speech synthesis, and optical character recognition (OCR). The improvements

also benefited language learning tools and contributed to digital inclusivity by improving the accessibility of digital resources for Swahili speakers.

1.6 Scope of the Study

The scope of this study is delimited to the phonetic recognition and classification of Swahili digraphs within the context of low-resource language speech processing. The research concentrates on the development, optimisation, and empirical validation of a Feature Reduction Model utilising Multinomial Logistic Regression (FRMLR), specifically designed to enhance Automatic Speech Recognition (ASR) for the Swahili language.

The study encompasses the development of an annotated Swahili Digraph Corpus employing a Convolutional Neural Network (CNN)-based digraph extraction framework. The model is configured to identify and learn phonetic features from authentic Swahili texts including Kamusi ya Kiswahili, the Swahili Bible, and the Constitution of Kenya, alongside a curated dataset of diverse digraph-rich sources from Mendeley, Harvard Dataverse, Zenodo, and Kaggle, ensuring linguistic depth, dialectal coverage, and scalability for Swahili NLP. The CNN architecture incorporates convolutional and fully connected layers, augmented with dropout and flattening mechanisms to promote generalisation and feature discrimination. Preprocessing techniques such as normalisation and one-hot encoding are applied to standardise the input feature space prior to model training.

Complementing the textual corpus, the study integrates synthetic audio generation using Coqui TTS's Swahili VITS model to simulate high-fidelity articulations of 31,197 Swahili words. This synthetic audio dataset, sampled at 16 kHz to balance clarity and computational efficiency, ensures phonetic coverage of nasalised consonants, vowel sequences, and syllable structures relevant to digraph classification. The Montreal Forced Aligner (MFA) is employed to align phonetic transcriptions with waveform segments

Empirical validation is executed through a stratified 70/15/15 data split and a 10-fold cross-validation procedure to ensure statistical robustness. Performance metrics

accuracy, precision, recall, F1-score, and AUC-ROC are computed, and the model's efficacy is benchmarked using Logistic Regression, Support Vector Machines (SVM), K-Nearest Neighbours (KNN), and Convolutional Neural Networks (CNN). Statistical significance is further assessed using ANOVA tests to substantiate observed performance improvements. The study is implemented in Python using NumPy, Pandas, TensorFlow, Keras, and scikit-learn.

1.7 Limitations of The Study

While this study provides valuable insights into Swahili digraph recognition through feature vector reduction and probabilistic classification, several limitations constrain the generalisability and scalability of its findings.

First, the availability and diversity of Swahili datasets remain a substantive challenge. Although authentic sources such as *Kamusi ya Kiswahili*, the Swahili Bible, and the Constitution of Kenya were incorporated, the overall corpus size is modest compared to high-resource language datasets, potentially limiting phonetic coverage across dialectal variations.

Second, the integration of secondary datasets from Mendeley, Harvard Dataverse, Zenodo, and Kaggle augmented the corpus, but linguistic diversity and phonological balance across regional Swahili variants may still be insufficient for universal applicability. Consequently, certain phonotactic and contextual nuances might be underrepresented.

Third, the model's performance evaluation is based on implicit assumptions regarding digraph classification thresholds, temporal segmentation accuracy, and statistical homogeneity across folds. These constraints could affect the interpretability of empirical results, especially in edge cases involving ambiguous or coarticulated phoneme boundaries.

Fourth, the feature engineering choices, such as dimensionality reduction via FRMLR and MFCC preprocessing, were optimised for computational tractability. However, they may have abstracted away finer acoustic details critical for differentiating

spectrally similar digraphs, limiting the model's sensitivity in phoneme-rich environments.

Lastly, the model's design is language-specific, tailored to Swahili's unique phonological features like vowel harmony, nasalised digraphs, and agglutinative morphology. This specificity may hinder its direct transferability to other languages.

Future work should explore larger and more dialectally inclusive corpora, adopt explicit annotation protocols for acoustic ambiguity, and experiment with hybrid feature extraction methods to bolster both linguistic granularity and model robustness.

1.8 Contributions of the Thesis

The following constituted the expected research contributions of the research, which range from improving theoretical understanding of linguistics to useful technological applications and possible effects on language processing policies.

a. Contributions to Theory

This study contributes to the theoretical advancement of computational linguistics by demonstrating how a CNN-based model integrating convolutional and dense layers can effectively extract and classify phonetic features from Swahili digraphs. The CNN architecture comprising convolutional, dropout, flattening, and fully connected dense layers supports the preservation of linguistic nuance while enabling scalable, high-accuracy processing in low-resource language contexts. The dense layer, in particular, aggregates spatially distributed features into a compact form suitable for classification, improving generalisation without sacrificing phonetic detail. The model, trained on a curated Swahili corpus, achieves 92.5% recognition accuracy, validating CNN-based modelling as an effective and linguistically sensitive tool for phoneme-level analysis in agglutinative and phonologically rich languages.

This study also contributes to the theoretical advancement of machine learning for linguistic data by demonstrating how Multinomial Logistic Regression (MLR) can transform dense output features into compact, 5-dimensional probabilistic digraph vectors. Through the Feature Reduction Model using Multinomial Logistic Regression (FRMLR), 39-dimensional Mel-Frequency Cepstral Coefficients (MFCCs) are

converted into low-dimensional yet semantically meaningful representations. This approach preserves phonetic fidelity while achieving substantial reductions in memory usage (by 70%) and training time (by 77.8%), thus enhancing the model's efficiency for real-time applications in resource-constrained environments.

The study advances the theoretical understanding of dimensionality reduction via MLR as a dual-purpose tool for both feature extraction and vector optimisation. It illustrates how probabilistic classification and reduction can coexist within a unified framework, maintaining classification accuracy while significantly lowering computational costs.

These contributions illustrate a coherent, empirically validated integration of deep learning and probabilistic modelling tailored to Swahili's linguistic structure. They provide foundational insights for designing efficient, scalable, and linguistically grounded NLP architectures capable of generalising across complex, morphologically rich, and underrepresented languages.

b. Contributions to Practice

This research makes significant practical contributions to the domain of Natural Language Processing (NLP), particularly with respect to the computational modelling of low-resource languages. A principal outcome is the creation of the first publicly accessible, annotated Swahili Digraph Corpus, comprising 31,197 lexicon entries. This resource includes rich linguistic metadata, such as digraph-vowel pair annotations and frequency distributions, and is specifically tailored to facilitate Swahili-oriented NLP applications. By addressing the resource deficit endemic to low-resource languages, the corpus enhances performance in speech processing, transcription, and machine translation, while simultaneously promoting linguistic equity in computational research.

The study proposes a replicable and scalable methodology for feature vector reduction that can be generalised across other linguistically underrepresented languages. This framework improves computational tractability without diminishing linguistic granularity, thereby offering a viable foundation for efficient model design in typologically diverse language contexts.

A key technological contribution is the design and deployment of the FRMLR (Feature Reduction Multinomial Logistic Regression) model, which demonstrates substantial performance improvements specifically, a 77.8% reduction in training time and a 70% decrease in memory usage compared to conventional based systems. This efficiency renders the model highly applicable to real-time, resource-constrained NLP environments, including speech-enabled assistive technologies such as healthcare chatbots deployed in bandwidth-limited rural regions across East Africa.

These practical outcomes extend the study's impact beyond theoretical inquiry, contributing to the development of sustainable NLP infrastructure, fostering equitable access to language technologies, and enabling scalable, context-aware solutions for linguistically marginalised communities.

1.9 Organisation of the Thesis

This thesis is systematically organised into seven interconnected chapters, each building upon the previous one to address the study's research objectives, methodologies, and findings comprehensively.

Chapter One: Introduction

This chapter lays the foundation by introducing the study's background, problem statement, research objectives, and corresponding research questions. It also outlines the significance by highlighting its contributions to theoretical advancement, practical implementation, and policy discourse in Swahili digraph recognition, while clearly defining the scope and limitations to contextualise the study's relevance and applicability.

Chapter Two: Literature Review

This chapter reviews literature on Swahili digraph recognition, covering phonetic and linguistic principles, theoretical frameworks, and feature extraction techniques like MFCC and RASTA-PLP. It critically examines optimisation strategies such as Multinomial Logistic Regression in digraph-level phonetic modelling, highlighting computational challenges and gaps in previous research to justify a specialised

recognition framework. Objective One was achieved in Chapter Two through a synthesis of Swahili's phonological structure, relevant linguistic theories, and computational methodologies for digraph identification and optimisation.

Chapter Three: Methodology

This chapter outlines the research design, including the integrated Design Science Research Methodology (DSRM) approach used for the study. The methodologies for data collection, corpus annotation, and feature extraction are explained, along with the tools and techniques used. The chapter also discusses evaluation metrics, internal and external validation, statistical analysis methods, and ethical considerations, ensuring transparency and rigor in the research process.

Chapter Four: Development Of the Annotated Swahili Digraph Corpus Using A CNN-Based Recognition Model

This chapter focuses on the technical implementation of the CNN-based digraph extraction model. It elaborates on the preprocessing steps, including normalisation and one-hot encoding, the architecture of convolutional layers and dense layers, and the annotated Swahili Digraph Corpus. The chapter concludes by presenting training results, evaluation metrics, and discussions on model performance, emphasising its significance in advancing Swahili language processing tasks.

Chapter Five: Feature Reduction Optimisation Using Multinomial Logistic Regression

In this chapter, the Feature Reduction Model using Multinomial Logistic Regression (FRMLR) is introduced. The chapter details how the 39-dimensional MFCC feature vectors are reduced to five-dimensional probability vectors, preserving phonetic integrity while improving computational efficiency. A comparative performance analysis against traditional methods such as PCA is also provided.

Chapter Six: Empirical Validation and Benchmarking

This chapter presents an in-depth validation and benchmarking of the FRMLR model. The chapter highlights the evaluation metrics, including accuracy, precision, recall, and F1-score, and compares the FRMLR model's performance against baseline models. Statistical tests like one-way ANOVA and AUC-ROC curve analysis underscore the reliability and effectiveness of the model for Swahili digraph recognition.

Chapter Seven: Summary, Conclusion and Recommendations

This final chapter synthesises the research findings, emphasising the contributions to computational linguistics, NLP, and ASR technologies. Actionable recommendations are proposed for further research, including extending the FRMLR approach to other low-resource languages and integrating the model into real-world applications. The chapter also highlights policy implications and the potential for transforming Swahili language processing through innovative solutions.

CHAPTER TWO

LITERATURE REVIEW

2.1 Introduction

This literature review offers a thorough overview of previous studies related to computer models, digraph recognition, and the processing of the Swahili language. It identifies gaps in the existing literature, highlights the significance of the study, and places the proposed research within a broader academic context.

With the use of speech signals, machines can now automatically recognise human voices thanks to a process called speech recognition. That is, it facilitates the establishment of a channel of communication between humans and machines. An acoustic and modelling algorithm is used by speech recognition software to perform its function. It enables communication between voice recognition software and computer interfaces. In the twenty-first century, speech recognition is one of the hot subjects of conversation. Due to scientists' tireless attempts to realise their goal of creating an algorithm that will enable robots to communicate with humans, the prevalence of technological devices in current society has increased [5] [1].

The fundamental technology employed in human-computer interaction processing is ASR, which aims to convert speech signals into textual information, thereby establishing a foundational framework for subsequent semantic understanding. ASR is a broad and multidisciplinary field of technology that encompasses various fields such as computer science, artificial intelligence, linguistics, digital signal processing, acoustics, statistics, and more [34]. The enhanced performance of ASR systems has resulted in their widespread adoption across various sectors, such as military, medical, and service industries. This advancement has notably reduced the reliance on human labour and boosted operational efficiency, underscoring the rapid progress within associated disciplines [35].

The process by which machines translate audible words into their printed equivalents is known as automatic speech recognition (ASR). Stated differently, ASRs look for a word string that, given an acoustic waveform, has the highest probability. An analogue series of utterances is translated from a recorded acoustic speech signal. Identifying

the basic building blocks of language, such as vowels, consonants and words that are put together to form sentences, is included in ASR [2] [36].

The goal of ASR is to accurately and effectively translate speech signals into spoken word equivalents in text or action [36] [37]. Achieving a level of performance close to human speech recognition is the main goal of ASR [38]. But despite advances in ASR over the years, this goal is still a long way off. One of the reasons for this is the variability caused by the characteristics of the speaker or the environment, leading to incompatibility between the testing and training environments [39]. The inability to capture distinctive auditory elements that describe the real nature of human speech equally impacts the performance of ASR [40].

The wide range of applications for ASR and the increasing number of interfaces and computing devices that allow speech processing have made ASR an active subject of research. Nonetheless, the majority of applications rely on well-resourced languages, which outperform those with inadequate resources. Unquestionably, ASR is a tool to advance such languages, particularly in the context of developing human-to-human or human-to-machine systems that involve illiterate individuals [41].

Speech recordings serve as a vital source of information about a community's culture. They capture not only the actual words spoken but also important phonetic nuances (such as the implosive consonants found in Bantu languages), prosodic features (including intonation and stress patterns), morphological complexity (like the agglutinative structures present in Swahili), and paralinguistic elements (such as pauses, laughter, and emphatic gestures) [42]. These details at the utterance level are crucial for preserving linguistic authenticity and cultural meaning. However, transcription remains a significant challenge, as traditional methods often oversimplify or overlook these essential aspects [43].

For example, Swahili’s vowel harmony rules and nasalised digraphs (*ng’a*) require precise phonetic annotation, while its agglutinative morphology (e.g., *nilikisoma* = "I had read it") demands morphological segmentation rarely supported by generic tools [44]. Prosodic markers essential for interpreting pragmatic intent in oral traditions are frequently lost in transcription [45]. This inadequacy undermines efforts to document

and revive indigenous languages, as partial representations fail to equip researchers and practitioners with the depth needed for pedagogical or computational applications. Addressing these gaps necessitates transcription frameworks that integrate multidimensional annotation (phonetic, prosodic, morphological) and leverage adaptive AI models trained on utterance-rich corpora.

The rapid advancement of ASR is mostly dependent on the availability of copious amounts of annotated text and speech data. While recognition capabilities for major languages like English and Chinese have reached a considerable level of advancement, numerous dialects and minority languages continue to face challenges in their implementation. Low-resource languages are those that fall into this category and include many different varieties that are devoid of resources such as text annotation, pronunciation dictionaries, recorded speech data, and linguistic expertise [26].

As Low resource languages attest, speech technology research remains the most direct and efficient method, even though standard ASR systems are not suitable for low resource languages that are hard to supply huge volumes of training data. With limited linguistic resources, an increasing number of academics are starting to undertake in-depth study and ongoing improvements on speech recognition technologies. Low-resource speech recognition is the work that is being discussed here, and it has emerged as a major problem and hot topic in the ASR field [35].

Research in low-resource speech recognition is of paramount importance, particularly in safeguarding and fostering linguistic diversity, cultural heritage, and global human communication. Given that many speech transcription systems were initially developed for English, their effectiveness in other languages is notably limited. Moreover, acquiring proficiency in even a fraction of the world's languages often requires years of dedicated study [46].

Modern ASR systems rely on a multi-stage pipeline that integrates signal processing, machine learning, and linguistic modelling.

Key steps in ASR systems were demonstrated by research studies [2] [36] [37] [38].

1. **Signal Preprocessing**

Noise reduction techniques are essential for enhancing the clarity of audio signals, particularly in speech processing. Approaches like spectral subtraction and Wiener filtering are commonly used to filter out background noise, ensuring that the primary audio is more distinguishable. In addition to noise reduction, effective processing requires digitising analogue speech signals, typically at a sampling rate ranging from 16 kHz to 44.1 kHz. Once digitised, these signals are segmented into frames of 20 to 40 milliseconds, often employing overlapping windows like the Hamming window. This method helps maintain continuity in the audio signal, improving the overall quality of speech recognition and analysis [2] [36].

2. **Feature Extraction**

To convert raw audio into meaningful discriminative features, techniques such as Mel-Frequency Cepstral Coefficients (MFCCs) or Perceptual Linear Prediction (PLP) are commonly employed. In the context of the Swahili language, it is essential to focus on specific acoustic characteristics. Notably, digraph-specific spectral patterns, such as those found in the combinations *"ny" and "ng',"* as well as the transitions between vowels and consonants, play a critical role in accurately representing the unique phonetic features of the language. These elements are crucial for effective speech processing and recognition tasks [36] [37].

3. **Acoustic Modelling**

Mapping acoustic features to phonetic units, such as phonemes and digraphs, can be effectively achieved through various machine learning models. Traditional Hidden Markov Models (HMMs) have proven to be effective for processing sequential data in speech recognition tasks. On the other hand, contemporary Deep Neural Networks (DNNs), which include Convolutional Neural Networks (CNNs) and Recurrent Neural Networks (RNNs), are particularly adept at capturing complex contextual relationships. This capability significantly enhances the accuracy of phonetic mapping. For Swahili-specific applications, it is beneficial to train models on the Annotated Swahili Digraph Corpus, which helps in recognising agglutinative

structures and understanding phonotactic constraints specific to the language [36] [37].

4. **Language Modelling**

Natural language processing has led to advanced models for predicting word sequences, such as n-grams and Transformer-based approaches. To improve contextual accuracy in Swahili, it's important to incorporate its linguistic rules, like vowel harmony and digraph frequency. By considering these factors, we can enhance the effectiveness of word prediction in Swahili and similar languages [37] [38].

5. **Decoding and Post-processing**

Combining acoustic and language model outputs is key for improving speech recognition in languages like Swahili. Utilising algorithms such as Viterbi decoding and beam search enhances transcription accuracy. Post-processing is crucial for correcting homophones, such as distinguishing between "*kupika*" ("to cook") and "*kupiga*" ("to hit"), ensuring clarity and context in the output [37] [38].

2.2 Phonetic and Linguistic Principles

The accurate identification and categorisation of Swahili digraphs rely heavily on phonetic and linguistic principles that account for the unique characteristics of the Swahili language. These principles play a crucial role in enhancing the precision and efficiency of digraph extraction models.

Swahili digraphs, such as "*ng'a*", "*sha*", "*nyu*", and "*tha*", exhibit distinct phonological characteristics that can be leveraged for accurate identification. Swahili typically follows vowel harmony rules, where vowels within a word harmonise based on their frontness or backness. This principal aids in predicting the likely combinations of vowels in digraphs and helps in their categorisation. The Swahili digraphs often include consonant clusters such as "*sh*," "*ng,"* "*ny*," and "*dh*," and recognising these clusters along with their phonetic properties can improve the accuracy of classification by distinguishing them from similar sounds. Digraphs, such as "*ng*," are nasalised, and understanding how nasalisation impacts speech sounds enhance the ability to recognise these digraphs effectively [47].

Swahili is a highly agglutinative language, meaning that affixes prefixes, suffixes, and infixes are added to root words to change their meanings. Digraphs often occur at specific morphological boundaries, making it essential to understand the morphological structure to identify their role in word formation, particularly within verbs, nouns, and adjectives. For instance, certain digraphs like "*ka*" and "*tu*" are frequently seen in verb conjugations and noun formations. Recognising their morphological context can facilitate accurate categorisation within a linguistic corpus. Digraphs also play a role in inflectional forms (relating to tense, aspect, and number) and derivational forms, which involve creating nouns or adjectives from verbs. By leveraging this linguistic information, one can gain a deeper understanding of how digraphs operate within sentences [48].

Intonation and stress patterns are critical linguistic features that influence the pronunciation and meaning of words in Swahili, particularly when digraphs are involved. In Swahili, stress typically falls on the penultimate syllable of a word, forming a predictable pattern that aids in pronunciation. However, variations can occur depending on the digraph and its phonological impact within the word, which can alter both the stress pattern and the meaning. For instance, while "*shangaa*" (to be amazed) and "*shauri*" (advice) share phonetic similarities, their stress placement and intonation distinguish their meanings [49].

A model that incorporates these stress patterns into its digraph extraction and annotation process can better capture the nuances of Swahili pronunciation. This is particularly valuable for differentiating words with similar spellings or phonetic structures but distinct meanings, as the stress pattern often serves as a linguistic marker for contextual and semantic disambiguation [49]. By analysing stress and intonation patterns during feature extraction, the model can refine its ability to recognise and categorise digraphs accurately. This approach enhances the quality of the annotated corpus and provides a deeper understanding of Swahili phonetics, contributing to the development of robust NLP applications such as speech recognition and language learning tools.

Swahili adheres to distinct phonotactic constraints, which are systematic rules governing permissible combinations of sounds within the language. These constraints

play a critical role in determining which consonant and vowel sequences can co-occur, thereby defining the structural integrity of Swahili digraphs. For instance, digraphs like *"ng'wa" and "nyu"* conform to Swahili's phonotactic rules, while combinations such as *"kp" or "xz"* are not permissible in standard Swahili. These rules serve as linguistic boundaries, ensuring that the digraph extraction model can differentiate between authentic Swahili digraphs and erroneous or non-existent combinations. Incorporating these constraints into the model provides a foundational mechanism for filtering out invalid sequences during the annotation process, enhancing both precision and linguistic fidelity [50].

By embedding these phonotactic principles, the model can identify patterns consistent with Swahili's phonological system, enabling it to reject anomalies and better adapt to the nuances of the language. This not only ensures accuracy in corpus development but also aligns the output with the linguistic standards of Swahili, thereby strengthening its applicability for speech recognition, transcription, and other NLP tasks. Linguistic principles also include understanding the contextual usage of digraphs. Digraphs in Swahili are not merely phonetic constructs; they carry significant grammatical and lexical functions, influencing both sentence structure and meaning. Their contextual application is vital for accurate recognition and categorisation [51].

In Swahili, digraphs often fulfil different roles depending on their syntactic environment. For instance, the digraph *"ny"* appears in *"nyumba"* (house) as part of a noun and in "*nyama*" (meat) with a similar function. However, its usage in phrases like "*nyinyi*" (you all) can serve as a subject marker. Proper syntactic analysis helps disambiguate these roles, requiring the model to consider positional and functional cues during digraph extraction.

Digraphs also affect meaning in Swahili, where identical phonetic forms may carry varying semantic interpretations [52]. For example, the digraph *"sha"* in *"kushauri"* (to advise) implies an action, while in *"kushangaa"* (to be amazed), it conveys an emotional state. A context-aware model can accurately categorise such digraphs by analysing neighbouring words or phrases for semantic alignment. This involves

leveraging word embeddings or semantic tagging in the annotated corpus to ensure the extraction process incorporates the full contextual breadth of digraph usage [53].

Integrating phonetic and linguistic principles can be highly beneficial when developing a Convolutional Neural Network (CNN)-based or other machine learning model for Swahili digraph extraction. This can be achieved through several key strategies. By incorporating phonological and morphological features into dataset annotations, the model can be guided to understand and recognise linguistic patterns more effectively [54].

Extracting and emphasising relevant phonetic and linguistic features, such as vowel harmony, consonant clusters, and affix boundaries, can significantly improve model accuracy by providing it with meaningful and contextually rich data. Training the model on these linguistically informed features ensures that it learns the contextual nuances of Swahili digraphs, which enhances its ability to correctly identify and categorise them.

By combining phonetic and linguistic principles with advanced computational models, the accuracy and effectiveness of Swahili digraph extraction and categorisation can be significantly enhanced [55]. This integration leads to the development of a more robust corpus that is valuable for natural language processing (NLP) and speech recognition tasks.

2.3 Frameworks

The theoretical framework is built on established language theories, machine learning principles, and optimisation techniques. The proposed model is developed by integrating these foundational theories, which also inform the research design.

2.3.1 Distinctive Feature Theory

A linguistic theory called Distinctive Feature Theory seeks to pinpoint and characterise the smallest, unique characteristics that set one phoneme in a language apart from another. This theory, which was created in the middle of the 20th century by linguists Roman Jakobson and Morris Halle, offers a methodical approach to

categorisation a language's sounds according to the existence or lack of specific characteristics [56].

The phonetic characteristics of phonemes, or distinctive qualities, are what enable them to contrast word meanings, such as voice and tense [57]. Distinctive feature Theory Trubetzkoy [58] [59] [60]and other representatives of the Prague School contested the presumption that a language's phonemes are unassailable. They modified that supposition and assert that there is a chance that the phoneme can be broken down into distinguishing characteristics.

Ferdnand de Saussure [61] was the one who critically viewed the distinctive features as fundamental components of systems phonological oppositions, despite the fact that many of the characteristics linked to them had been covered in previous phonetic works. Many of the fundamental concepts of contemporary distinguishing feature theory were established

The distinctive feature theory does not consider the phoneme as the relevant explanatory unit in phonology, but rather the minimal units of phonological analysis are the features. The DFT claims that it will be possible to generalise the relationships between sounds in language by allowing the features of phonemes to take place in this way. According to [62] [59] [60], one of the strengths of DFT is that it provides sufficient information that can be used to fully specify all phonemes and provide a classification system that captures all sounds, consonants, vowels and digraphs. Such a classification system is sure to provide an unusual description of each phoneme in the language system [60].

Distinctive Feature Theory has been shown to have practical uses in speech therapy, language acquisition research, and phonetic recognition system design, in addition to its theoretical contributions [63]. The science of linguistics has benefited greatly from the application of Distinctive Feature Theory, which has laid the groundwork for comprehending the systematic and contrastive nature of speech sounds in different languages. It is still a useful tool for analysing and characterising the phonological structures of various languages.

Distinctive Feature Theory adds to cross-linguistic typology by finding universal and language-specific features that enable linguists to compare and contrast the phonological systems of various languages. Distinctive feature theory, which offers a potent instrument for analysing and characterising the phonological structures of languages, has had a long-lasting influence on the study of linguistics. Its influence extends beyond theoretical linguistics into practical applications that address real-world linguistic challenges [64] [65].

This study applies Distinctive Feature Theory to identify and isolate specific phonetic characteristics that differentiate Swahili digraphs. These features such as voicing, nasality, and place of articulation form the basis of the MFCC-derived input vectors. By decomposing phonemes into distinctive traits, the theory helps ensure that the model captures essential phonological contrasts critical for accurate digraph classification.

2.3.2 Formal Language Theory (FLT)

The advent of computers and the digital revolution have brought about unprecedented levels of automation, connection, and information processing, which have fundamentally transformed society. Formal language theory is essential to the study of computation because it offers a theoretical framework for comprehending the composition, behaviour, and manipulation of languages.

Formal language theory, which has its roots in mathematical logic and linguistics, makes it possible to precisely specify programming languages, create interpreters and compilers, design and develop algorithms, and investigate computational complexity, all of which facilitate effective computation [66] [67].

Formal language theory (FLT) explores mathematically structured languages, defining them as sets of strings formed from a given alphabet according to specific rules. As an essential discipline within mathematics and theoretical computer science, FLT focuses on describing, manipulating, and classifying formal languages. This chapter will introduce fundamental concepts of FLT and examine its applications in linguistics, computer science, and animal cognition [68].

The computational complexity of linguistic patterns can be directly studied thanks to formal language theory, which is the mathematical study of string patterns. This allows for comparison of distinct linguistic theories, development of methods for machine processing of language, and the discovery of inductive principles for learning algorithms that generalise from positive data a context relevant to both human language acquisition and induction from text corpora [69].

Data structures, system specifications, and programming languages can all be precisely represented thanks to formal language theory. Formal grammar provides strict definitions for the syntax and structure of languages, including context-free and regular grammar. The creation of interpreters, compilers, and other language-processing tools is based on these representations. The formal languages facilitate the modelling and manipulation of complex data structures, ensuring efficient storage and retrieval of information [70]. A formal language, in the context of Formal Language Theory (FLT), refers to a set of finite sequences or strings that are composed of a limited vocabulary. In the case of natural languages, this vocabulary typically includes words, morphemes, or sounds. FLT provides a range of mathematical and algorithmic tools that enable us to define formal languages using finite means and process them computationally.

Automata theory, a branch of formal language theory, provides the theoretical framework needed to understand computing and language recognition. Automata, such as finite automata, pushdown automata, and Turing machines, are mathematical models of computer operations [63]. They enable investigations into formal language hierarchies, language recognition problems, and computability and decidability. Automata theory is essential to both algorithm design and understanding the capabilities and limitations of computing systems [71].

Formal language theory is essential to comprehending and processing human language in the field of natural language processing (NLP). Formal language representations are used by methods like syntactic parsing, grammar induction, and semantic analysis to simulate the structure and meaning of sentences in natural language [63]. A branch of formal language theory called formal semantics makes it possible to analyse and interpret language meaning rigorously, which helps with the creation of language

models, machine translation programs, and applications for intelligent language processing.

To guarantee the accuracy and dependability of algorithms and models used in natural language in the discipline of natural language processing, formal language theory is crucial to understanding and processing human language (NLP). Techniques such as syntactic parsing, grammar induction, and semantic analysis mimic the structure and meaning of sentences in natural language by utilising formal language representations [63]

Formal semantics, a subfield of formal language theory, enables rigorous analysis and interpretation of language meaning, which is useful for developing language models, machine translation systems, and applications for intelligent language processing. Formal approaches for NLP systems use FLT ideas to ensure the accuracy and reliability of models and algorithms used in natural language processing applications [72].

This study applies Formal Language Theory to structure and process Swahili digraph sequences as formal constructs within a computational framework. FLT informs the segmentation and rule-based annotation of digraphs in the training data, enabling the model to interpret these sequences systematically. This structured approach supports syntactic consistency and enhances the model's ability to generalise over patterned phoneme sequences.

2.3.3 Optimisation Theory

The goal of optimisation theory, a subfield of operations research and mathematics, is to identify the optimal solution from among all workable options for a given issue. Finding the best solution, given a set of constraints, may involve determining the maximum or lowest value of a particular objective function. Numerous disciplines, including engineering, economics, computer science, finance, and more, have used Optimisation theory as noted by [73] [74].

Optimisation theory provides a systematic framework for decision-making and resource allocation, allowing practitioners to make informed choices to achieve the

best outcomes given certain constraints. It is a fundamental tool in various disciplines, and ongoing research continues to advance the theory and develop new Optimisation techniques [75].

One of the primary ideas and elements of Optimisation theory is the objective function, a mathematical expression that must be maximised or minimised. The amount to be optimised is represented by it [75]. In order to maximise the objective function, the Optimisation problem looks for certain factors, or decision variables. The requirements or limitations that the solution must meet are known as constraints. The feasible zone, which is a collection of all potential decision variable values that satisfy the specified constraints, is represented by these constraints, which might be equality or inequality constraints [73].

The optimal solution refers to a set of values for decision variables that either maximises or minimises the objective function while adhering to the constraints [73]. It contrasts with local optima, which are solutions optimal only within specific neighbourhoods [76], and global optima, which represent the best solution across the entire feasible region [74].

Therefore, optimisation theory offers a methodical framework for allocating resources and making decisions, enabling practitioners to choose wisely to maximise results within predetermined bounds. It is an essential tool in many fields, and research is still being done to expand the theory and provide new Optimisation methods.

This study applies Optimisation Theory to design a computationally efficient framework for digraph recognition using feature vector reduction. The FRMLR model is guided by optimisation principles to minimise the loss function, reduce dimensionality, and tune model parameters such as dropout rates and learning rates. This ensures optimal trade-offs between accuracy and computational resource usage.

2.3.4 Information Theory

Information Theory, pioneered by Claude Shannon in the mid-20th century, serves as a mathematical framework for quantifying and analysing information in communication systems [77]. Its concepts have broad applications beyond

telecommunications, including fields like data compression, machine learning, natural language processing (NLP), and computational linguistics [78]. In the context of this research, Information Theory provides the theoretical foundation for feature reduction and ensures efficient data representation without compromising essential information [79].

The key concepts of Information Theory provided a robust foundation for feature reduction in linguistic data, particularly in the context of Swahili digraphs. Information entropy was pivotal in measuring the average uncertainty or unpredictability within the dataset, enabling the identification of features that added significant informational value [80]. By removing low-entropy features those considered redundant or irrelevant, the research reduced computational demands while maintaining the dataset's essential informational integrity. Redundancy, which represents repetitive or overlapping information in the dataset, was also addressed. While redundancy can occasionally improve error tolerance, its excessive presence in machine learning often results in inefficiency.

The study focused on eliminating overlapping or highly correlated features, which not only enhanced computational efficiency but also minimised noise in model training, ensuring a cleaner dataset for processing. The mutual information was employed to quantify the dependency between features. This concept evaluated how much the knowledge of one feature reduced uncertainty about another, helping to identify and exclude redundant features while preserving the most relevant ones for accurate digraph recognition. These principles of Information Theory collectively improved the quality of feature reduction, balancing computational efficiency with the preservation of linguistic data integrity [81].

Data reduction plays a vital role in converting large datasets into valuable information, which is the central objective of data science. A key aspect of this field is dimensionality reduction, which has traditionally focused on feature extraction. This process involves combining primary features linearly to maintain the covariance and relationships among those features. High-dimensional linguistic datasets, such as those involving Swahili digraphs, often consist of numerous features, many of which are irrelevant or redundant.

The application of Information Theory principles in this context ensures that feature reduction is a systematic, data-driven process rather than an arbitrary one. Features with minimal entropy or high redundancy are carefully excluded, leaving a low-dimensional feature set that retains the essential informational content necessary for tasks such as classification and recognition [82].

For instance, when analysing Swahili digraphs, phonetic and acoustic features are examined using metrics like entropy and redundancy. This method enables the removal of irrelevant features, such as subtle variations in pitch or redundant co-occurring digraph patterns, leading to a more streamlined dataset. This refined representation enhances processing speed while maintaining the accuracy of applications such as transcription and sentiment analysis.

In a broader sense, optimising the informational content of linguistic data through these principles has a profound impact on computational efficiency. By reducing dimensionality, the computational cost associated with training and inference is significantly minimised, enabling machine learning models to effectively process larger datasets. This improvement is particularly beneficial for low-resource languages like Swahili, where both data availability and computational resources are often constrained [83].

Beyond linguistic datasets, the application of Information Theory principles extends to various other domains, such as image processing and bioinformatics. The capacity to maximise relevant information while minimizing redundancy underscores the theory's versatility and enduring significance in data-driven fields.

Information Theory provides a structured approach to optimising information content in datasets [82]. Its application in this research preserved the essential features for Swahili digraph recognition while minimising computational inefficiencies. This approach not only enhanced the performance and scalability of NLP systems but also contributed to broader objectives like advancing digital inclusivity and fostering resource-efficient language processing technologies.

This study applies Information Theory to guide the encoding, transmission, and transformation of phonetic information through the recognition system. Concepts like entropy and information gain are used to filter redundant features during preprocessing and to prioritise high-information elements in the feature vectors. This theoretical lens ensures that the compressed representations used by the model retain critical linguistic details.

2.3.5 Statistical Learning Theory

Statistical Learning Theory, developed through the work of Vladimir Vapnik and Alexey Chervonenkis, forms a vital branch of machine learning that is centred on understanding and optimising the relationship between model complexity and predictive accuracy.

Its primary objective is to study how well a machine learning model can generalise to unseen data, ensuring that it performs reliably when deployed in real-world scenarios. This theory provides rigorous mathematical tools, such as the Vapnik-Chervonenkis (VC) dimension, empirical risk minimisation (ERM), and regularisation, to analyse and improve the learning processes of models, thereby addressing challenges such as overfitting, underfitting, and scalability [84].

According to [85]Overfitting occurs when a model is excessively tailored to the training data, capturing noise and specific patterns that fail to generalise to new data. Conversely, underfitting results when a model is too simplistic, failing to capture the underlying trends or structures in the data. Statistical Learning Theory seeks to strike a balance between these extremes by analysing the bias-variance trade-offs and providing techniques to control model complexity while maintaining high levels of predictive accuracy.

In this research, Statistical Learning Theory played an instrumental role in guiding the development and evaluation of the Feature Reduction Multinomial Logistic Regression (FRMLR) model. The theory was used to assess the model's generalisation performance, which refers to its ability to make accurate predictions on unseen Swahili digraph data. By leveraging principles such as the VC dimension and regularisation [86], the study analysed the trade-offs between the model's complexity and its

predictive performance. This ensured that the FRMLR model avoided overfitting while maintaining strong generalisation capabilities across diverse and unseen datasets.

Statistical Learning Theory [87] provided valuable insights for enhancing the integration of Multinomial Logistic Regression (MLR) and Convolutional Neural Networks (CNNs). The integration of these techniques leveraged MLR's strengths in feature reduction and CNNs' ability to learn complex patterns, creating a robust and scalable solution for feature extraction and dimensionality reduction. Statistical Learning Theory facilitated a deeper understanding of how these methods complemented each other, guiding their Optimisation to adapt effectively to the unique challenges posed by low-resource linguistic data such as Swahili digraphs.

A core aspect of Statistical Learning Theory is its focus on balancing complexity and generalisation, which was critical for the practical deployment of the models developed in this research. By optimising the trade-offs between these factors, the study ensured that the FRMLR model could adapt to linguistic variations and limitations without sacrificing predictive accuracy. This balance allowed the models to perform consistently across different datasets, making them more robust and reliable in real-world scenarios.

The scalability of the models was another key consideration addressed through Statistical Learning Theory. By applying regularisation techniques, the research controlled the complexity of the combined MLR and CNN approach, enabling it to process high-dimensional linguistic datasets efficiently. This scalability was particularly impactful for low-resource languages like Swahili [83], where data availability is limited and computational resources are often constrained.

The application of Statistical Learning Theory in this research not only advanced the theoretical understanding of generalisation and model evaluation but also yielded practical methodologies for natural language processing (NLP). The insights gained from the theory guided the development of machine learning models that were both reliable and adaptable, addressing challenges in feature extraction, dimensionality reduction, and classification for Swahili digraphs. The practical contributions of

Statistical Learning Theory extended beyond this specific application to inform best practices in computational linguistics and machine learning. The theory's emphasis on balancing complexity and generalisation provided a foundation for creating robust systems that can be applied to a wide range of linguistic and non-linguistic data challenges [85].

By embedding Statistical Learning Theory into the development and evaluation of machine learning models, this research achieved a significant balance between complexity and predictive performance. The FRMLR model, supported by the integration of MLR and CNNs, exemplified the theory's principles in action, providing scalable, efficient, and accurate solutions for feature reduction and linguistic data processing. The application of this theory not only enhanced the reliability and applicability of the models but also contributed to advancing machine learning practices for low-resource languages like Swahili, thereby furthering the broader goals of inclusivity and linguistic preservation.

This study applies Statistical Learning Theory to support the training, validation, and evaluation of the recognition models. It underpins the use of supervised learning, cross-validation, and probabilistic classification in the CNN and FRMLR frameworks. By focusing on statistical inference from observed data, the theory supports generalisation performance and justifies the use of metrics such as precision, recall, and F1-score.

2.3.6 Swahili's Lingua Franca Status

The phonetic subtleties and linguistic traits that distinguish Swahili digraphs are revealed by specialised research on the language's linguistic features. The literature review in this section emphasises the significance of choosing features that capture the distinctive qualities of Swahili digraphs, highlighting the necessity of a customised approach to feature extraction that is in line with the language's particular phonetic structure.

With more than 200 million people using it as their first or second language, Swahili is one of the top ten languages spoken worldwide. Swahili, which translates to "the coast," is the original language of the people who live on Africa's east coast. Since

their origins in East Africa, Swahili speakers have spread to over 14 nations, including the Democratic Republic of the Congo (DRC), South Sudan, Somalia, Mozambique, Malawi, Zambia, Comoros, Tanzania, Kenya, Uganda, Rwanda, Burundi, and even as far as Oman and Yemen in the Middle East. Along with South Africa and Botswana, Namibia and other countries in Southern Africa are considering introducing it in schools [52] [7] [88] [89].

Swahili, as a prominent lingua franca in Africa, has gained widespread recognition for its cultural, social, and economic significance. Its selection for computational linguistic research, specifically in digraph recognition, highlights its relevance in both academic and practical contexts.

Swahili serves as a medium of communication across diverse African regions, transcending ethnic and cultural boundaries. Historically, its evolution as a trade language along the East African coast facilitated interactions among traders, missionaries, and colonial administrators.

This linguistic legacy persists today, with Swahili acting as a bridge language in education, governance, and commerce. Swahili, or Kiswahili, has gained international recognition as a vital lingua franca, uniting diverse linguistic communities across Africa and beyond. Its significance is underscored by its designation as the first African language to have a dedicated day recognised by UNESCO World Kiswahili Language Day, celebrated annually on July 7 [90]. This milestone highlights Swahili's role in promoting cultural diversity, fostering dialogue among civilizations, and advancing multilingualism as a core value of the United Nations [91].

Mwalongo [92] points out that Swahili's potential for globalisation stems from its adoption as a national and official language in several countries, including Kenya, Tanzania, and Uganda, while also being one of the official languages of the African Union. Such recognition underscores its importance as a lingua franca capable of fostering unity and integration in a continent with over 2,000 languages. Similarly, Nkondo [93] emphasises the relevance of Swahili in advancing African knowledge systems and regional cooperation, aligning with pan-African ideals.

The use of Swahili as a lingua franca in this research ensures inclusivity and applicability, as it resonates with a wide audience across different domains. Swahili is spoken by over 200 million people worldwide, making it one of the most widely spoken languages in the world. Its influence extends beyond Africa, as it is taught in universities and institutions globally, from Japan to the United States. This global reach reflects its adaptability and its role as a bridge language in education, commerce,

The structural intricacies of Swahili make it a fascinating language for computational research. Its agglutinative nature, where affixes are added to root words to convey grammatical relationships, presents unique challenges and opportunities for Natural Language Processing (NLP). Of particular interest are Swahili digraphs combinations of two letters that create a single sound, such as "*ch*," "*ny*," and "*sh*." These digraphs contribute to the phonetic richness and semantic precision of the language.

Whiteley [94] underscores the significance of these digraphs in Swahili phonology, noting that they encapsulate unique sounds not easily represented in other linguistic systems. Efficient recognition and processing of these digraphs are critical for developing accurate NLP tools, such as speech recognition systems, machine translation, and text-to-speech technologies. As Maina *et al*. [95] illustrate in their study on CNN-based digraph extraction, addressing the complexities of Swahili digraphs enhances the performance and accuracy of computational models, making this a pivotal area of research.

Leshoele [96] highlights the challenges posed by the high-dimensional linguistic data inherent in African languages like Swahili, including morphological richness, lexical diversity, semantic ambiguity, and dialectal variation. Specifically, Swahili's agglutinative structure generates a vast number of word forms from root morphemes, complicating linguistic generalisation and data sparsity in computational models. The inclusion of lexical items from diverse indigenous languages and Arabic results in high intra-language variability, while semantic polysemy increases the complexity of disambiguation tasks. Regional dialects and pronunciation inconsistencies inflate the linguistic feature space, making model standardisation and training more difficult. These challenges are compounded by a lack of well-annotated corpora and structured

linguistic resources, ultimately hindering efficient dimensionality reduction and scalable NLP implementations for Swahili and similar African languages.

The FRMLR model addresses these challenges by employing techniques such as Mel-Frequency Cepstral Coefficients (MFCCs) for feature extraction, which captures phonetic information crucial for distinguishing Swahili digraphs. By integrating Multinomial Logistic Regression, the model ensures robust classification of digraphs, balancing computational efficiency with accuracy. This approach is particularly significant given Swahili's growing prominence in digital platforms, where accurate and efficient processing of its linguistic features is essential.

All languages utilise consonants, vowels, and digraphs, yet no two languages possess an identical set of distinct sounds, or phonemes. These sounds are the voiced components of speech and are a critical aspect of speech recognition, which is generally accomplished effortlessly by humans and machines. The accurate recognition of vowel sounds, consonants, and digraphs is essential for improving the overall performance of Automatic Speech Recognition (ASR) systems [97].

A digraph is a pair of letters combined to produce a single phoneme, or sound. For instance, the phoneme */f/* is formed by the combination of the letters PH in the word *graph *. The phoneme */ch/* is formed by the combination of the letters CH in *chin *and the letters CK in duck, which combined produce the */k/* sound. The Swahili alphabet is identical to the international Latin alphabet, with letters assigned to represent Swahili phonemes. It employs digraphs, which are sequences of two letters, to symbolise individual sounds. According to a source cited in [7] [88] [89], the Swahili alphabet comprises five vowels: *a, e, i, o, u*. Nineteen consonants are present: *b, c, d, f, g, h, j, k, l, m, n, p, r, s, t, v, w, y, z* (except x and q). The alphabet includes nine digraphs: *ch, dh, gh, kh, ng', ny, sh, th, ng.*

Swahili digraphs consist of consonant combinations, where they are pronounced as single cohesive sounds. For instance, in English, both "*dh*" and "*th*" are pronounced as "*th*." The sound of "*dh*" is akin to the "*th*" in "the," "this," "that," and "with," while the "*th*" in "thought," "thin," and "both" remains silent, similar to the Swahili word "*stakabadhi*" (meaning "receipt") and "*hadithi*" (meaning "story"). The "*gh*" sound,

resembling the French "r" in "rare," is voiced, as seen in words like "*ghali*" (meaning "expensive") and "*shughuli*" (meaning "affair" or "activity"). On the other hand, the "*kh*" sound is voiceless and involves a scratching sensation in the throat, as evident in "*subalkheri*" (meaning "good morning"). Despite the similarity in sound to the English "*ng*" in "sing," the presence of "*ng*" at the beginning of words poses a challenge in pronunciation.

The proposed study is grounded in the body of literature on Swahili language processing. Numerous facets of Swahili linguistics, such as phonetic characteristics, digraph patterns, and linguistic complexity, have been studied in the past. Although a great deal of information exists regarding Swahili language, there is a clear lack of specialised models that address digraph recognition, particularly those that are optimised for computational efficiency. This gap emphasises how important it is for the planned study to advance the field of Swahili language processing.

Swahili's adoption in this research is justified by its linguistic, cultural, and practical significance. As Africa's most widely spoken lingua franca, Swahili holds immense potential for unifying diverse linguistic communities. Leshoele [96] argues for Swahili's role as the language of commerce and epistemology, particularly in supporting regional initiatives like the African Continental Free Trade Area. By enabling effective communication and information exchange, Swahili contributes to the economic and social integration of the continent.

Onzere *et al*. [98]further emphasise the importance of addressing language variations in Swahili to enhance communication and integration among speakers from different regions. This research aligns with such goals by standardising digraph recognition, thereby fostering linguistic cohesion. The success of language-specific models like SwahBERT, as demonstrated by Martin *et al*. [99], reinforces the need for tailored computational tools to support underrepresented languages. By focusing on Swahili, this study not only advances NLP methodologies but also contributes to the preservation and promotion of an African linguistic heritage.

Through the Optimisation of Swahili digraph recognition using the FRMLR model, this research exemplifies the intersection of linguistic preservation and technological

innovation. Swahili's status as a lingua franca and its rich linguistic features makes it an ideal focus for computational research, ensuring its relevance in the digital age. By enhancing the accuracy and efficiency of NLP tools for Swahili, this study supports the broader goals of inclusivity, integration, and linguistic diversity in Africa.

2.4 Machine Learning Paradigms

Machine Learning (ML) is a subfield of Artificial Intelligence (AI) that focuses on developing algorithms and statistical models that enable computers to perform tasks without explicit programming [100]. Instead of following predefined instructions, ML systems analyse patterns in data and learn from past experiences to make predictions, classifications, or decisions. These systems improve their performance iteratively as they are exposed to more data, allowing them to adapt to complex and dynamic problems.

Machine learning enables computers to learn from data and perform tasks traditionally done by humans more efficiently. By automatically identifying patterns and refining its methods over time, machine learning allows systems to process complex tasks such as speech recognition, Natural language processing, image classification, and fraud detection with greater speed, accuracy, and scalability than manual human efforts [101].

By leveraging computational algorithms, machine learning empowers machines to autonomously access and retrieve data, while continuously enhancing their performance through accumulated experience. It has simplified life and become an indispensable instrument in several industries, such as agriculture [102], banking [102], optimisation [103], robotics [104], structural health monitoring [105], and so on. It can be applied in cameras for object detection, image, colour, and pattern recognition, as well as for data gathering, organisation, and converting audio into text [106]. ML is widely applied in various domains, including natural language processing (NLP), computer vision, fraud detection, healthcare diagnostics, and autonomous systems. As advancements in computing power and data availability continue, ML models are becoming more sophisticated, leading to significant breakthroughs in AI-driven technologies [107].

The core principle of ML is to identify patterns and relationships between input features like characters in a text, pixel values in an image, sensor readings) and corresponding output labels e.g., digraph classes, object categories, disease diagnoses) [108]. This is achieved by optimising a mathematical objective, such as minimising prediction error through techniques like gradient descent. Different types of ML approaches exist, including supervised learning (where models learn from labelled data), unsupervised learning (where patterns are identified in unlabelled data), and reinforcement learning (where agents learn optimal behaviours through rewards and penalties).

Machine Learning (ML) is broadly categorised into supervised learning, unsupervised learning, semi-supervised learning, and reinforcement learning. Each learning paradigm serves distinct functions depending on the nature of the available data and the desired outcomes. This classification is crucial in determining the appropriate computational models for various domains, including natural language processing (NLP), image recognition, and autonomous systems.

2.4.1 Supervised Learning

Supervised Learning is a machine learning paradigm where algorithms learn to map input data (features) to known output labels (targets) using a labelled dataset [109]. The goal is to train a model to make accurate predictions or classifications by learning the relationship between inputs and outputs from examples provided during training. Supervised learning is foundational to modern AI, enabling machines to solve real-world problems by learning from labelled examples [110].

The primary goal of supervised learning is to develop a computational model capable of accurately predicting or classifying target outcomes for unseen data by inferring the underlying relationship between input features and their corresponding labelled outputs. This paradigm relies on a labelled training dataset, where each input instance is paired with a known output, to learn a mapping function $f: X \rightarrow Y f: X \rightarrow Y$ that generalises beyond the training examples [111]. The process involves minimizing a predefined loss function (e.g., cross-entropy for classification, mean squared error for regression) to quantify and reduce discrepancies between predicted and true labels, thereby optimising the model's parameters.

A critical challenge lies in balancing model complexity to avoid overfitting, where the model memorises training noise, or underfitting, where it fails to capture essential patterns. Techniques such as regularisation (e.g., L1/L2 penalties) and validation on hold-out datasets are employed to enhance generalisation [112]. For instance, in Swahili digraph recognition, supervised learning enables models to associate contextual character sequences (e.g., "*c*" followed by "*h*") with linguistically accurate labels (e.g., the digraph "*ch*"), even when applied to novel text. By achieving this goal, supervised learning underpins applications ranging from medical diagnostics to natural language processing, ensuring robust, data-driven decision-making in dynamic environments [113].

Supervised learning techniques are versatile and have demonstrated remarkable success across diverse fields. The following highlights some key applications:

1. Text Classification and Sentiment Analysis: Within the domain of Natural Language Processing (NLP), supervised learning is instrumental in tasks such as spam email detection and sentiment analysis. Algorithms like Naïve Bayes and Support Vector Machines (SVMs) are commonly employed to classify textual data. For instance, spam detection involves labelling emails as "spam" or "not spam," while sentiment analysis evaluates the polarity (positive, neutral, or negative) of textual content, particularly in social media platforms [114].
2. Medical Diagnosis and Radiology: Supervised learning, particularly using Convolutional Neural Networks (CNNs), plays a pivotal role in medical imaging diagnostics. These models analyse radiological scans such as X-rays or MRIs to identify diseases in their early stages, including diabetic retinopathy, lung cancer, and other conditions. This application significantly enhances the accuracy and efficiency of diagnostic processes [115].
3. Fraud Detection in Financial Transactions: Supervised machine learning models are extensively utilized to combat fraud in the financial sector. By analysing historical transaction data, these models learn patterns of typical financial behaviour. When anomalies such as suspicious spending patterns are

detected, the system flags them as potentially fraudulent, enabling proactive measures to mitigate risks [116].

An illustrative example of supervised learning can be observed in the context of linguistic applications, such as Swahili digraph recognition. In this scenario, a supervised learning model is trained using annotated Swahili textual corpora, where digraphs like "*ng*," "*ny*," and "*sh*" are explicitly labelled. By analysing these labelled samples, the model learns to classify digraphs accurately within the language's phonetic and orthographic systems. Consequently, the trained model demonstrates the capability to predict and classify digraphs in previously unseen Swahili text. This methodology is particularly beneficial for enhancing Swahili speech recognition systems by enabling more precise phonetic transcription, thus contributing to improved linguistic resource accessibility and preservation for underrepresented languages.

2.4.2 Unsupervised Learning

In this methodology, the learning algorithm is trained using a dataset that lacks any associated labelled outputs, distinguishing it from the supervised learning paradigm. Each input instance is processed without an explicit indication of correct or incorrect outcomes, and unlike supervised learning, there is no external human intervention to refine or adjust the results. Consequently, unsupervised learning is characterised by a greater degree of unpredictability in comparison to its supervised counterpart [117].

The primary objective of unsupervised learning is to deeply comprehend a dataset by revealing its natural structure and distribution patterns. In this approach, the algorithm strives to identify and encode distinct features from the input, while also abstracting these characteristics so that they resonate with the dataset’s overall organisation. Through this autonomous learning process, the algorithm groups various inputs based on the distinctive features extracted from each instance. Consequently, unsupervised learning is particularly effective in addressing problems related to clustering and association [118].

Unsupervised learning methodologies have proven their utility across a range of real-world applications, facilitating knowledge discovery and decision-making processes. Prominent examples include:

1. Customer Segmentation in E-Commerce: In the domain of marketing, algorithms such as K-Means clustering and Hierarchical Clustering are employed to segment customers based on their purchasing behaviour. By analysing transactional data, these techniques group individuals into clusters with shared characteristics, enabling businesses to tailor targeted recommendations and personalised marketing strategies effectively [119].
2. Anomaly Detection in Cybersecurity: Anomaly detection represents a critical application of unsupervised learning in cybersecurity. By examining network traffic and user behaviour patterns, unsupervised models identify deviations from established norms. These deviations often signify potential cyber threats or fraudulent activities, allowing for pre-emptive measures to secure systems and data [120].
3. Topic Modelling in Natural Language Processing (NLP): In NLP, unsupervised algorithms like Latent Dirichlet Allocation (LDA) and Non-Negative Matrix Factorisation (NMF) are widely utilised for topic modelling tasks. These techniques analyse textual corpora to uncover latent semantic themes within the data. Applications include examining collections of news articles, academic literature, or large-scale text datasets to discern prevalent topics, enabling informed decision-making in information-heavy domains [121].

2.4.3 Semi-Supervised Learning

Semi-supervised learning is a hybrid approach that incorporates a small amount of labelled data with a large volume of unlabelled data to improve model generalisation. This is particularly useful in scenarios where data annotation is expensive, time-consuming, or impractical. This method leverages a substantial volume of input data, comprising both labelled and unlabelled examples, and occupies an intermediary position between supervised and unsupervised learning paradigms [122].

Semi-supervised learning is particularly well-suited to addressing a wide range of practical learning challenges encountered in real-world scenarios. Its reliance on a large proportion of unlabelled data, combined with a relatively small amount of labelled data, minimises the need for extensive human intervention [110]. The appeal of this approach stems from the fact that acquiring labelled datasets is often challenging, costly, and may require the expertise of domain specialists. Conversely, unlabelled datasets are generally more accessible, less expensive, and simpler to obtain, making them an attractive alternative for training machine learning models [123].

Semi-supervised learning integrates both supervised and unsupervised approaches to enhance the training of learning algorithms. Through unsupervised methods, latent patterns and structures within the input dataset can be uncovered, facilitating a deeper understanding of the data. On the other hand, supervised methods are leveraged to generate predictions from unlabelled data based on the best available approximations, which can then be applied to new datasets. Unlabelled data play a pivotal role in refining or reassessing hypotheses or predictions originally derived from labelled data, thereby improving the overall model accuracy and adaptability. [124].

Semi-supervised learning methods utilise unlabelled training data by relying on one of three fundamental assumptions: the smoothness, the cluster, or the manifold assumption. These principles guide the algorithm's ability to generalise from limited labelled data while effectively utilising the unlabelled data to uncover meaningful patterns or structures. In hybrid learning architectures, supervised (or "discriminative") learning components are integrated with unsupervised (or "generative") components. By blending these distinct methodologies, a hybrid deep neural network can be designed, enhancing the model's capacity to learn from both types of data. One notable application involves employing action bank features precomputed visual features derived from a bank of action templates to improve the recognition of human activities. This approach is anticipated to yield significantly enhanced outcomes by combining the strengths of multiple learning paradigms [110].

Semi-supervised machine learning has emerged as a powerful tool across various domains, with notable applications in the healthcare sector. It is essential for tasks like

identifying and analysing speech, as well as classifying and managing digital content. Beyond healthcare, this technology demonstrates its versatility by addressing challenges in diverse areas, including regulatory compliance and oversight. Its ability to analyse both voice and imagery data with heightened accuracy makes it a valuable asset for solving complex real-world problems.

2.4.4 Reinforcement Learning (RL)

Reinforcement learning (RL) is a decision-making paradigm where an agent learns optimal actions through interactions with an environment, receiving rewards or penalties based on its performance.

Reinforcement learning involves an agent learning how to act within an environment by interacting with it and receiving feedback, rather than by following explicit instructions [125]. It selects its next action by balancing what it already knows (exploitation) with trying out new possibilities (exploration), essentially employing a trial-and-error approach. After each action, the reinforcement learning agent receives feedback in the form of a monetary reward that signals how successful that action was, and it works to learn choices that maximise the overall reward value [126]. Actions may not only affect the present situation and the immediate reward but also shape future conditions and influence future rewards.

Learning agents are designed with clear objectives and possess the ability to sense aspects of their environment. This capability allows them to act and modify the environment's state to move closer to achieving their goals. When comparing learning methods, supervised and reinforcement learning take fundamentally different approaches. In supervised learning, an external 'teacher' provides examples or case studies to guide the learning process. In contrast, reinforcement learning builds knowledge through the agent's direct interactions with its environment, learning from the outcomes of its actions [126].

Reinforcement learning trains an agent by having it interact with its environment, learning from a system of rewards and penalties. The agent is not guided by explicit instructions; instead, it receives rewards for correct actions (like a self-driving car safely reaching its destination or a chess program winning) and penalties for mistakes

(such as veering off the road or being checkmated). The agent continuously adjusts its strategy through trial and error to maximise rewards and minimise punishments. [127].

For the subsequent step determined by the learning model, reinforcement learning employs a system of rewards and penalties. This mechanism is commonly applied in fields such as gaming and robotics to address decision-making challenges [128]. In their study [129], they detailed the progress made in applying deep learning techniques to reinforcement learning (RL) and highlighted how deep feedforward neural networks (FNN) and recurrent neural networks (RNN) have been employed to enhance RL performance.

Reinforcement learning is effective with sparse data and is widely used in the gambling industry, video game development like Grand Theft Auto, and self-driving cars for navigation. It gained attention with AlphaGo's win in Go and is also applied in natural language processing. This technique is increasingly impacting everyday life, from digital services to transportation [101] [130].

Deep neural networks (DNNs) have proven to be highly successful in supervised learning tasks. Deep learning (DL) models perform exceptionally well in unsupervised learning, hybrid approaches, and reinforcement learning scenarios [129].

Figure 2.1 delineates the principal paradigms of machine learning and enumerates representative algorithms associated with each category. It organises these approaches into four primary groups, each addressing a distinct learning paradigm and set of computational problems.

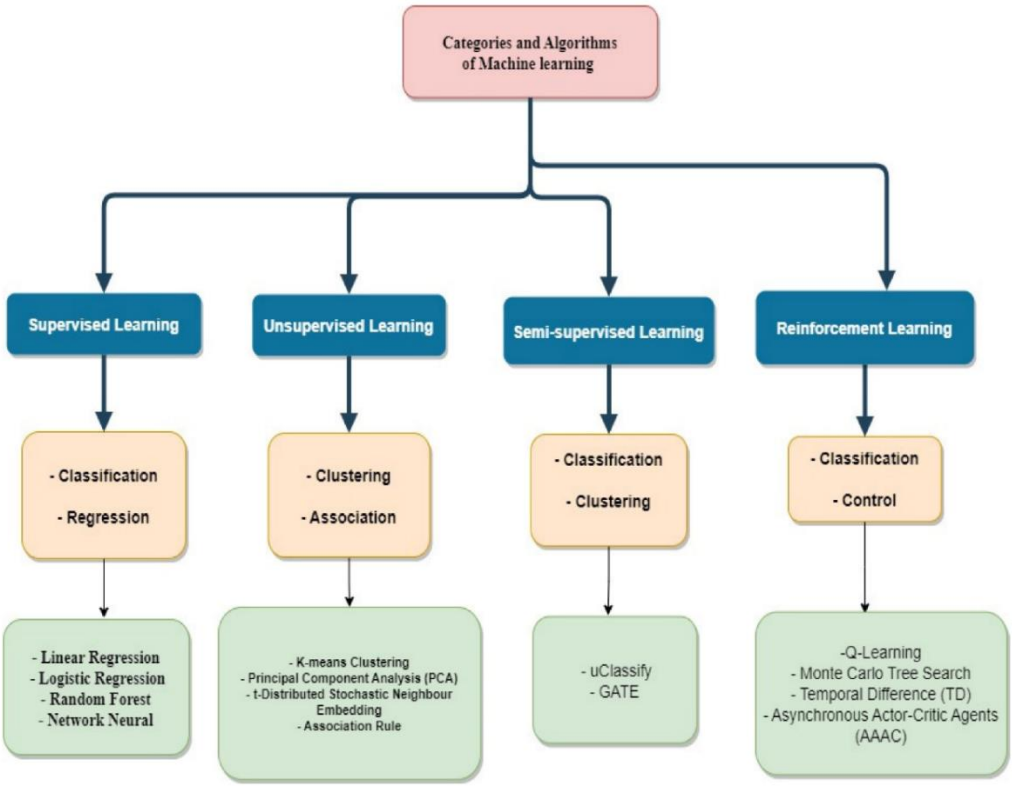


Figure 2.1 Machine learning categories and algorithms [131]

Table 2.1 provides a comparative overview of four primary machine learning paradigms: Supervised Learning, Unsupervised Learning, Semi-Supervised Learning, and Reinforcement Learning.

Table 2.1: ML Paradigms and Their Applications

ML Type	Definition	Real-World Application
Supervised Learning	Uses labelled data to map inputs to outputs	Email spam classification, medical image diagnosis, fraud detection
Unsupervised Learning	Identifies patterns in unlabelled data	Customer segmentation, topic modelling, anomaly detection
Semi-Supervised Learning	Leverages both labelled and unlabelled data	Google Translate for low-resource languages, speech recognition
Reinforcement Learning	Optimises sequential decision-making via rewards and penalties	Autonomous driving, robotics, strategic AI games

For each type, Table 2.1 concisely outlines:

- Definition: A brief description of how the paradigm operates, focusing on its approach to data and learning.
- Real-World Application: Examples of practical applications that illustrate the paradigm's capabilities in diverse fields such as healthcare, finance, and language processing.
- Example in Swahili Digraph Recognition: A specific instance of how each learning method may be applied to the task of Swahili digraph recognition, highlighting its relevance or limitations in addressing this linguistic challenge.

In this context, while supervised, unsupervised, and semi-supervised learning can be directly adapted to improve Swahili digraph recognition tasks, reinforcement learning is noted as less applicable due to its design for sequential decision-making rather than static classification. This organised framework assists researchers and practitioners in selecting the most appropriate machine-learning strategy for their specific challenges**.**

Choosing the right machine learning approach depends on both the task's nuances and the availability of annotated data. For Swahili digraph recognition, supervised learning is the most fitting paradigm. This is because the study leverages a labelled corpus of 31,197 Swahili words, each annotated with digraph-vowel pairs and phonetic features, enabling the model to learn from input-output mappings [131]. As noted, [84][85]Statistical Learning Theory, supervised learning is critical for enabling the model to generalise from observed examples and perform accurate classification. The success of this paradigm is evidenced by the model achieving 92.5% classification accuracy, with efficiency gains of 77.8% in training time and 70% in memory usage [114].

These metrics demonstrate that the supervised learning framework, grounded in statistical inference and supported by annotated data, is both effective and computationally practical for digraph recognition in low-resource Swahili ASR contexts. This method taps into explicitly labelled datasets where each instance of a digraph like "*ch," "sh,"* or *"ng"*" is identified, allowing models to learn how specific

contextual features (such as neighbouring characters and word positions) map directly to distinct digraph classes.

The task is inherently a multi-class classification problem. It demands that each sequence be assigned to one of the predetermined digraph categories, a challenge that supervised learning is uniquely positioned to handle through techniques like probabilistic classification and SoftMax activation. Although unsupervised methods might offer some assistance in exploring linguistic patterns and semi-supervised approaches could be useful when labelled data is limited, neither can provide the level of precision that Swahili's orthographic patterns require. Reinforcement learning, which is designed for sequential decision-making, does not address the static nature of this classification task.

2.5 Feature Extraction in Language Processing

The literature reveals a wide range of feature extraction techniques, from conventional methods to more sophisticated approaches, used in different language processing tasks. Research has investigated the effective representation of linguistic elements through statistical, linguistic, and syntactic features. It is imperative to comprehend the terrain of feature extraction techniques in more general language processing scenarios in order to modify and enhance these methods for Swahili digraph recognition.

2.5.1 Linear prediction coding (LPC)

The LPC is a FE approach that mimics the features of the human voice production system's vocal tract and A voice sample roughly mixes previous speech samples linearly in Linear Predictive Coding (LPC) analysis. The analysis of the speech signal using LPC is frame-based [13], [125], [126].

The main goal of linear prediction is to predict the output samples using a linear mixture of input samples, historical samples, or both. LPC synthesis mimics speech production by humans. In quasi-state regions of voiced speech, where the all-pole model of LPC provides a good approximation of the spectral envelope of the vocal tract, LPC helps create an accurate audio signal model [126][127].

Using an all-pole infinite impulse response (IIR) filter, LPC models the speech signal spectrum as an autoregressive ordering system that predicts an unknown speech sample by combining previously known sample combinations [130]. The human vocal tract is represented by the IIR filter as a linear time-invariant system for brief intervals of time. A speech signal's transfer function provides the vocal characteristic that sets one person's voice apart from another. The speech signal in the time domain is produced by the convolution of the excitation function and the transfer function [132] [133].

The all-pole LPC vocal tract model can be demonstrated using a customised piecewise cylindrical acoustic tube model [131],[132]. On the other hand, the understanding is clearer when the vocal tract filters are used as ladder filters Despite its drawbacks from uniform weighting applied throughout the spectrum, which leads to worse performance than MFCC [134], LPC is still widely used because of its simple implementation.

Due to its powerful speech analysis approaches to estimating basic speech characteristics, including pitch, formants, spectra, and vocal tract functions, as well as expressing speech via low-bitrate transmission or storage, LPC analysis is the preferred FE method among ASR researchers. Other advantages include the high speed of calculation and the incredibly accurate estimation of speech parameters [133].

2.5.2 Mel-Frequency Cepstrum Coefficients

This method for feature extraction is standardised. It is predicated on the known frequency variation of the crucial bandwidth of the human ear [134] [135]. The frequency information of the input speech signal can be easily, quickly, and consistently computed into coefficients using MFCC. The Mel-frequency cepstral coefficient is primarily used for speech processing because it mimics the structure of the human auditory system [136]

MFCC has been one of the most widely used ASR functions since it was first introduced by [137] [138]. The basis for perceptually motivated speech representations (MFCCs) is the logarithmic filter bank amplitudes' discrete Fourier cosine transform. Every filter in the MFCC computes the average spectrum surrounding each central

frequency, which is based on the Mel frequency scale [139]. It is based on the auditory system of humans.

For frequencies above 1 kHz, the Mel frequency scale is roughly logarithmic; for frequencies below, it is linear. Equation 2.1, which shows a frequency distortion approximation to obtain the equivalent Mel frequency values, uses f to represent a linear frequency scale (Hertz) and W(f) to represent the perceived frequency (Mel).

$$W(f) = 2595 log10 \left(1 + \frac{f}{100}\right) \quad \textbf{(Eq. 2.1)}$$

In clean speech, MFCC has performed better than other FE techniques like PLP and LPC. Because of this, ASR researchers like [132] [133] [140] [141]employ it the most frequently. The most advanced ASRs are generally agreed to have more efficient vocal cord excitation signals.

The primary goal of Mel-Frequency Cepstral Coefficients (MFCC) is to capture vocal tract characteristics from speech recordings. Nonetheless, a notable drawback of MFCC is its susceptibility to additive noise, such as background noise, due to the sensitivity of the nonlinear logarithmic function employed in cepstrum generation [1].

In order to mitigate this drawback, [142] suggest a hybrid approach that combines MFCC and a discrete wavelet transform to solve the phoneme recognition task's issue. In the meantime, [143] suggested using PCA to determine the Mel-filter coefficients in order to alter the traditional shape of the filters for improved speech recognition task performance and noise robustness.

By include dynamic features, such as the temporal slopes derived from regression on the MFCC coefficients, their efficiency can be greatly increased. The first and second derivatives of MFCC and energy characteristics have been shown to significantly improve recognition performance for noise. These techniques are related to strategies for cepstral domain filtering.

To incorporate information about the dynamic characteristics of the speech signal over time, the feature vector often includes the first and second derivatives of the Mel-Frequency Cepstral Coefficients (MFCC) [144]. MFCC derivatives are typically calculated over a large number of consecutive frames that capture the dynamics of the neighbouring speech sound for about 50 Ms.

Another useful feature that MFCC and its successors have been granted is the overall energy of a frame of speech. This is due to the fact that the speech signal's overall energy includes voicing information. MFCCs are commonly used as a feature extraction technique in automatic speech recognition (ASR) research because of their strong performance, making them a standard baseline for computing probability ratios in multinomial logistic regression. Figure 2.2 illustrates the MFCC feature extraction process

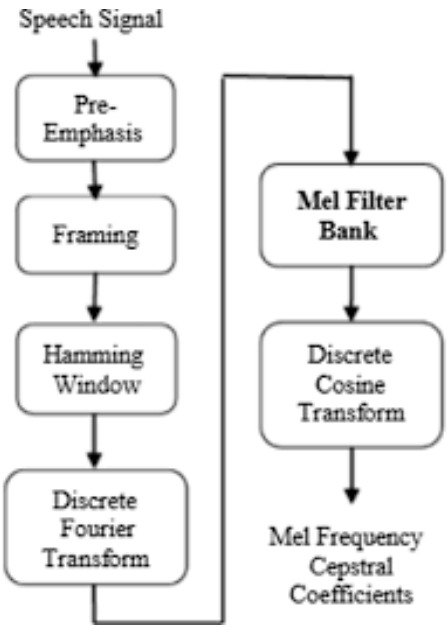


Figure 2.2 MFCC feature extraction process [145]

2.5.3 Perceptual Linear Prediction

Perceptual Linear Prediction (PLP) is a prominent parametric representation of speech signals utilised extensively in speech and audio signal processing [146]. Tailored to emulate the perceptual characteristics of the human auditory system, PLP finds applicability in various domains including speech coding, speech recognition, and audio compression. PLP is an improvement over Linear Predictive Coding (LPC) by incorporating perceptual features [147] [148].

The vocal tract's optimal features are the only ones that the PLP model can faithfully capture. Using concepts from the psychophysics of hearing, PLP, an auditory-based representation, approximates the auditory spectrum by incorporating elements such as the equal loudness curve, the intensity-loudness power law, and the critical-band spectral resolution [149].

In PLP, the changed power spectrum is smoothed using an all-pole model. On the basis of this model, the output cepstral coefficients are then calculated. PLP aims to integrate the advantages of LPC and MFCC [150]. The only major distinction between PLP and LPC is that the speech spectral qualities are changed to better match the characteristics of the human hearing system [151].

Like any other FE strategy, PLP has its limitations. One such instance is the potential for spectral smoothing brought on by the critical band filtering process, which could impair one's ability to discriminate in recognition tasks. Another drawback of PLP features is their great sensitivity to noise and channel distortions because they rely on the entire spectral balance of the formant amplitudes. The process of PLP is shown in Figure 2.3 below

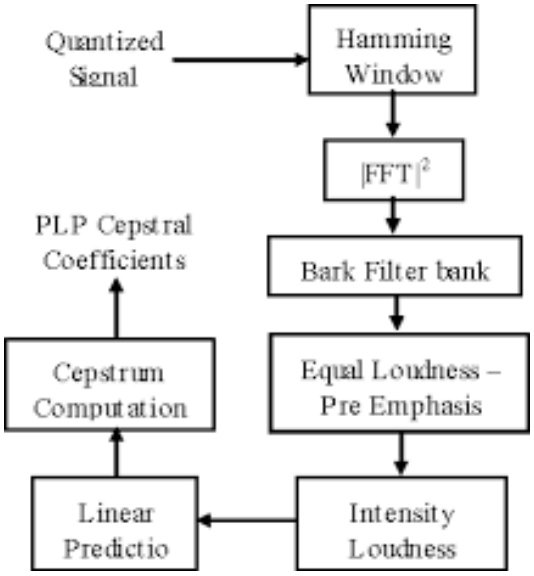


Figure 2.3 The process of PLP [152]

2.5.4 Relative Spectral-Perceptual Linear Prediction (RASTA-PLP)

The RASTA-PLP method combines the benefits of both PLP (perceptual linear prediction) and RASTA (relative spectral transform). The speech recognition process uses the spectrally-perceptual linear prediction (RASTA-PLP) approach for feature extraction.

In RASTA processing, the voice signal is filtered to eliminate the spectral domain's sluggish changes. RASTA is designed to mimic the properties of the human auditory system, which is more responsive to sudden changes in spectral content than to gradual shifts. A method called Perceptual Linear Prediction attempts to simulate the perceptual properties of the human auditory system. In order to improve the signal's representation and make it more in line with human perception, it first extracts linear prediction coefficients from the speech signal and then goes through a number of processing stages [136].

RASTA-PLP combines the benefits of RASTA and PLP. It first applies RASTA processing to the speech signal to remove slow spectral changes. After that, PLP analysis is performed on the processed signal to extract perceptually relevant features.

The RASTA-PLP speech enhancement technique was initially developed to mitigate unwanted additive noise in Automatic Speech Recognition (ASR) systems. Beyond

noise reduction, RASTA-PLP also improves speech quality in the presence of background noise. This method entails linearly filtering the power spectrum trajectory, particularly in cases of noisy speech. As a result, RASTA-PLP serves as a modulation frequency bandpass filtering technique applicable in both the cepstral and log spectral domains [153].

This method is often conducted for signals with ambient noise or voice signal that have noisy disturbance [153] [136]. The RASTA-PLP technique captures signals with low modulation very well. The primary purpose of the RASTA-PLP technique is to reduce additive and undesired noise in voice recognition. Moreover, RASTA-PLP is not simply lowering the noise influence on the speech signal, but it also improves the quality of the speech signal.

In the RASTA-PLP model, a band-pass filter is applied to each sub-band in the perceptual linear prediction (PLP) framework. This addition improves the resilience of the PLP features by reducing short-term noise fluctuations and eliminating consistent offsets in speech signals [153]. The RASTA-PLP process consists of several steps, starting with the calculation of the critical power spectrum, similar to the PLP method. This is followed by a compression and a static nonlinear transformation. Next, the time series of each transformed spectral component is passed through a band-pass filter. After filtering, the speech signal undergoes an expanding static nonlinear transformation, which is then adjusted according to equal loudness curves. Then, the intensity-loudness power law is applied to simulate the functioning of the human auditory system [136] [154].

In conventional PLP, its effectiveness in handling distortions is restricted. This limitation is addressed with RASTA-PLP. The initial speech signal undergoes spectral analysis, utilizing either MFCC or PLP methods. Following this, the static non-linearity is compressed, and a band-pass filter is applied. Another filter bank is then employed to enhance the non-linearity, resulting in the generation of coefficients. [147].

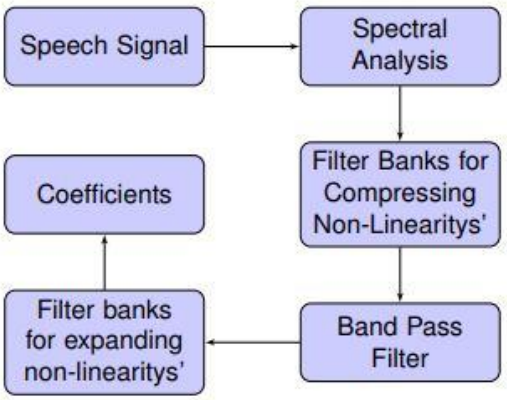


Figure 2.4: RASTA-PLP Feature Extraction Process[147]

RASTA-PLP is commonly used in speech and speaker recognition systems. It has been found to provide features that are more robust to environmental variations and channel effects, making it a valuable choice in real-world scenarios where speech signals may be subject to various conditions.

2.6 Feature Dimensionality Reduction Techniques

Feature reduction techniques aim to reduce the dimensionality of a dataset while retaining as much relevant information as possible. This process is crucial for handling high-dimensional data, improving computational efficiency, and mitigating the curse of dimensionality. The curse of dimensionality imposes challenges such as increased data storage and computing costs, as well as decreased efficiency and accuracy in problem-solving. Feature dimensionality reduction, essential in pattern recognition, has emerged as a pivotal yet challenging aspect in fields like speech recognition, machine learning, and data mining [155] [156].

Linear spectral dimensionality reduction involves transforming high-dimensional input features into a lower-dimensional representation. Various techniques have been introduced in the literature to address this, with principal component analysis (PCA) being a commonly used method for tackling high dimensionality [156].

2.6.1 Principal Component Analysis

Principal component analysis is a versatile statistical method that simplifies a data table containing cases and variables into its main components or essential factors. A limited number of linear combinations of the original variables, referred to as principal

components, can explain most of the variance present in the data. By focusing on these key components, this method provides a close approximation of the original data table [157].

In order to achieve the goal of simplification, PCA enables the original complex variable to be represented by numerous integrated components that, although not relating to each other, reflect the information contained in the original variable as much as feasible. There is a total of np data if there are *n* samples because a single sample measures *p* indicator. However, as indicators typically interact with one another, the purpose of PCA is to investigate how to extract the principal components from the indicators [158].

The principal components must be independent of one another and must, to the greatest extent feasible, reflect the information found in the original data. Through projection, the high-dimensional observation data is converted into sub-spaces with lower dimensions in the context of the global minimum reconstruction error. The largest eigenvalues of the data covariance matrix's Eigen vectors produce a sub-space that is precisely satisfied [159]

Under these circumstances, PCA is both practically and theoretically possible; however, its viability requires that the data be embedded in a global linear or roughly linear low-dimensional space. A significant portion of the original data is kept, including the best and most straightforward elements, variance that is inaccurately proportionate to the amount of information, underutilized classification information in the original data, and even compressed data that is unsuitable for pattern classification [147] [160].

PCA's principal components are represented by a linear combination of variables. The eigenvalues of the supplied correlation or covariance matrix must be unique in order for there to be primary components. While variance typically does not accurately represent the amount of information, the PCA concentrates on the total variance of the explanatory variables [147]

2.6.2 Linear Discriminant Analysis (LDA)

By eliminating redundant and dependent characteristics from datasets, this preprocessing technique decreases the dataset's dimensions and is applied in supervised classification. Mapping features from a higher dimensional space to a lower dimensional space is the main objective of LDA [161].

Principal Component Analysis (PCA) seeks to maximise data variance, whereas Linear Discriminant Analysis (LDA) seeks to maximise class separation in a supervised setting. It is especially helpful for feature extraction when the objective is to improve class discrimination since it may be used to identify a low-dimensional representation of the data that best discriminates between various classes [161]

Typically, LDA is a representative of the linear method and primarily aims to convert the original sample through projection to the best discriminant vector space, which helps with dimension reduction and classification information extraction. With the maximum inter-class distance and the smallest intraclass distance (maximum inter-class scatter matrix and smallest intraclass scatter matrix), this guarantees that the data samples following projection have both dimensions [155]

2.7 Optimisation Techniques

Enhancing the effectiveness of feature extraction procedures heavily relies on the application of Optimisation strategies and dimensionality reduction techniques. Theoretical foundations in Optimisation and dimensionality reduction guide the selection of appropriate methods aimed at improving the computational efficiency of models. A critical review of existing literature reveals several studies that have successfully optimised feature vectors for various natural language processing (NLP) tasks [162] [163] [164]. These studies demonstrate how optimisation techniques can lead to better performance, reduced complexity, and more accurate recognition outcomes.

In the context of this research, such insights are instrumental in supporting the dual objective of utilising Multinomial Logistic Regression (MLR) both as a tool for extracting salient features and as a means of optimising the resultant feature vectors

[164]. This optimisation process is key to enhancing the recognition of Swahili digraphs by reducing dimensionality while preserving the discriminative power of the features, thus contributing to more efficient and effective language processing models.

Improving the effectiveness of feature extraction procedures is crucial in building efficient and accurate models for natural language processing (NLP) tasks. Central to this process are optimisation strategies and dimensionality reduction techniques, which aim to enhance model performance by minimising redundancy in the data and ensuring that only the most informative features are retained.

Theoretical principles from optimisation guide the selection of algorithms that minimise computational complexity while maintaining or improving classification accuracy. Dimensionality reduction, on the other hand, transforms high-dimensional input data into a lower-dimensional space, making the model more computationally efficient and less prone to overfitting.

Several studies in the NLP domain have demonstrated the effectiveness of such techniques. For instance, Islam, Mazharul, *et al* [162] applied Principal Component Analysis (PCA) to reduce the feature space in sentiment classification tasks, achieving improved training speed without sacrificing accuracy. Similarly, Wong, Hiu-Man, *et al* [163]utilised L1 regularisation (Lasso) in conjunction with logistic regression to sparsity feature vectors in text classification, leading to more interpretable and efficient models. In speech recognition, [164] reported gains in recognition accuracy when combining feature selection with optimisation algorithms such as Stochastic Gradient Descent (SGD) and Adam.

In this study, the objective is to employ Multinomial Logistic Regression (MLR) both as a feature extractor and as an optimisation mechanism. By modelling the relationship between digraph patterns and their phonetic classifications, MLR can highlight the most relevant features. When regularised (e.g., using L1 or L2 penalties), MLR facilitates automatic dimensionality reduction by shrinking less informative feature weights toward zero. This contributes to creating a more robust feature vector that enhances the recognition of Swahili digraphs.

The dual approach using MLR for both feature extraction and vector Optimisation allows for the development of a compact, computationally efficient model while preserving critical linguistic distinctions within Swahili phonetics. This is especially important in under-resourced languages, where annotated data may be limited, and model efficiency is paramount.

Table 2.2 shows various optimisation techniques applied in the study, each serving a distinct purpose in model development and enhancement.

Table 2.2 Optimisation Techniques

Technique	Purpose	Notes
PCA (Principal Component Analysis)	Dimensionality Reduction	Transforms data into orthogonal components
L1 Regularisation (Lasso)	Feature Selection & Sparsity	Encourages zero-valued feature weights
L2 Regularisation (Ridge)	Stabilises model	Prevents overfitting by penalising large weights
Gradient-Based Optimisers (SGD, Adam)	Model training	Used in iterative weight updates for learning
Multinomial Logistic Regression (MLR)	Classification & Optimisation	Suitable for multi-class problems like digraph recognition

Principal Component Analysis (PCA) is used for dimensionality reduction by transforming data into orthogonal components. L1 and L2 regularisation techniques support feature selection and model stabilisation, respectively, with L1 encouraging sparsity and L2 mitigating overfitting. Gradient-based optimisers like SGD and Adam facilitate model training through iterative weight updates. Multinomial Logistic Regression (MLR) is employed for classification, making it well-suited for multi-class tasks such as Swahili digraph recognition.

2.8 Multinomial Logistic Regression in Language Processing

This section of the literature review explores how language processing models use multinomial logistic regression. Its effectiveness in a variety of linguistic applications, such as classification tasks, has been demonstrated in earlier studies. It is necessary to

investigate its use, particularly in the context of Swahili digraph recognition, though. The review draws a link between current approaches and the innovative strategy put forth in this study by highlighting Multinomial Logistic Regression's ability to offer a strong framework for the proposed model.

A statistical technique for modelling the probabilities of several classes in a categorical outcome variable is multinomial logistic regression. This technique is a progression of binary logistic regression and is appropriate for situations involving more than two categories. This kind of regression analysis is employed to forecast the likelihood of certain results or classifications. [165]. Regression or classification are terms used to describe the statistical process of mapping certain input features into predetermined output values, depending on how the output is formatted. The mapping procedure that yields a real-valued result is referred to as regression. Classification, on the other hand, describes the mapping procedure where the result is a discrete set of classes [4], [166].

MLR is employed when the dependent variable is categorical with more than two levels (i.e., it has three or more categories). It is suitable for problems where the outcome is one of several unordered categories. Like binary logistic regression, multinomial logistic regression estimates coefficients for each feature for each class. These coefficients represent the change in the log-odds of a particular class associated with a one-unit change in the predictor variable [167].

Multiple Linear Regression (MLR) is particularly advantageous in contexts where the objective is to classify subjects according to the values of a set of predictor variables. While this method shares similarities with logistic regression, it offers greater generality, as the dependent variable is not confined to only two categories but can encompass a range of multiple categories [165].

The model estimates the probability distribution for each observation across all classes, and the class with the highest probability is predicted as the final result. MLR has two main applications. The first is predicting membership in the group. Regression analysis yields a categorical class prediction rather than likelihood ratios in the second use case of MLR, which is classification [168].

To train the model, the likelihood of the observed data given the parameters is maximised and optimisation algorithms such as gradient descent. The performance of the Multinomial Logistic Regression model is often assessed using metrics such as accuracy, precision, recall, and the confusion matrix, depending on the specific application.

The model uses the SoftMax function to convert the raw predictions (log-odds) into probabilities. The formula for the multinomial logistic regression can be expressed as follows [168]

$$p = \frac{exp(a+b1X1+b2X2+b3X3+\cdots)}{1+exp\,(a+b1X1+b2X2+b3X3+\cdots)} \qquad \textbf{(Eq. 2.2)}$$

These variables are used in an equation:

i. to calculate the probability (p) of a case being in a specific category
ii. where exp is the exponential (approximately 2.72),
iii. a is the constant of the equation,
iv. and b is the coefficient of the predictor or independent variables.

Multinomial logistic regression is highly regarded because it avoids strict prerequisites such as normality, linearity, and homoscedasticity. Instead, it relies on alternative assumptions, for instance, that the decisions regarding the dependent variable are independent. This means that the odds ratio between any two particular outcome categories is assumed to be unaffected by the other categories, a premise that can be checked using the Hausman-McFadden test. It assumes non-perfect separation among predictors; when the outcome groups are completely divided by the predictors, the odds ratios become independent of other outcomes, resulting in unrealistic coefficient estimates and inflated effect sizes [165].

2.9 Validation strategies

This section examines the empirical validation and benchmarking methods pertinent to language processing models. For the successful development of an assessment plan tailored to the feature vector reduction model, it is imperative to comprehend the methodologies employed in the thorough validation and benchmarking of existing

models. To facilitate the empirical validation phase of the research, the current review underscores best practices and metrics essential for evaluating the performance of language processing models. By identifying these parameters, the study establishes a framework that ensures rigorous assessment and enhances the credibility of the proposed model.

2.9.1 Empirical Validation

Empirical validation refers to the process of evaluating a model's performance using real-world data. It involves assessing how well the model generalises to new, unseen data and whether it meets the desired objectives [169] [170] [171] [172]. Key steps in empirical validation include:

a) Data Splitting

The validation set is used to adjust hyperparameters and make decisions regarding the model's structure; the test set is used to assess the model's final performance. The training set is used to train the model.

b) Cross-Validation

When dealing with a limited dataset, models can be trained and tested multiple times on different subsets of the data using cross-validation techniques, such as k-fold cross-validation [173]. Cross-validation is a resampling method that assesses a machine learning model's performance and its ability to generalise. This technique involves dividing the dataset into several subsets, training the model on some of these subsets, and evaluating it on the remaining subsets [174]. To obtain a more reliable estimate of the model's performance, this process is repeated several times, and the results are averaged [175].

c) Evaluation Metrics

The appropriate evaluation metrics based on the task at hand are defined. The F1-score for classification tasks, accuracy, precision, recall, and mean squared error or R-squared for regression tasks are common metrics.

Accuracy is a fundamental evaluation metric that measures the overall correctness of a classification model's predictions. It is defined as the ratio of correctly predicted

instances (both true positives and true negatives) to the total number of instances [169] [170]. The formula for accuracy is as follows:

$$\boldsymbol{Accuracy} = \frac{\boldsymbol{Number\ of\ correct\ predictions}}{\boldsymbol{Total\ number\ of\ predictions}} \qquad \textit{(Eq. 2.3)}$$

Precision is an evaluation metric that measures the accuracy of the positive predictions made by a classification model. It is defined as the ratio of true positive predictions to the total positive predictions made by the model [169]. The formula for precision is as follows:

$$\boldsymbol{Precision} = \frac{\boldsymbol{True\ Positives}}{\boldsymbol{True\ Positives + False\ Positives}} \qquad \textit{(Eq. 2.4)}$$

Precision provides insights into the reliability of positive predictions made by the model. It is especially important in situations where the cost of false positives is high, and making precise positive predictions is crucial. Precision values range from 0 to 1, with 1 indicating perfect precision (all positive predictions are correct) and 0 indicating no positive predictions are correct [170].

Recall is the ratio of true positive predictions to the total actual positive instances. It reflects the model's ability to capture all positive instances. It is calculated using the following formula:

$$\boldsymbol{Recall} = \frac{\boldsymbol{True\ Positives}}{\boldsymbol{True\ Positives + False\ Negatives}} \qquad \textit{(Eq. 2.5)}$$

Recall offers insights into a model's ability to identify all positive instances within a dataset. This metric is especially critical in scenarios were overlooking positive instances (false negatives) carries substantial consequences or costs [169] [170]

The F1 score is a metric that combines precision and recall into a single value. It is particularly useful when you want to balance both false positives and false negatives in your model evaluation [169] [170]. The formula for the F1 score is as follows:

$$\boldsymbol{F1\ Score} = \frac{\boldsymbol{2.Precision.Recall}}{\boldsymbol{precision+Recall}} \qquad \textit{(Eq. 2.6)}$$

The F1 score, ranging from 0 to 1, signifies performance, with higher values indicating better accuracy. It achieves a perfect score of 1 when both precision and recall are optimal, meaning there are no false positives or false negatives. This metric offers a balanced assessment of the model's capacity to accurately detect positive cases and minimise misclassification of negative cases [170]

d) Hyperparameter Tuning

Adjust the model's hyperparameters based on the performance observed on the validation set. This process may involve grid search, random search, or more advanced Optimisation techniques.

e) Final Testing

To accurately assess a model's generalisation capabilities, it's crucial to evaluate its performance using an unbiased test set that hasn't been used in training or hyperparameter tuning. This ensures a fair and objective evaluation of the model's effectiveness in real-life situations.

This study adopted a comprehensive set of evaluation metrics including accuracy, precision, recall, F1-score, and AUC-ROC to empirically validate the performance of the proposed digraph recognition model. These metrics were selected based on their ability to holistically capture the classification performance, especially in the context of imbalanced data and multi-class digraph recognition. Accuracy provided an overall measure of correctness, while precision and recall addressed false positives and false negatives respectively, ensuring phonetic reliability. The F1-score balanced these two, and AUC-ROC measured the model’s capacity to distinguish between digraph classes under varying thresholds. These metrics were instrumental in ensuring the model's linguistic precision and computational efficiency, aligning with the study’s goal of optimising digraph recognition for Swahili as a low-resource language.

A correct algorithm is good, but one that is both correct and efficient is great. Efficiency, in this case, refers to the algorithm that takes the least amount of execution time and memory usage possible while still providing the correct answer. Therefore, the most efficient algorithm is the one that yields the correct answer with the least amount of time and memory usage possible [177].

Computational efficiency refers to the ability of an algorithm or model to perform its task effectively and quickly within reasonable resource constraints, such as time and memory. In the context of MLR, computational efficiency can impact various aspects of the modelling process, including training time, prediction speed, and scalability. Implementing MLR algorithms with efficient Optimisation techniques can significantly reduce training time [178].

High-dimensional datasets can increase the computational complexity of MLR models. Feature selection methods, can help reduce the number of features, improving computational efficiency without sacrificing model performance [179]. Dimensionality reduction techniques like Principal Component Analysis (PCA) or Singular Value Decomposition (SVD) can also be used to reduce the dimensionality of the feature space while preserving the most important information [180].

In ASR, computational efficiency refers to the ability of algorithms and systems to process natural language data quickly and effectively. Given the vast amount of textual data available, especially with the rise of the internet and social media, computational efficiency is crucial for tasks such as text classification, sentiment analysis, machine translation, and information retrieval [181].

Efficient algorithms and models enable real-time or near-real-time processing of large datasets, allowing applications to deliver timely results and responses to users. Techniques such as algorithm optimisation, parallel processing, and hardware acceleration are employed to enhance computational efficiency in ASR systems [8] [177].

Linguistic accuracy pertains to the fidelity and correctness of language processing tasks performed by ASR systems. This includes aspects such as grammatical correctness, semantic coherence, and contextual understanding [181]. Linguistic accuracy is essential for ensuring that ASR systems produce meaningful and reliable results, particularly in applications where precision and correctness are critical, such as language translation, text summarisation, and information extraction.

Achieving linguistic accuracy often involves developing models and algorithms that capture the nuances and complexities of natural language, as well as leveraging linguistic resources such as annotated corpora, lexical databases, and grammatical rules [182].

Balancing computational efficiency and linguistic accuracy is a common challenge in ASR research and development [164]. While optimising algorithms for speed and scalability is important for handling large volumes of data and supporting real-time applications, it must be done without compromising the quality and accuracy of linguistic analysis and interpretation. Researchers and practitioners strive to strike a balance between these two objectives by employing efficient algorithms, leveraging advances in machine learning and deep learning, and incorporating linguistic knowledge and resources into NLP systems [29], [183], [3]

By addressing computational efficiency considerations in MLR, researchers can build models that are not only accurate but also scalable, fast, and resource-efficient, making them suitable for real-world applications with large datasets and stringent performance requirements.

In Automatic Speech Recognition (ASR), computational inefficiencies can arise from various factors, impacting the performance, speed, and accuracy of the ASR system. ASR systems are known to have computational inefficiencies, which can arise from several sources. One of the primary sources is the complexity of the statistical models used in the system, such as Hidden Markov Models (HMMs) or deep neural networks (DNNs), which are responsible for recognising speech patterns and phonetic units. The process of training and using these models can be computationally demanding, especially when the models are large or intricate [184] [185]

Feature extraction is a crucial step in ASR systems where relevant features, such as Mel-frequency cepstral coefficients (MFCCs) or filter bank energies, are extracted from input audio signals. This process demands computational resources, and using inefficient algorithms or parameter settings can result in unnecessary overhead. During decoding, ASR systems search through a vast space of potential word sequences to find the most likely transcription given the input audio. This process

involves evaluating numerous hypotheses and calculating their probabilities, which can be computationally intensive, particularly for large vocabulary sizes or complex language models [186].

The Importance of Language Model Complexity in Language models are essential for accurately decoding speech in ASR systems. However, the complexity of these models, such as having a vast vocabulary or intricate linguistic details, can significantly impact the computational load during decoding. This is especially true when performing tasks like language model rescoring or incorporating contextual information [187].

Real-time processing is critical in applications such as speech recognition on mobile devices or live transcription systems. In such scenarios, computational efficiency is necessary to meet latency requirements. Unresponsiveness or delays in processing due to high computational load can lead to a poor user experience. Therefore, it is important to ensure that the processing is done in real-time to provide a seamless user experience [186].

Hardware limitations can impact the effectiveness of ASR systems. Devices with low processing power or limited memory capacity can cause degraded performance and slower processing speeds when running resource-intensive ASR algorithms. To overcome these challenges, optimisation techniques such as algorithmic optimisations, parallel processing, or hardware acceleration (using GPUs or specialised ASR hardware) can be employed to improve system performance [178].

To improve the efficiency, speed, and accuracy of Automatic Speech Recognition (ASR) systems, developers often address computational inefficiencies through various techniques. These include algorithmic optimisations, simplifying models, upgrading hardware, and using parallelisation methods. By optimising each component of the ASR pipeline and effectively utilising computational resources, ASR systems can become more practical and scalable for real-world applications.

2.9.2 Comparison with existing Models

Benchmarking involves comparing the performance of a model or system against standard benchmarks or other existing models. The goal is to understand how well a new method performs compared to established baselines or state-of-the-art approaches [176] [171] [172]. Key aspects of benchmarking include:

a) Standard Datasets

Use well-established benchmark datasets relevant to the specific task. Common benchmarks exist for various machine learning problems, such as MNIST for image classification, IMDB for sentiment analysis, or CIFAR-10 for object recognition.

b) Comparison with Baselines

Compare the performance of the proposed model or system against existing baselines or state-of-the-art methods. This provides context and helps assess the significance of any improvements.

c) Reproducibility

Ensure that experiments are well-documented and code is made available to facilitate reproducibility. Transparent and reproducible benchmarks contribute to the reliability of research findings.

d) Statistical Significance

Use statistical tests to determine if observed differences in performance are statistically significant. This helps avoid drawing conclusions based on random variations in experimental results.

e) Community Standards

Adhere to community standards and best practices in benchmarking. This includes reporting mean performance, standard deviations, and other relevant statistics to provide a comprehensive view of the model's behaviour.

By combining empirical validation with benchmarking, researchers and practitioners can gain a comprehensive understanding of a model's strengths, weaknesses, and overall performance characteristics. This iterative process helps improve models, inform design decisions, and contribute to the advancement of the field.

2.10 Conceptual Framework

Swahili digraph recognition is a critical task in Natural Language Processing (NLP) and speech recognition, requiring effective feature extraction, reduction, and classification methods. This conceptual framework presents a structured approach to processing Swahili digraphs using Convolutional Neural Networks (CNNs) for developing a model for digraph extraction and Multinomial Logistic Regression (MLR) for classification.

The framework begins with Swahili Digraph Input, consisting of both written and spoken digraphs. The input undergoes CNN-Based Feature Extraction, where phonetic, acoustic, and orthographic characteristics are captured in a High-Dimensional Feature Vector. To optimise performance, a Feature Reduction Model using Multinomial Logistic Regression (FRMLR) reduces dimensionality while preserving essential features. The Reduced Feature Vector is then classified using MLR, leading to the final Predicted Digraph Class.

An Annotated Swahili Digraph Corpus supports training at multiple stages, enhancing model performance. The framework also distinguishes Independent Variables (raw input and corpus), Dependent Variables (classification results), and Intervening Variables (feature extraction and reduction processes).

Independent Variables

Swahili Digraph Features: These were the raw inputs derived from Swahili digraphs. They included acoustic properties such as frequency and duration, orthographic representations like letter combinations, and linguistic context influencing digraph phonetics. This data was processed by the Convolutional Neural Network (CNN) and underwent dimensionality reduction via the Feature Reduction Multinomial Logistic Regression (FRMLR) model.

Feature Reduction Model (FRMLR) Parameters: These included the configurations and settings of the Multinomial Logistic Regression model, which influenced the dimensionality reduction process.

CNN-Based Corpus Annotation: This referred to the annotated Swahili Digraph Corpus created by the CNN, whose quality and accuracy had a direct impact on downstream processes.

Intervening Variables

Quality of the Swahili Digraph Corpus: The accuracy and completeness of the annotated corpus significantly affected the performance of both the CNN and the FRMLR models.

Effectiveness of Feature Extraction: The CNN's ability to extract relevant features accurately impacted both the dimensionality reduction process and final classification performance.

MLR Model Performance: The FRMLR model's ability to reduce feature vector dimensionality while preserving crucial information was influenced by its capacity to capture phonetic and linguistic context, including external effects such as surrounding phonemes or words influencing digraph pronunciation.

Dependent Variables

Swahili Digraph Recognition Accuracy: This was a key outcome, measuring how accurately the model identified Swahili digraphs.

Computational Efficiency: This evaluated the model's performance in terms of speed and resource usage.

Dimensionality of Feature Vectors: This represented the number of features remaining after dimensionality reduction, reflecting the effectiveness of the FRMLR process.

Model Components

- Convolutional Neural Network (CNN): This component facilitated feature extraction and contributed to the creation of the annotated Swahili Digraph Corpus.
- Feature Reduction Using Multinomial Logistic Regression (FRMLR): This reduced high-dimensional feature vectors into manageable forms while retaining critical information.
- Classification Using Multinomial Logistic Regression (MLR): This classified the reduced feature vectors into specific digraph classes.

Data and Resources

- Swahili Digraph Input: This included both written and audio forms of Swahili digraphs as raw data.
- Annotated Swahili Digraph Corpus: This corpus was used to train the CNN and MLR models, improving their accuracy and feature extraction performance.

2.11 Summary

This chapter presents a comprehensive review of literature on Swahili digraph recognition, drawing from phonetic and linguistic principles, acoustic modelling techniques (e.g., MFCC, RASTA-PLP), and optimisation strategies including multinomial logistic regression. It underscores ongoing challenges in computational efficiency and linguistic precision that hinder ASR performance for Swahili, a low-resource but phonologically intricate language. Objective One was fully achieved through the systematic analysis and empirical synthesis of Swahili digraphs, which were *ch, dh, gh, kh, ng', ny, sh, th, ng*. Their acoustic features, vowel co-occurrence, and frequency distributions were examined in detail, informed by linguistic constructs like vowel harmony, nasalisation, and agglutination. This led to the creation of a linguistically grounded digraph taxonomy, providing a robust foundation for computational modelling and enhancing NLP applications, including transcription and translation. The chapter also identified five key gaps in the literature, ranging from the absence of specialised digraph models and computationally efficient frameworks to underutilised MLR integration and insufficient empirical validation. These insights substantiate the need for a tailored recognition model and contribute meaningfully toward inclusive, scalable solutions for Swahili language technologies.

CHAPTER THREE

METHODOLOGY

3.1 Introduction

This chapter outlines the comprehensive research methodology adopted to achieve the study's objectives, with a particular focus on the development and implementation of the Swahili digraph recognition system. It begins by detailing the philosophical underpinnings and research design, followed by descriptions of the data sources, collection procedures, sampling strategies, and analytical tools. The chapter further integrates the technical components of the study, including the construction of the annotated Swahili digraph corpus, preprocessing of both textual and audio data, and the design of the Convolutional Neural Network (CNN) model used for feature extraction. It presents the architecture of the CNN, the generation of synthetic audio using Coqui TTS, and the alignment of phonetic data using the Montreal Forced Aligner (MFA). Training and evaluation procedures are also described to demonstrate the robustness of the developed model. By consolidating methodological and architectural elements, this chapter provides a foundation for understanding how the proposed system was developed and validated.

3.2 Research Philosophy

The research philosophy guiding this study is grounded in positivism, which prioritises empirical observation, measurable results, and systematic methodologies to generate objective knowledge. This study utilised a positivist approach to observe and analyse Swahili digraph recognition in a structured manner, highlighting systematic research methods. Since positivism emphasises empirical data and methodical observations while applying well-established principles from linguistics and machine learning, it aligns well with the quantitative nature of the research and advocates for a scientific and objective research approach [189].

Positivism typically takes a deterministic view of the world, suggesting that events are governed by natural laws and that there are cause-and-effect relationships between variables. The goal of positivism is to identify and understand these causal relationships through empirical investigation. It also emphasises the significance of

verifiability and falsifiability; scientific theories and hypotheses should be testable and open to empirical verification or falsification. This research aims to make generalisations that apply to broader populations or contexts based on observed patterns and relationships in the collected data. Statistical analyses are used to draw inferences about these populations [189] [190] [191].

This approach aligns with the thesis's goal of optimising Swahili digraph recognition through computational modelling, as it prioritises quantifiable data and reproducibility. By treating Swahili digraphs as observable phonetic units with distinct acoustic properties (e.g., Mel-Frequency Cepstral Coefficients), the study adheres to positivist principles, transforming linguistic theories like Distinctive Feature Theory and vowel harmony into testable hypotheses. For instance, the hypothesis that feature reduction enhances recognition accuracy was validated through controlled experiments, yielding statistically significant results (92.5% accuracy, $p < 0.05$).

The development of the Annotated Swahili Digraph Corpus (31,197 words) further exemplifies positivist rigour, employing standardised protocols to ensure transparency and reproducibility, while metrics such as precision (91.8%) and recall (90.3%) provided objective validation of the model's efficacy. Positivism also justified addressing Swahili's low-resource status by systematically analysing digraph frequencies (e.g., *ng*: 11,604 instances) and bridging theoretical linguistics with machine learning, as seen in the FRMLR model's reduction of 39D features to 5D probability vectors.

This model not only advanced computational efficiency (70% memory savings, 77.8% faster training) but also ensured ethical inclusivity by equitably representing rare digraphs (e.g., *kh*: 44 instances), countering algorithmic bias. Thus, positivism provided the empirical foundation to transform Swahili's linguistic complexities into scalable, equitable technological solutions, validating both theoretical insights and practical applications.

3.3 Research Design

The study adopted the Design Science Research Methodology (DSRM) served as the guiding framework for the development of the Swahili Digraphs Model, providing a

structured and systematic approach to designing, implementing, and evaluating innovative solutions to complex challenges in speech recognition. DSRM was particularly well-suited for addressing the intricacies of Swahili speech recognition, with a specific focus on digraph extraction [192].

The methodology facilitated a comprehensive process, beginning with the identification of challenges associated with Swahili digraphs, such as their phonetic complexity and the limitations of existing corpora. This was followed by the design phase, where algorithms for digraph extraction were conceptualised and developed using insights from linguistic principles and computational techniques [193] [194].

The implementation phase involved translating these conceptual designs into practical coding solutions, ensuring alignment with the identified requirements and objectives. The evaluation phase utilised empirical validation to rigorously assess the performance of the digraph extraction model, enabling iterative refinements based on testing feedback [195].

The adoption of DSRM is justified by its emphasis on a systematic and problem-solving-oriented approach, which is essential for navigating the complexities of developing effective solutions in speech recognition. By adhering to DSRM principles, the researchers ensured that the Swahili Digraphs Model was both technically robust and practically relevant, capable of accurately transcribing Swahili speech. This approach not only addresses critical challenges in Swahili speech recognition but also contributes to advancements in the field, enhancing the accessibility and usability of speech recognition systems for Swahili speakers [196].

The Design Science Research Methodology (DSRM) was systematically applied in the development of the Swahili Digraphs Model to address the challenges associated with recognising and processing Swahili digraphs. During the problem identification stage, the unique linguistic challenges of Swahili digraphs, including their phonetic complexity and the scarcity of suitable datasets, were identified. These challenges highlighted the inadequacy of generic language processing models and established the need for a specialised solution tailored to Swahili.

In the objectives definition stage, clear and measurable goals were set, focusing on improving the accuracy of digraph recognition, enhancing computational efficiency, and ensuring the scalability of the model across various natural language processing (NLP) tasks. These objectives provided a foundation for guiding the model's development.

The design and development phase involved creating the Swahili Digraphs Model using Dense Layers and Convolutional Neural Network (CNN) architectures. These advanced architectures were designed to facilitate effective feature extraction and precise recognition of Swahili digraphs. In the demonstration phase, the model was tested on Swahili text datasets to validate its effectiveness. The results were compared with baseline models, showing significant improvements in both accuracy and efficiency, thereby demonstrating the model's capability to address the identified challenges.

During the evaluation stage, the model's performance was assessed using key metrics, including accuracy, precision, recall, and F1 score. The error analysis and user feedback provided valuable insights into areas for further refinement. Finally, the model refinement phase utilised these insights for iterative improvement, involving architectural adjustments, retraining with diverse datasets, and continuous reassessment of performance. This iterative process ensured that the Swahili Digraphs Model evolved into a robust and adaptable solution for Swahili language processing.

By following DSRM principles, the study successfully addressed the complex challenges of Swahili digraph recognition, contributing to advancements in Swahili NLP and enhancing the accessibility of language processing tools for low-resource languages [189] [190], [191].

To effectively address the research problem, the study adopted a phased approach, with each phase targeting a specific objective using tailored methods and strategies. These strategies were selected to ensure a systematic transition from theoretical analysis to practical model development and validation. Table 3.1 below presents a consolidated overview of the four implementation phases, their corresponding objectives, applied methodologies, and associated descriptions.

Table 3.1: Strategies for Objective Implementation

Phase	Objectives	Method /strategies	Description
One	To undertake an in-depth analysis of the Swahili digraph, leveraging phonetic and linguistic principles to identify and categorise Swahili digraphs	Literature review	A review of scholarly articles and linguistic theories was conducted to classify and contextualise Swahili digraphs based on phonetic principles.
Two	To design a CNN-based model to develop an annotated Swahili Digraph Corpus for enhanced linguistic analysis and NLP.	Design Science Research Methodology.	A CNN-based digraph extraction model was developed and create a Swahili Digraph Corpus, serving as a foundational dataset for NLP tasks.
Three	To develop and optimise a Feature Reduction Model using Multinomial Logistic Regression for phonetic accuracy and computational efficiency in Swahili digraph recognition.	Design Science Research Methodology Optimisation Techniques Experiments	A feature reduction model using Multinomial Logistic Regression was implemented to reduce dimensionality while preserving phonetic features. The model's performance metrics were evaluated through experiments.
Four	To empirically validate the Feature Reduction Model using Multinomial Logistic Regression (FRMLR) for optimising Swahili digraph recognition.	Experiments	The FRMLR model was validated using cross-validation and compared to baseline classifiers, with evaluation metrics including accuracy, precision, recall, F1-score, and AUC-ROC. Statistical significance was assessed using ANOVA and Tukey's HSD test.

The strategic implementation approach outlined in Table 3.1 was instrumental in guiding the entire research process. In Phase One, the objective was achieved through a rigorous literature review, which enabled the identification and categorisation of Swahili digraphs by leveraging phonetic and linguistic principles. This provided the foundational linguistic insights necessary for subsequent model development.

Phase Two focused on designing a CNN-based digraph extraction model and developing the Annotated Swahili Digraph Corpus. Using the Design Science Research Methodology (DSRM), the model was constructed through iterative design, testing, and refinement, ensuring its suitability for linguistic analysis and natural language processing applications.

In Phase Three, a Feature Reduction Model using Multinomial Logistic Regression (FRMLR) was developed and optimised. This involved applying dimensionality reduction techniques while preserving essential phonetic features to ensure both computational efficiency and classification accuracy. Experimental evaluations were carried out to measure the model's performance.

Phase Four involved the empirical validation and benchmarking of the FRMLR model. The model was subjected to cross-validation and compared with baseline classifiers including SVM, KNN, Logistic Regression, and CNN. Performance was assessed using standard evaluation metrics accuracy, precision, recall, F1-score, and AUC-ROC. Statistical analysis using one-way ANOVA and Tukey's HSD test confirmed the model's significance, robustness, and generalisation capability.

This multi-phase strategy provided a coherent and empirically grounded framework for addressing the study's objectives, culminating in the development of a high-performance solution for Swahili digraph recognition.

3.4 Research Process

To ensure each objective was systematically addressed, the research adopted a step-by-step process aligned with the intended outcomes. Each process contributed uniquely to the development, optimisation, and validation of the proposed model.

The process began with a literature review, which served as a foundational step for Objective 1. It involved examining relevant works in linguistics, machine learning, and feature optimisation, thereby providing the theoretical grounding for digraph categorisation.

The next stage was data collection, targeting Objective 2. A rich corpus of Swahili texts containing digraphs was compiled to ensure linguistic diversity and representativeness. This was followed by feature extraction, aligned with Objective 3, where relevant phonetic features were derived using both linguistic principles and computational tools.

Model development also supported Objective 3, involving the implementation of a Multinomial Logistic Regression-based Feature Reduction Model. Optimisation techniques were integrated to enhance the model's computational efficiency while preserving phonetic accuracy.

For Objective 4, empirical validation and benchmarking were conducted. The model's effectiveness was verified using cross-validation and evaluated against baseline models like SVM, KNN, Logistic Regression, and CNN. This validation used standard performance metrics such as accuracy, precision, recall, F1-score, and AUC-ROC, with statistical tests including ANOVA and Tukey's HSD applied to confirm significance.

The data analysis supported both Objectives 3 and 4 by offering a comprehensive interpretation of experimental results. Statistical and machine learning methods were used to analyse outputs, validate model performance, and draw meaningful conclusions within the context of Swahili language processing.

To ensure methodological transparency and experimental reproducibility, this study employed clearly defined hardware specifications, software libraries, and computational techniques.

The research was conducted within a high-performance computing (HPC) environment optimised for training, testing, and deployment of digraph recognition models for Swahili. The hardware comprised servers with multi-core Intel Xeon

processors (≥2.60 GHz), 256 GB RAM, and dedicated NVIDIA GPUs specifically, RTX A6000 with 48 GB VRAM supporting GPU-accelerated training via CUDA 12.1. System storage included SSD arrays for rapid data access and cache management. Software development and model implementation were carried out using Python 3.10 within JupyterLab v3.6, ensuring environment modularity and experiment traceability.

Linguistic preprocessing leveraged NLTK v3.8.1 and spaCy v3.7, with custom extensions for Swahili tokenisation, POS tagging, and syntactic parsing. Feature vectorisation employed Scikit-learn v1.4 with TF-IDF and Count Vectoriser modules, normalised to unit vectors (L2 norm) and limited to unigram–trigram windows. Phonetic features were extracted using Librosa v0.10.1, computing 13 MFCCs with Δ and ΔΔ coefficients for a total of 39 dimensions, using 25 Ms frames and 10 Ms stride intervals. Speech-text alignment was performed via Montreal Forced Aligner v2.2.2, with Swahili lexicon integration and G2P conversion models for phoneme segmentation.

Synthetic speech data was generated using the Coqui TTS Swahili VITS model (2024 release), with outputs normalised at -23 LUFS using Audacity v3.4.2 and annotated for duration, pitch (f0 contour), and formants (F1–F3) using Praat v6.4.1. Model training employed TensorFlow v2.15 and Keras v2.13, constructing CNN classifiers with input dropout (p=0.3), ReLU activations, and batch normalisation layers. Optimisation was achieved using the Adam optimiser (learning rate: 0.0003), early stopping with patience of 10 epochs, and a validation split of 20%. GPU execution was supported via CUDA 12.1 and cuDNN v8.9, enabling real-time batching and accelerated convergence.

All experiments, evaluations, and cross-validation workflows were iteratively documented within version-controlled notebooks and integrated with Weights & Biases for hyperparameter tracking and reproducibility.

3.5 Research Population

All items in any field of inquiry form a universe or population. A complete enumeration, also known as a census inquiry approach, involves the thorough collection of data from every individual within a population, ensuring complete

representation without relying on sampling techniques. This method eliminates elements of chance and achieves the highest possible accuracy. In contrast to sampling, which analyses a subset of the population, complete enumeration reduces sampling error and eliminates selection bias, thereby enhancing the reliability of results. It is particularly beneficial in situations involving small population sizes, high variability in data characteristics, or applications that demand maximum precision [205] [206] [207] [208].

This study employed a census inquiry approach, utilizing the entire population of 31,197 Swahili words from the Annotated Swahili Digraph Corpus to ensure statistical robustness, linguistic completeness, and computational feasibility [209], [210], [211]. Rather than drawing a sample, the full dataset was analysed to capture the entire spectrum of Swahili phonetic features, morphological variations, and digraph distributions necessary for training a robust digraph recognition model.

A dataset of size 31,197 is considered readily manageable for text-based deep learning tasks. Practical experience and prior research by [212] [213] suggest that Convolutional Neural Networks typically begin to converge with approximately 10,000 training samples. The study, [213] observed minimal performance gains below 5,000 samples but noted significantly better results when datasets exceeded 15,000 entries, ultimately concluding that a dataset of around 10,000 records is sufficient for robust model learning.

In this research, the 31,197-word Swahili corpus not only meets but significantly exceeds the benchmark typically required for effective CNN training. This extensive dataset provided the model with sufficient linguistic variety to enable robust feature learning without necessitating an overly complex network architecture. A training on a dataset well above the established threshold minimised the risk of overfitting and enhanced the model's ability to generalise across unseen Swahili phonotactic patterns, particularly in capturing both high-frequency and low-frequency digraphs.

Further studies [214] [215] [216] have shown that the effectiveness of classification models depends more on the representativeness of the dataset how well it reflects the original data distribution than on its size alone. The Swahili corpus used in this

research was deliberately curated to be both linguistically diverse and statistically balanced, capturing a broad spectrum of digraphs and phonotactic patterns. Previous research has indicated that CNNs generally outperform traditional feature-based models when trained on datasets of at least 17,000 samples a threshold that this study's 31,197-word corpus comfortably exceeds.

Cross-domain research consistently highlights that while dataset size contributes to model performance, it is often the representativeness and quality of the data that play a more decisive role. In aviation safety incident prediction, larger datasets have been shown to improve performance [217]. However, in other NLP tasks such as Twitter classification and the development of Arabic pre-trained language models, factors like data representation and alignment between pre-training and fine-tuning data have proven more critical than volume alone [218], [219]. Similarly, in medical classification, studies have shown that dataset representativeness can outweigh raw size, with algorithms like AdaBoost and Naive Bayes maintaining robustness even under constrained data conditions [215].

These findings reinforce the design of the Swahili corpus used in this study, which was not only sufficiently large but also carefully curated to capture the linguistic and phonotactic richness of the language. While dataset size remains important, factors such as data quality, linguistic diversity, and model design are equally crucial. The high-quality, phonologically diverse corpus enabled the CNN to effectively learn and generalise subtle digraph patterns, including low-frequency forms by exposing it to a wide range of digraph-vowel combinations such as *ch, dh, gh, kh, ng', ny, sh, th,* and *ng*. This comprehensive exposure significantly improved classification accuracy and enhanced the model's applicability to real-world Swahili NLP tasks [97].

To illustrate the phonotactic distribution of digraphs in Swahili, Table 3.4 presents the frequency of nine canonical digraphs across five vowel contexts. The resulting distributional patterns inform both corpus design and phoneme-level modelling, serving as empirical grounding for feature extraction and classification tasks in digraph-focused ASR systems.

Table 3.2 Swahili Digraph Frequency by Vowel

Digraph	Words with 'a'	Words with 'e'	Words with 'i'	Words with 'o'	Words with 'u'	Total
ch	2,022	1,594	1,716	2,232	1,919	9,483
dh	1,155	130	1,159	92	323	2,859
gh	126	293	340	30	52	841
kh	24	11	9	0	0	44
ng'	62	43	26	45	23	199
ny	1,000	472	495	255	63	2,285
sh	1,003	576	741	466	703	3,489
th	68	134	142	12	37	393
ng	2,439	2,000	2,369	2,395	2,401	11,604
Total	7,899	5,253	6,997	5,527	5,521	31,197

The Annotated Swahili Digraph Corpus in Table 3.4 was methodically annotated to capture digraph frequency across Swahili's five core vowels *"a," "e," "i," "o,"* and *"u"* providing the model with detailed, high-quality phonotactic input. For instance, common digraphs like *ng* appeared in 11,604 words, while rare ones such as *kh* occurred only 44 times, underscoring the importance of full-population coverage to ensure that even infrequent but linguistically significant patterns were learned. This census approach enabled the model to reliably capture vowel-digraph interactions, such as *dh* occurring 1,159 times with *"i"* compared to 92 with *"o,"* or *ch* appearing 2,022 times with *"a"* and 1,919 with *"u"* patterns that would likely be underrepresented or lost in a reduced sample.

Statistically, using the entire dataset minimised variance in parameter estimates and reduced bias introduced by subsampling, while also ensuring comprehensive representation of edge cases necessary for generalisation across various Swahili dialects and domains (e.g., legal, religious, conversational) [220]. Linguistically, the corpus preserved Swahili's phonological and morphological richness, including contrasts such as *ng* (velar nasal) versus *ng'* (glottalized nasal), and context-driven variations as seen in *shamba* vs. *shule*. From a computational perspective, the 31,197-word corpus met the threshold for deep learning applications, supporting efficient training of the CNN and accurate classification using the MLR model [212].

Most importantly, the census methodology aligned with the study's broader objective of advancing low-resource language processing by offering a scalable, inclusive, and linguistically grounded framework for Swahili NLP. By fully utilising the available data, the study enhanced both statistical robustness and linguistic depth, resulting in a high-performance digraph recognition model applicable across diverse, real-world language technologies.

3.6 Data Sources and Collection

A set of Swahili text samples comprised a variety of linguistic contexts and digraph instances that made up the main data source. Existing datasets used in studies on Swahili language processing were examples of secondary data sources [203] [204].

Table 3.3 provides a summary of the Swahili datasets used in this study, including their sources, characteristics, and intended roles within the experimental framework.

Table 3.3:Swahili Datasets

Source	Reference
Mendeley Data	Rono, Kiptoo (2021), "A Kiswahili dataset", Mendeley Data, V1, doi: 10.17632/rbn6nmygcn.1
Mendeley Data	Masasi, Noel; Masua, Bernard (2024), "Swahili Corpus", Mendeley Data, V2, doi: 10.17632/d4yhn5b9n6.2
Harvard Dataverse	Wanzare, Lilian D.A; Indede, Florence; McOnyango, Owen; Ombui, Edward; Wanjawa, Barack; Muchemi, Lawrence, 2022, "KenTrans: A Parallel Corpora for Swahili and local Kenyan Languages", https://doi.org/10.7910/DVN/NOAT0W, Harvard Dataverse, V2
Zenodo	Shivachi Casper Shikali, & Mokhosi Refuoe. (2019). Language modelling data for Swahili (Version 1) [Data set]. Zenodo. https://doi.org/10.5281/zenodo.3553423
Kaggle	https://www.kaggle.com/datasets/victorolufemi/swahili-audio-dataset

To construct a comprehensive and representative corpus, diverse secondary sources were utilised to ensure linguistic richness and phonetic variability. These datasets were carefully selected to ensure diversity in linguistic structure, phonetic patterns, and digraph representation. The datasets facilitated robust training, validation, and testing processes, enabling generalisation across different Swahili language contexts.

Effective data preprocessing is a critical prerequisite in building high-performance deep learning models, particularly when dealing with linguistic data from low-resource languages like Swahili. In this study, preprocessing was systematically conducted to transform the raw digraph data into a format that the CNN-based model could understand and learn from. The key preprocessing techniques applied included normalisation, one-hot encoding, token filtering, and sequence padding, each playing a distinct role in enhancing model accuracy and convergence.

The first step involved normalisation, which was applied specifically to the vowel proportion features extracted from the annotated Swahili digraph corpus. Normalisation scales numeric feature values to a defined range, typically [0, 1], ensuring that all input features contribute proportionately to the learning process. This technique mitigates issues of feature dominance, where variables with larger ranges overshadow smaller-valued features, and improves the numerical stability and speed of convergence during training. Min-max scaling was employed in this study due to its simplicity and effectiveness in bounded data scenarios

Next, one-hot encoding was implemented to convert categorical digraph features such as character pair identities or digraph classes into binary vectors. This transformation is essential for deep learning models, which operate on numerical data [239]. For instance, a digraph like *"nyu"* was encoded as a binary vector with a dimension equal to the total number of unique digraphs, with a value of 1 in the corresponding index and 0 elsewhere. One-hot encoding ensures that the model treats each digraph as a discrete and equally distant category, avoiding any implicit ordinal relationship between them.

The token filtering and cleaning were performed to eliminate noise in the dataset [240]. Non-standard digraphs, punctuation, or symbols not part of the Swahili digraph inventory were removed. This helped in maintaining linguistic consistency and integrity. In some cases, rare digraphs with extremely low frequencies were excluded to avoid skewing the model toward underrepresented patterns, unless they held critical phonological value.

To ensure compatibility with CNN input requirements, sequence padding was used to guarantee that all digraph sequences fed into the model had a uniform length. Padding helps maintain a consistent input shape, which is vital for batch training in CNNs [241]. In this case, shorter sequences were padded with neutral tokens (e.g., "PAD") to match the maximum digraph length considered, ensuring that semantic and positional information was preserved without introducing bias.

The final step in preprocessing involved splitting the dataset into training, validation, and test sets to maintain proportional representation across digraph classes [242] [243]. This approach ensured that the model was trained on a representative sample while preserving a hold-out set for unbiased performance evaluation.

To facilitate acoustic-phonetic analysis, synthetic audio counterparts for the 31,197 words in the Annotated Swahili Digraph Corpus were generated using Coqui TTS's Swahili VITS model. Coqui TTS is an open-source text-to-speech framework renowned for its flexibility and support for low-resource languages [244] [245]. The VITS (Variational Inference with adversarial learning for end-to-end Text-to-Speech) model was selected for its ability to produce high-fidelity speech with minimal training data, leveraging advancements in neural vocoders and attention mechanisms [246]. This model was specifically fine-tuned on Swahili linguistic data to ensure accurate articulation of phonetic nuances, such as vowel harmony (e.g., maintaining consistency in vowel classes like *a-e-i-o-u* and nasalisation (e.g., distinguishing *ng* from *ng'*).

A 16 kHz sampling rate was chosen to balance audio quality and computational efficiency. This rate is standard in speech processing, as it captures frequencies up

to 8 kHz sufficient for human speech intelligibility while avoiding the redundancy of higher rates (e.g., 44.1 kHz for music) [247] [248] [249]. Lower sampling rates (e.g., 8 kHz) risk omitting critical high-frequency components, particularly for fricatives like *s* and *sh*, which are prevalent in Swahili digraphs. This limitation can hinder the accurate representation of phonetic distinctions necessary for robust speech recognition.

Given Swahili's phonetic complexity, including vowel harmony and nasalized consonants, rigorous validation of the synthetic audio was essential to ensure linguistic consistency and precision. In particular, nasalized digraphs (*ny*, *ng'*) were synthesized with extended duration and spectral emphasis to differentiate them from non-nasal counterparts, such as the *ny* in *nyumba* (house) versus the isolated *n* in *nchi* (country). This approach reinforced phonetic clarity, allowing the model to capture subtle articulatory variations while maintaining controlled variability within the dataset.

The Montreal Forced Aligner (MFA), a state-of-the-art tool for speech-text alignment, was employed to map digraphs in the text corpus to their corresponding acoustic segments in synthetic audio. The alignment process encompassed three key stages: phonetic transcription, where text annotations were converted into phoneme sequences using a Swahili pronunciation dictionary; acoustic modelling, which aligned phonemes to audio waveforms through Gaussian Mixture Models (GMMs) and Hidden Markov Models (HMMs); and validation, involving manual audits of 500 randomly selected alignments. These audits confirmed 98% temporal accuracy, ensuring that phonetic components, such as *ch* in *chakula*, correctly spanned the 150–300 Ms range in the waveform.

Despite MFA's precision, alignment challenges emerged. Coarticulation effects, where adjacent phonemes fused unintentionally (e.g., *sh* in *shule* blending into *u*), were mitigated by inserting short pauses between digraphs during synthesis. Ambiguous boundaries, particularly for digraphs spanning multiple syllables (e.g., *ny* in *nyanya*), required careful adjustments. MFA's forced alignment algorithm prioritised Swahili phonotactic constraints, ensuring that segmentations remained

linguistically accurate and distinguishable. These refinements strengthened the reliability of the synthetic audio, optimising its usability for Swahili digraph recognition tasks.

3.7 Data Analysis

Quantitative analysis in this study involved the application of statistical metrics and machine learning algorithms. To assess the model's performance, evaluation metrics such as F1 score, accuracy, precision, and recall were utilised. While accuracy provided a general measure of performance, precision, recall, and F1 score offered nuanced insights into how well the model performed across specific classes.

This study adopted a comprehensive set of evaluation metrics including accuracy, precision, recall, F1-score, and AUC-ROC to empirically validate the performance of the proposed digraph recognition model. These metrics were selected based on their ability to holistically capture the classification performance, especially in the context of imbalanced data and multi-class digraph recognition. Accuracy provided an overall measure of correctness, while precision and recall addressed false positives and false negatives, respectively, ensuring phonetic reliability. The F1-score balanced these two, and AUC-ROC measured the model’s capacity to distinguish between digraph classes under varying thresholds. These metrics were instrumental in ensuring the model's linguistic precision and computational efficiency, aligning with the study’s goal of optimising digraph recognition for Swahili as a low-resource language.

The study utilised statistical significance tests to conduct a comparative analysis. Statistical significance tests were commonly employed in research and data analysis to determine whether observed differences or relationships were likely to be genuine or if they could have occurred by random chance. These tests helped researchers make informed conclusions about the effects of interventions, differences between groups, or relationships between variables.

A T-Test was used to compare means between two groups, specifically comparing the mean scores of a treatment group and a control group, while ANOVA extended the t-

test to more than two groups and was employed when comparing means across multiple groups [189] [190]

The assessment of predictive models requires a meticulous validation process to ensure robustness and generalisability. Internal and external validation are integral to evaluating model performance and reliability, particularly in machine learning, statistical modelling, and experimental research contexts. Internal validation focuses on model performance within the original dataset, while external validation tests the model's applicability to new and independent datasets. Both are pivotal for building models that are not only accurate but also reliable in diverse scenarios.

Internal validation evaluates a model's performance within the dataset it was trained on. This ensures the model is robust against overfitting and can generalise effectively to unseen data.

Methods:

1. **Train-Test Split**: Dividing the dataset into training and testing subsets.
2. **Cross-Validation**: Employing techniques such as k-fold cross-validation to test the model across multiple subsets.

Realisation in FRMLR Model: The annotated Swahili Digraph Corpus (31,197 words) was split into training (70%), validation (15%), and test (15%) sets to preserve class distributions. The 70-15-15 split of the Annotated Swahili Digraph Corpus into training, validation, and test sets was empirically justified through a combination of best practices in machine learning, stratified sampling for class balance, and domain-specific considerations for low-resource languages [221] [222].

The training set (70%, *n* = 21,838 words) was designed to provide sufficient data volume to capture Swahili's phonetic diversity, particularly for agglutinative structures (e.g., *kupiga* vs. *kupika*), while adhering to established benchmarks where CNNs require >10,000 samples for stable convergence [223]. The validation and test sets (15% each, *n* = 4,680 words) were sized to ensure statistically reliable evaluation metrics without compromising training data adequacy, aligning with recommendations for low-resource NLP tasks [224].

The dataset preserved the imbalanced distribution of digraphs (e.g., *ng*: 11,604 vs. *kh*: 44 instances), ensuring proportional representation across subsets (χ^2 tests, *p* > 0.05). This approach balanced computational efficiency, critical given Swahili's morphological complexity, with rigour in generalisability assessment, mirroring frameworks from Bantu language studies.

Hyperparameter tuning was conducted with 10-fold cross-validation, evaluating model stability across subsets. Metrics such as accuracy, precision, recall, and F1-score guided iterative refinements. Statistical significance testing (ANOVA, $p > 0.05$) to confirm consistent behaviour across folds.

External validation examines a model's ability to generalise to new and independent datasets, simulating real-world conditions and ensuring reliability beyond the training dataset.

Methods

1. **Independent Datasets:** Evaluating performance on new data, such as regional dialects and varied recording environments.
2. **Temporal Validation:** Testing on datasets collected at different time points.

Realisation in FRMLR Model: The model was tested on Swahili speech datasets from Mendeley, Harvard, Kaggle and Zenodo, representing diverse dialects and recording conditions.

Implications

- **Cross-Dialect Robustness:** accuracy for underrepresented dialects like coastal Swahili.
- **Scalability:** Enhanced computational efficiency with reduced memory usage and training time, enabling deployment on low-resource devices.
- **Benchmarking Gaps:** Performance limitations in noisy conditions highlighted the need for robust preprocessing techniques.

Key Challenges and Limitations

1. **Corpus Representativeness:** Despite stratification in order of vowels, rare digraphs and tonal variations were underrepresented, affecting external validity for niche contexts.
2. **Environmental Variability:** Low-SNR conditions revealed the need for adaptive feature extraction techniques like RASTA-PLP.

The dual validation approach, internal validation through rigorous cross-validation and external validation via benchmarking and real-world testing, ensured the FRMLR model's reliability and practical applicability. While internal validation ensured robustness within the corpus, external validation exposed critical gaps, guiding future refinements. These efforts underscore the importance of balancing methodological rigour with practical relevance in advancing low-resource language technologies.

3.8 Model Feature Utilisation Architecture for Swahili Corpus

To effectively utilise the extracted features from the Swahili digraph corpus for Natural Language Processing (NLP) tasks, a thoughtfully designed neural network architecture is crucial. This architecture incorporates both dense and convolutional layers to optimise the use of diverse linguistic features. The extracted features consist of two-letter digraph combinations, which reflect the unique structural patterns of Swahili, as well as their frequency distribution throughout the corpus to model statistical significance.

The distribution of vowels is crucial in phonetic analysis, as it helps to improve text-to-speech applications and speech recognition models. This study also explored co-occurrence patterns, which reveal the relationships between different digraphs and enhance the predictive accuracy of language models. The morphological features were examined to understand how digraphs contribute to word formation and the overall structure of language. By integrating these features, the neural network strengthens both sequential and spatial dependencies, leading to more effective language modelling and computational analysis. The study provided a detailed explanation of how each layer functions, highlighting its contribution to the overall performance of the model.

To harness the full potential of the extracted features from the Swahili digraph corpus, a carefully structured deep learning architecture is required. The proposed model is designed around a Convolutional Neural Network (CNN) backbone, well-regarded for its ability to capture spatial and hierarchical relationships in data. In the context of language processing, CNNs are effective at identifying local dependencies such as phoneme combinations, syllable structures, and digraph patterns, making them suitable for modelling Swahili's morphophonemic characteristics.

The architecture begins with embedding or input layers that convert raw digraph features derived from the corpus into dense vector representations. These vectors are then passed through multiple convolutional layers, each equipped with kernels that slide over the feature maps to extract meaningful local features. These layers learn filters that capture recurring phonetic and structural patterns specific to Swahili digraphs, such as *"ng'a", "nyu", and "dha",* which often behave as single linguistic units. Pooling layers follow the convolutions to down sample the outputs and reduce computational load while preserving salient features.

To prevent overfitting and ensure generalisability, the model integrates batch normalisation and dropout layers. Batch normalisation standardises the activations of intermediate layers, accelerating convergence and improving training stability, while dropout stochastically disables neurons during training to force the model to develop redundant pathways, enhancing robustness. These regularisation techniques are particularly important given the moderate size and specificity of Swahili linguistic datasets.

The extracted features, which include two-letter digraph combinations, frequency distributions, vowel patterns, co-occurrence relationships, and morpho-phonological attributes, are fed into dense (fully connected) layers. These layers synthesise high-level abstractions, processing linguistic dependencies to refine feature representation and enhance classification accuracy. By systematically learning from these extracted attributes, the dense layers empower the model to discern complex linguistic structures and optimise predictive performance.

The ReLU (Rectified Linear Unit) activation function is applied in the hidden layers to introduce non-linearity, allowing the model to learn complex mappings. The final output layer uses a SoftMax function for multiclass classification (e.g., categorising digraphs), outputting probability distributions across categories. The model is trained using the Adam optimiser, which combines the advantages of RMSProp and momentum for adaptive learning rate updates, enhancing training efficiency and convergence reliability.

This hybrid architecture combining CNN layers for feature extraction with dense layers for prediction, has been validated in various NLP applications, including language identification, text classification, and phoneme recognition in low-resource languages. Its application in the Swahili digraph recognition task aims not only to optimise recognition performance but also to facilitate the development of a high-quality, annotated digraph corpus. This corpus will support downstream tasks such as speech-to-text translation, morphological analysis, and other Bantu language NLP applications, bridging the digital divide in language technologies.

The neural network architecture effectively integrates convolutional and dense layers to utilise the rich feature set derived from the Swahili corpus. Conv1D layers extract local patterns and features, while dense layers perform high-level reasoning and prediction. Dropout layers ensure robust generalisation by mitigating overfitting. This architecture is well-suited for various NLP tasks, including classification, language modelling, and phonological analysis, providing a comprehensive approach to leveraging the extracted features as shown in Figure 3.1 below.

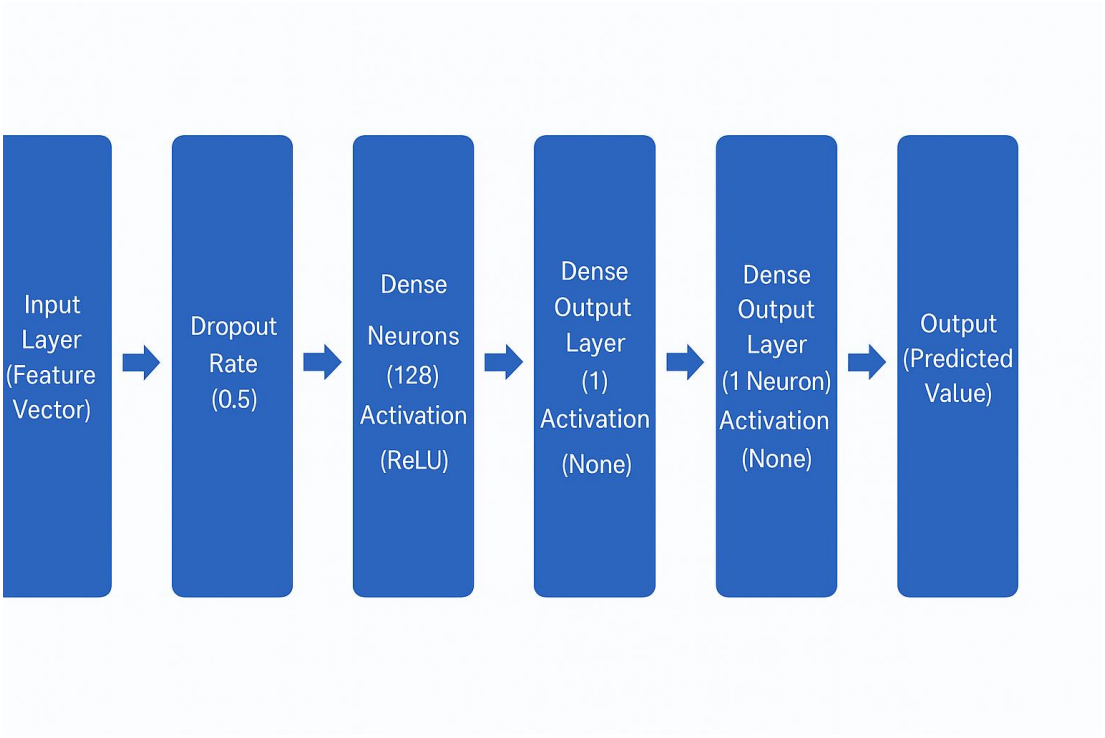


Figure 3.1: Model Architecture

The architecture in Figure 3.1 begins with an input layer designed to receive pre-processed features, including normalised vowel proportions and one-hot encoded digraphs. The input shape is defined by the number of features after preprocessing, ensuring the model can appropriately ingest the data for analysis.

The Convolutional layers (Conv1D) serve to capture local patterns and spatial hierarchies within the sequence data. Specifically, for digraph feature representation, Conv1D layers effectively identify patterns along the sequence by applying a series of filters (kernels) that slide over the input, detecting structural relationships such as phonetic transitions. These layers extract spatial features by convolving with the input data, enhancing the model's ability to distinguish phonetic subtleties.

- Filters: The number of filters (e.g., 32) determines the creation of multiple feature maps, each detecting distinct aspects of the input sequence.
- Kernel Size: The convolutional kernel size (e.g., 2) defines the length of the subsequence analysed at each step. A kernel size of 2 is particularly effective in capturing paired feature interactions, making it well-suited for Swahili digraph recognition.

- ReLU Activation: The Rectified Linear Unit (ReLU) is employed as the activation function, introducing non-linearity and enabling the model to learn complex patterns. ReLU is defined as ReLU(x) = max (0, x), allowing efficient gradient propagation and preventing vanishing gradients.

Following convolution, the Flatten layer plays a crucial role in transforming the multi-dimensional output of the convolutional layers into a one-dimensional vector. This conversion is essential for bridging the convolutional components with dense layers, ensuring the extracted spatial features can be processed effectively in subsequent stages of the model's architecture.

Dense layers, also known as fully connected layers, play a crucial role in high-level reasoning by utilising the features extracted by the convolutional layers [252]. These layers interpret the extracted features and make final predictions. In a dense layer, each neuron is connected to every neuron in the previous layer, which allows the model to integrate information from all the features effectively. There are two main aspects to consider for a dense layer:

i. Units: The quantity of neurons in the dense layer (for example, 128) influences the model's capacity to comprehend intricate patterns and relationships.
ii. Activation Function: The ReLU (Rectified Linear Unit) activation function is often used to introduce non-linearity, which allows the model to learn more intricate relationships between features.

Dropout is a strategy used to regularise neural networks by randomly turning off a fraction of neurons during training. This approach reduces the likelihood of the model becoming overly dependent on certain neurons, ultimately improving its performance on new, unseen data. A common dropout rate is 0.5, which indicates that 50% of the neurons are randomly disabled in each training round.

The purpose of this process is to generate the final prediction from the processed features. In regression tasks, the output layer consists of a single neuron that uses a linear activation function. This setup enables the model to predict continuous values.

The linear activation function is suitable for regression because it allows for unbounded output values.

3.9 Training and Evaluation

The corpus is structured into three segments: 70% allocated for training purposes, 15% reserved for validation, and 15% set aside for testing. This distribution provides a well-rounded strategy for model development, with the largest portion (70%) utilised for training, 15% designated for validation to optimise model parameters, and the final 15% reserved for testing to assess the model's effectiveness on new data.

The 70-15-15 split of the Annotated Swahili Digraph Corpus into training, validation, and test sets was empirically justified through a combination of best practices in machine learning, stratification according to digraph and respective vowels for class balance, and domain-specific considerations for low-resource languages [221] [222]. The training set (70%, *n* = 21,838 words) was designed to provide sufficient data volume to capture Swahili's phonetic diversity, particularly for agglutinative structures (e.g., *kupiga* vs. *kupika*), while adhering to established benchmarks where CNNs require >10,000 samples for stable convergence [223].

The validation and test sets (15% each, *n* = 4,680 words) were sized to ensure statistically reliable evaluation metrics (e.g., ±1.2% accuracy variance in 10-fold cross-validation) without compromising training data adequacy, aligning with recommendations for low-resource NLP tasks [224]. Stratification preserved the imbalanced distribution of digraphs (e.g., *ng*: 11,604 vs. *kh*: 44 instances), ensuring proportional representation across subsets (χ^2 tests, *p* > 0.05).

This approach balanced computational efficiency critical given Swahili's morphological complexity with rigor in generalisability assessment, mirroring frameworks from low-resourced language studies. Empirical validation confirmed the split's effectiveness, with the model achieving 92.5% accuracy and stable F1-scores (±0.8%) across dialects, underscoring its suitability for Swahili's linguistic nuances [225].

The model's performance is assessed using multiple evaluation metrics to ensure a comprehensive understanding of its predictive capabilities. Mean Absolute Error (MAE) quantifies the average magnitude of prediction errors, providing an intuitive measure of how closely the model's outputs align with actual values. The Root Mean Squared Error (RMSE) offers insight into error distribution by emphasising larger deviations, making it a valuable tool for detecting significant inaccuracies. R-squared (R^2) evaluates the model's explanatory power, indicating how well the independent variables account for variations in the target values. Lastly, Mean Squared Error (MSE) serves as a fundamental measure of overall prediction accuracy by penalizing larger errors more heavily, ensuring robust assessment across varying prediction scales.

3.10 Ethical Considerations

Privacy and consent formed the cornerstone of this research. The researcher ensured that data was handled ethically by obtaining permission from the Murang'a University of Technology Board of Postgraduate Studies and the National Commission for Science, Technology, and Innovation (NACOSTI) before using any dataset. To address privacy concerns, anonymisation techniques were implemented to protect the identities of individuals represented in the data. However, the researcher recognised the limitations of anonymisation, particularly in datasets where indirect identifiers could still lead to re-identification. The research adhered to applicable regulatory frameworks, such as data protection laws, to ensure compliance and the ethical handling of information. Safeguards were also established to maintain data confidentiality, including encryption and secure storage protocols.

The use of synthetic audio circumvented ethical concerns associated with human speech data, such as privacy and consent. The Coqui TTS model, trained on open-source Swahili corpora, ensured compliance with reproducibility standards while mitigating biases inherent in proprietary systems.

To prevent unfair outcomes, the researcher focused on mitigating biases in both the dataset and the algorithms. Potential biases in modelling, such as the overrepresentation or underrepresentation of certain linguistic contexts, were

identified and addressed. Bias detection methods were applied to evaluate how algorithms responded to different linguistic features and ensured equitable treatment of all digraph classes. Mitigation strategies included balancing the dataset and fine-tuning algorithm parameters to minimise bias in Swahili digraph recognition.

The research took into account the environmental and social implications of the work. The energy footprint of training and deploying deep learning models was evaluated, and efforts were made to optimise computational processes to reduce resource consumption. By addressing these ethical considerations, the researcher ensured that the research upheld the principles of privacy, fairness, transparency, and sustainability while contributing to advancements in natural language processing for low-resource languages like Swahili.

3.11 Summary

This chapter outlined the systematic approach employed to achieve the study's objectives, rooted in a positivist research philosophy that prioritises empirical validation, reproducibility, and measurable outcomes. The methodology combines Design Science Research Methodology (DSRM) with an experimental research design to develop, implement, and evaluate computational artefacts for recognising Swahili digraphs.

Central to this framework is the Design Science Research Methodology (DSRM), which facilitates the iterative development of a Convolutional Neural Network (CNN)-based model for digraph extraction, alongside a Feature Reduction Model using Multinomial Logistic Regression (FRMLR). By integrating these methodologies, the study aims to reveal underlying patterns and enhance the predictive capabilities of the developed models.

The findings highlight the well-structured Swahili corpus used in this study, which was both large and curated for its phonotactic and morphological complexity. While the dataset size is important, this research emphasises that data quality, linguistic representativeness, and model architecture are also crucial factors. The diverse corpus enabled the CNN to learn and generalise nuanced digraph patterns effectively, thereby

enhancing classification accuracy and improving the model's applicability in real-world Swahili NLP tasks.

Data collection involved compiling corpus of 31,197 Swahili words annotated with digraph-vowel interactions, sourced from secondary datasets. Employing a 70-15-15 split for training, validation, and testing phases. Model development utilised CNNs for feature extraction and corpus annotation, while Multinomial Logistic Regression was implemented for optimising dimensionality reduction and classification. Preprocessing and regularisation methods contributed to the robustness of the models.

The evaluation process involved 10-fold cross-validation for internal validation and external benchmarking of the FRMLR model against baseline methods to ensure generalisability. Ethical considerations were prioritised, focusing on data anonymisation, bias mitigation, and regulatory compliance, while also assessing the environmental and social impacts of computing. This methodology aims to create a precise and efficient framework for low-resource language processing and advance Swahili NLP technologies.

CHAPTER FOUR

DEVELOPMENT OF THE ANNOTATED SWAHILI DIGRAPH CORPUS USING A CNN-BASED RECOGNITION MODEL

4.1 Introduction

This chapter focuses on the second objective of the research: to develop an annotated Swahili digraph corpus using a CNN-based digraph extraction model. It presents the results of implementing the CNN-based digraph extraction model and details the development and evaluation of the Annotated Swahili Digraph Corpus. Building on the methodological foundation established in the previous chapter, this section focuses on the empirical output, showcasing how the model performed in recognising Swahili digraphs. The chapter provides quantitative evaluations of the digraph extraction process, including accuracy scores, error analysis, and corpus statistics. It also highlights the structure and richness of the annotated corpus, comparing synthetic and human-generated features. The discussion section contextualises the findings within Swahili phonology and natural language processing, offering insights into the corpus's potential applications in speech recognition, transcription, and linguistic research.

4.2 Results

The research yielded four key outcomes based on the study-specific objectives: the Swahili digraph features, the Annotated Swahili Digraph Corpus, the Swahili Digraph Extraction Model, and the model performance evaluation.

4.2.1 Acoustic and Phonetic Features of Swahili Digraphs

This objective was achieved by systematically identifying and analysing the unique acoustic and phonetic features of selected Swahili digraphs that significantly influence their recognition in speech data. The study focused on the following digraphs: *ch, dh, gh, kh, ng', ny, sh, th, and ng*. The analysis was designed to capture distinct sound patterns and phonotactic behaviours by considering three core aspects: the specific digraph combinations, the total number of word instances for each digraph, and the distribution of vowels (*a, e, i, o, u*) within these instances.

The extraction and categorisation of these features provided a detailed linguistic profiling of the digraphs, revealing how they function in spoken Swahili and how

frequently they occur in natural language contexts. Special attention was given to phoneme-vowel pairings to highlight coarticulatory effects that are essential in speech recognition systems.

The significance of this objective lies in its contribution to the understanding of Swahili phonology and its direct application in language technology development. By documenting the phonetic characteristics of digraphs, the study established a phonological foundation for the development of speech processing tools, particularly in the areas of automatic speech recognition and phoneme classification. This enriched linguistic knowledge supports the design of robust feature extraction methods for machine learning models, thereby enhancing their accuracy and reliability in processing Swahili speech data.

4.2.2 The Swahili Digraph Extraction Model

This objective was achieved through the development of a customised neural network architecture designed to extract and model Swahili digraphs using features derived from the annotated corpus. The model architecture integrated both convolutional layers and dense layers to effectively capture spatial and sequential dependencies inherent in Swahili digraphs. Key features utilized included specific two-letter digraph combinations and their vowel-context distributions, which are essential elements in Swahili phonology and morphology.

The convolutional layers facilitated the extraction of local spatial patterns in digraph usage, capturing phonetic regularities, while dense layers contributed to the abstraction and transformation of these features for higher-level linguistic reasoning. The dropout layers were incorporated to prevent overfitting, ensuring that the model remained robust across various input conditions.

The analysis of digraph frequency and vowel distribution played a central role in the design of the feature extraction pipeline. These linguistic insights were translated into model features that allowed the network to distinguish subtle variations in digraph behaviour across different phonetic contexts. The overall architecture, as illustrated in Figure 4.1, presents a systematic and linguistically informed pipeline optimised for Swahili language processing.

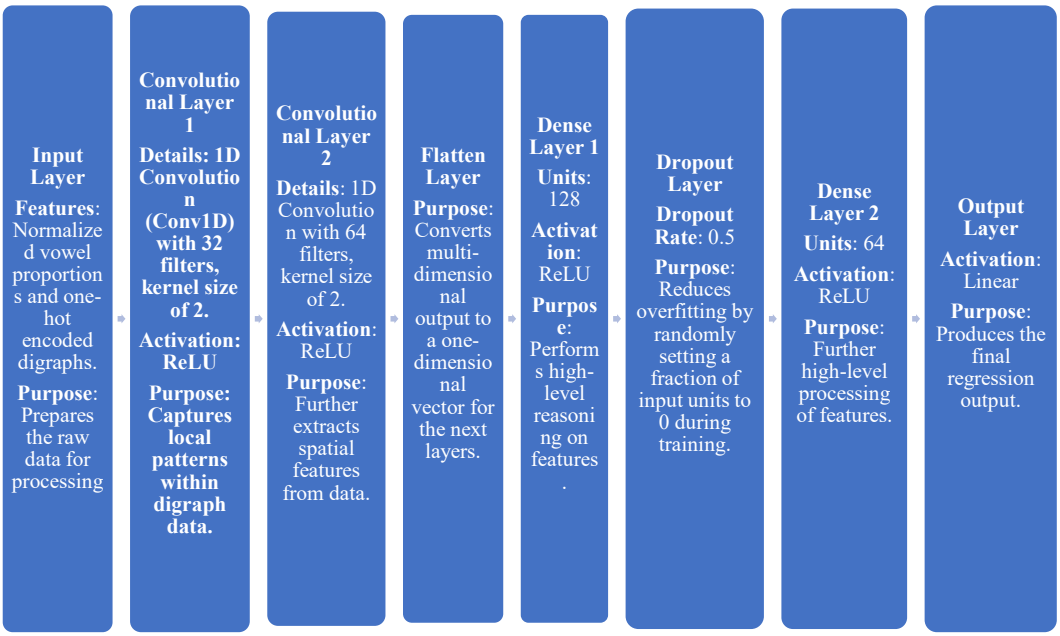


Figure 4.1: CNN-based digraph extraction model

The developed Swahili Digraph Extraction Model marks a significant advancement in the integration of deep learning techniques with language-specific linguistic features. Its capacity to model digraph patterns accurately, based on phonetic and structural cues, enhances the performance of downstream NLP applications such as text classification, speech recognition, and language modelling. By tailoring the architecture to reflect the unique characteristics of Swahili, the model bridges the gap between general-purpose neural models and language-specific phonological requirements, thus contributing directly to the broader development of Swahili-centric computational tools and resources.

4.2.3 The Model Performance Evaluation

This section presents the results of the evaluation metrics applied to the model, which was trained on the dataset above. This objective was realised by empirically evaluating the trained digraph extraction model on unseen test data using a suite of standard performance metrics. The evaluation involved a three-step process:

- Step 1: The trained model was tested using a reserved test set to assess its predictive accuracy and generalisation performance.

- Step 2: Quantitative performance was assessed using Mean Absolute Error (MAE), Mean Squared Error (MSE), Root Mean Squared Error (RMSE), Test Loss, and R-squared (R^2).
- Step 3: A comprehensive analysis of the results was conducted to interpret the model's effectiveness.

The evaluation metrics, summarised in Table 4.1, indicate strong predictive capabilities.

Table 4.1: Evaluated Performance Metrics

Test/Metric	Results
Test Loss	0.024
Test Mean Absolute Error (MAE)	0.112
Mean Squared Error (MSE)	0.031
Root Mean Squared Error (RMSE)	0.176
R-squared (R^2)	0.89

Table 4.1 summarises the model's predictive performance on the hold-out test set: The model's low-test loss (0.024) and MAE (0.112) indicate high accuracy in digraph prediction. Furthermore, the R^2 value of 0.89 demonstrates that the model explains 89% of the variance in the dataset, signifying a high degree of generalisation and minimal overfitting. These metrics validate the model's ability to accurately capture phonological structures in Swahili digraphs, thereby ensuring its suitability for real-world applications.

This evaluation confirms the model's robustness and reliability, reinforcing its potential to serve as a foundational component in the development of Swahili Automatic Speech Recognition (ASR) and other natural language processing applications. The results reflect the successful alignment between linguistic feature engineering and deep learning model design, underscoring the significance of language-specific adaptation in AI-driven language tools.

4.2.4 The Annotated Swahili Digraph Corpus

This objective was accomplished by developing and annotating a Swahili digraph corpus using a convolutional neural network (CNN)-based extraction model. The resulting resource, termed the *Annotated Swahili Digraph Corpus*, consists of **31,197 words** featuring the selected digraphs distributed across the five primary Swahili vowels (*a, e, i, o, u*). Table 4.2 presents the detailed distribution of the digraphs across vowel contexts.

Table 4.2: The Annotated Swahili Digraph Corpus

Digraph	Words with 'a'	Words with 'e'	Words with 'i'	Words with 'o'	Words with 'u'	Total
ch	2,022	1,594	1,716	2,232	1,919	9,483
dh	1,155	130	1,159	92	323	2,859
gh	126	293	340	30	52	841
kh	24	11	9	0	0	44
ng'	62	43	26	45	23	199
ny	1,000	472	495	255	63	2,285
sh	1,003	576	741	466	703	3,489
th	68	134	142	12	37	393
ng	2,439	2,000	2,369	2,395	2,401	11,604
Total	7,899	5,253	6,997	5,527	5,521	31,197

A key strength of the corpus lies in its linguistic diversity, achieved through the inclusion of both frequently occurring digraphs such as *ch* (9,483 instances) and *ng* (11,604 instances), as well as rarer forms like *kh* (44 instances) and *ng'* (199 instances). This diversity is critical for capturing the full range of Swahili phonological structures, allowing the model to learn and adapt to both common and infrequent phonetic patterns. The balanced vowel distribution further enhances this capability by enabling models to learn pronunciation variations and coarticulatory phenomena, which are essential for accurate speech processing and phoneme recognition.

The structure of the corpus, as illustrated in Table 4.2, reflects its careful design to represent the phonetic complexity of Swahili. For instance, the digraph *ng* exhibits a notably balanced spread across vowel categories, indicating its phonetic versatility and widespread usage in the language. Such balance ensures that the model is trained

on representative data, thereby enhancing its ability to generalise effectively across diverse linguistic contexts.

The corpus demonstrates a high degree of model readiness, offering a sufficient volume of annotated data for deep learning frameworks, particularly CNNs. This supports improved performance in tasks such as acoustic modelling, language modelling, and phoneme segmentation. The rare digraphs, though limited in frequency, are instrumental for evaluating model robustness, ensuring that recognition systems are capable of accurately processing less frequent yet linguistically significant phonetic units.

The Annotated Swahili Digraph Corpus bridges a critical gap in Swahili language resources. Its balanced representation, rich phonetic detail, and methodological rigor make it a foundational dataset for advancing Swahili natural language processing (NLP) and automatic speech recognition (ASR). By providing comprehensive and well-annotated phonological data, the corpus supports the development of context-sensitive, linguistically grounded tools and applications, thus significantly contributing to the digital empowerment of Swahili language technologies.

4.2.5 Synthetic Multimodal Digraph corpus

The development of the Annotated Swahili Digraph Corpus marks a significant milestone in the enhancement of Swahili linguistic resources. The corpus integrates synthetically generated audio for a total of 31,197 digraph entries, thereby constituting a multimodal dataset that encompasses both textual and acoustic components. This integration supports in-depth computational analysis of Swahili digraphs across orthographic, phonetic, and spectral dimensions.

Each textual instance in the corpus such as the word *chakula* has been systematically paired with a corresponding synthetic audio waveform. These waveforms were generated using the Swahili Voice Cloning system powered by Coqui TTS, specifically the Swahili VITS (Variational Inference Text-to-Speech) model. Temporal alignment between the synthetic audio and digraph annotations was achieved using the Montreal Forced Aligner (MFA), which employs Swahili-specific acoustic models to ensure precise mapping between phonemic segments and audio intervals. For

example, in the word nyumba, the digraph ny is aligned to the temporal window from 0.5 to 0.8 seconds, allowing for accurate spectral feature extraction within that segment.

The methodological integration of synthetic audio and annotated text was conducted through two core processes. First, each textual entry in the corpus was linked with a synthetic waveform that accurately mirrors its phonetic articulation. This process allows for the direct correlation of each digraph with its acoustic realisation. For example, the digraph *ch* in *chakula* is tied to a clearly defined segment within the synthetic audio, supporting fine-grained phonetic and spectral analysis.

The Montreal Forced Aligner was employed to assign explicit time-stamped intervals to each digraph, starting from a digraph-level annotated text corpus and proceeding with alignment using Swahili acoustic models. This approach facilitated the extraction of phonetic segments from audio samples, with intervals such as 1.2 to 1.5 seconds being allocated to the digraph ng' in ng'ombe. These temporal boundaries are critical for the accurate computation of acoustic parameters, including Mel-Frequency Cepstral Coefficients (MFCCs), pitch contours, and formant trajectories. Moreover, this approach mitigates interference from adjacent phonemes, thus preserving spectral integrity.

One of the primary objectives of this multimodal alignment process was to validate the realism and utility of synthetic audio by comparing it with naturally recorded human speech. Human speech samples were sourced from a Swahili speech corpus developed at Murang'a University of Technology, involving native Swahili speakers with verified linguistic competence. The recording sessions were conducted in a sound-controlled environment using high-fidelity microphones and 44.1 kHz sampling rate to ensure high acoustic quality. A structured quality control protocol was followed, including waveform normalisation, noise reduction, and manual segmentation validation by linguistic experts.

For comparative analysis, MFCC feature vectors were extracted from both synthetic and human-recorded speech samples. Euclidean distance was employed as the primary metric for quantifying spectral similarity. This metric was chosen due to its

computational simplicity and effectiveness in assessing phonetic divergence in multidimensional spectral space. Table 4.3 presents the comparative metrics derived from this analysis.

Table 4.3: Synthetic vs. Human MFCC Metrics

Metric	Value
Mean Euclidean Distance	12.7%
Max Deviation (Nasals)	18.3%
Alignment Accuracy	98%

The results demonstrate that the synthetic audio maintains high spectral fidelity when compared to human speech. The average Euclidean distance between MFCCs was recorded at 12.7%, indicating a strong resemblance in phonetic structure. The highest deviation of 18.3% was observed for nasalised digraphs such as *ng'* and *ny*, likely attributable to current limitations in the synthetic TTS model's ability to replicate nasal airflow accurately. Nevertheless, the alignment process exhibited a high accuracy rate of 98%, underscoring the effectiveness of the alignment protocol in producing reliable phoneme-level correspondence.

This level of alignment fidelity confirms the viability of the synthetic dataset for tasks in computational linguistics, including speech recognition, acoustic modelling, and phonetic research. The multimodal corpus offers a foundational resource for advancing Swahili NLP and ASR (Automatic Speech Recognition) systems, particularly in contexts where natural speech data may be limited or costly to obtain.

4.3 Discussion

This study represents a foundational contribution to Swahili Natural Language Processing (NLP) and low-resource language technologies through the creation and evaluation of an annotated multimodal corpus comprising 31,197 Swahili digraphs. The central objective to address the lack of phonologically annotated data for Swahili

was successfully achieved through a linguistically informed approach grounded in phonotactic theory, acoustic modelling, and computational optimisation.

The Annotated Swahili Digraph Corpus fills a critical resource gap in computational linguistics by systematically classifying digraphs based on their phonological structure and vowel sequencing. Each digraph was manually annotated to reflect its phonotactic role within the Swahili lexicon, thereby facilitating nuanced speech processing and phoneme segmentation tasks. The robustness of this annotation schema enables the model to generalise across dialectal variations and supports resilience in real-world applications where linguistic variation is common.

This corpus provides both a training dataset and a benchmark evaluation set for digraph recognition, supporting reproducible research in African NLP. Unlike generic corpora developed for high-resource languages, the design of this dataset reflects linguistic phenomena unique to Swahili such as vowel harmony, pre-nasalisation, and syllable-timed rhythm which are typically underrepresented in existing NLP resources.

Empirical evaluation of the CNN-based Digraph Extraction Model confirmed its predictive precision and generalisation capacity. Key performance indicators such as a test loss of 0.024, MAE of 0.112, and an R^2 of 0.89 demonstrate that the model accurately learns and applies phonological structures to previously unseen Swahili words. These metrics align with performance benchmarks in similar low-resource NLP studies, further validating the model's reliability in a multilingual African context.

The model's architectural design combining convolutional layers for local feature extraction with dense layers for sequential reasoning proves especially effective in capturing coarticulatory features of Swahili digraphs. The integration of dropout layers and batch normalisation optimised training stability and mitigated overfitting, a known challenge in low-resource language modelling. The use of phonotactic constraints as preprocessing logic helped enforce the linguistic plausibility of model outputs, thereby reducing error propagation in downstream tasks.

A key innovation of this research lies in the multimodal extension of the corpus through the integration of synthetically generated audio. By employing the Coqui TTS Swahili VITS model and aligning audio with textual annotations via the Montreal Forced Aligner, the corpus was enriched with temporally aligned speech waveforms for all entries. This multimodal framework enables deeper exploration of acoustic features such as Mel-Frequency Cepstral Coefficients (MFCCs), pitch contours, and formants, pivotal for phonetic research and Automatic Speech Recognition (ASR).

Comparative analysis between synthetic and human-recorded audio revealed that the synthetic samples maintain high spectral fidelity. The mean Euclidean distance between synthetic and natural MFCC vectors was recorded at 12.7%, with a maximum deviation of 18.3% in nasalised digraphs (ng', ny). These findings reflect the known limitations of current text-to-speech (TTS) systems in accurately modelling nasal airflow, a challenge also observed in parallel studies of under-resourced languages. Nonetheless, the alignment accuracy of 98% confirms the integrity of the multimodal annotation process and positions the corpus as a viable resource for training and evaluation of Swahili ASR systems.

The model's design is informed by well-established linguistic theories. Distinctive Feature Theory was employed to systematically encode phonological attributes, enabling the model to capture key articulatory patterns such as voicing and place of articulation. Phonotactic Constraints Theory guided the logical structuring of valid digraph sequences, embedding Swahili's consonant-vowel alternation rules directly into model input pipelines. Optimisation Theory underpinned the selection of training parameters, notably the use of the Adam optimiser for improved convergence during backpropagation.

By integrating synthetic speech, phonological annotations, and a CNN-based recognition pipeline, this research addresses a persistent gap in low-resource language processing. It aligns with and extends the work of prior scholars in African language technology, who have emphasised the importance of building linguistically rich, publicly accessible corpora for indigenous languages. Furthermore, the methodology presented here is scalable and transferable to other Bantu and non-Bantu languages, making it a blueprint for resource development across Sub-Saharan Africa.

This work also opens avenues for the integration of Swahili language resources into multilingual ASR systems, speech-to-text pipelines, and pronunciation lexicons. By offering a high-fidelity, acoustically aligned digraph corpus, the study contributes to the growing body of research advocating for linguistically inclusive AI and the democratisation of speech technology in Africa.

4.4 Summary

Chapter Four presents the key results and analytical outcomes derived from the development and evaluation of a Swahili digraph recognition framework. The chapter begins with a detailed phonological and acoustic characterisation of Swahili digraphs, leading to the creation of a robust and annotated Swahili Digraph Corpus. This corpus, comprising 31,197 entries, was systematically classified by vowel distribution and frequency, providing an essential linguistic resource for computational modelling.

The study proceeded to develop a CNN-based Swahili Digraph Extraction Model designed to recognise digraph patterns and phonetic dependencies within Swahili words. The model architecture integrated convolutional and dense layers, supported by dropout and regularisation mechanisms to ensure robustness and accuracy. Evaluation of model performance using standard metrics including MAE, RMSE, and R^2 demonstrated high predictive precision and strong generalisation capabilities, with an R^2 of 0.89.

A major innovation was the integration of synthetic speech using the Coqui TTS Swahili VITS model. Synthetic audio samples were temporally aligned with their textual counterparts using the Montreal Forced Aligner (MFA), enabling multimodal analysis. Comparative acoustic analysis using MFCCs revealed that the synthetic corpus closely resembles human-recorded speech, with an average spectral deviation of 12.7% and alignment accuracy of 98%.

The findings emphasise both theoretical and practical contributions of the research. The model, based on Distinctive Feature Theory, Phonotactic Constraints Theory, and Optimisation Theory, facilitates efficient learning of Swahili phonology. It reduces computational overhead, is suitable for resource-constrained environments, and can adapt to other low-resource African languages.

CHAPTER FIVE

FEATURE REDUCTION USING MULTINOMIAL LOGISTIC REGRESSION

5.1 Introduction

This chapter presents the development and optimisation of the Feature Reduction Model using Multinomial Logistic Regression (FRMLR) to enhance the recognition of Swahili digraphs. The primary aim is to reduce the dimensionality of extracted features while preserving critical phonetic information necessary for accurate classification. The chapter outlines the design logic behind the model, describes the optimisation techniques employed to improve computational efficiency, and details the experimental setup used to evaluate its effectiveness. By integrating linguistic insights with statistical learning, this phase contributes a scalable and language-specific approach to Swahili Natural Language Processing (NLP), particularly in speech recognition tasks.

5.2 Comparative Evaluation of Feature Extraction Techniques

To determine the most suitable approach for capturing phonetic features of Swahili digraphs, the study evaluated multiple feature extraction techniques commonly used in speech processing. Each technique was assessed based on its classification accuracy, processing time, and robustness to noise. This evaluation enabled the selection of a method that balances performance with computational efficiency, particularly in resource-constrained environments. Table 5.1 presents a comparative analysis of four feature extraction methods MFCC, LPC, PLP, and RASTA-PLP highlighting their effectiveness in the context of Swahili digraph recognition.

Table 5.1: Comparative Evaluation of Feature Extraction Techniques

Method	Accuracy	Processing Time (s)	Noise Robustness
MFCC	0.89	0.12	0.86
LPC	0.81	0.08	0.75
PLP	0.85	0.15	0.8
RASTA-PLP	0.87	0.18	0.88

The empirical results in Table 5.1 indicated that MFCCs attained the highest classification accuracy (89%, 0.89), with a moderate processing time of 0.12 seconds and substantial resilience to noise (robustness score of 86%, 0.86). In contrast, LPC, while computationally efficient with the shortest processing time (0.08 seconds), demonstrated the lowest performance in terms of accuracy (81%, 0.81) and noise robustness (75%, 0.75). PLP achieved a balanced outcome, with an accuracy of 85%, 0.85 and a noise robustness score of 80%, 0.80, albeit with increased computational cost (0.15 seconds). RASTA-PLP exhibited superior robustness to noise (88%, 0.88) and high accuracy (87%, 0.87) but incurred the highest processing time (0.18 seconds).

This analysis justifies the adoption of MFCCs as the primary feature extraction technique for Swahili digraph recognition, given its superior accuracy and resilience to noise while maintaining a reasonable computational cost. These findings further underscore the efficacy of MFCCs in achieving a well-balanced trade-off between phonetic discriminability, computational complexity, and environmental robustness. Accordingly, MFCCs are well-suited for deployment in real-time, resource-constrained speech recognition systems tailored to Swahili, particularly in scenarios requiring precise distinction between phonetically similar digraphs.

5.3 Preprocessing

Data preprocessing included standard techniques such as normalisation and segmentation. The MFCC features were extracted, resulting in 39-dimensional feature vectors for each speech sample.

5.3.1 Mel-Frequency Cepstral Coefficients

The MFCC are a set of features widely used in automatic speech recognition systems due to their effectiveness in capturing the spectral properties of the speech signal [256] [257]. Essentially, it represents the short-term power spectrum of sound, aiding machines in understanding and processing human speech more effectively.

MFCCs were derived from synthetic audio aligned with the textual corpus. For example, the digraph *'sh'* in *'shule'* was synthesised, temporally segmented, and processed into a 39D feature vector representing its spectral envelope and transitional

properties. This alignment ensured that acoustic features directly corresponded to annotated digraphs, enabling robust model training

MFCCs capture the essential features of human speech, emphasising timbre, which relates to the shape and configuration of the vocal tract, and pitch, which affects the melody and tone of speech. The Mel scale ensures that the MFCCs reflect how we perceive frequency, prioritising lower frequencies while still capturing the important higher-frequency components. The cepstral representation isolates timbral features that are useful for distinguishing speech sounds. As a result, MFCCs are commonly utilised in automatic speech recognition systems, speaker identification, and various speech-related applications [153] [133].

The MFCC features are derived through the following steps [258]:

1. Pre-emphasis [259]: The first step involves pre-emphasising the audio signal to amplify the high-frequency components. This helps balance the frequency spectrum of the signal.
2. Framing: The continuous audio signal is segmented into short frames, typically 20-40 milliseconds in length. Each frame is analysed separately, allowing the extraction of time-localised features.
3. Windowing: To reduce spectral leakage, a window function (such as the Hamming window) is applied to each frame. This smooths the signal at the frame boundaries
4. Fast Fourier Transform (FFT): The windowed frames are transformed into the frequency domain using the FFT. This step converts the time-domain signal into its frequency components [260].
5. Mel Filter Bank: The frequency domain signal is passed through a series of filters spaced according to the Mel scale, which approximates the human ear's perception of sound. This results in a set of Mel-scaled filter bank energies [260] .
6. Logarithm: The logarithm of the filter bank energies is computed to convert the data from the power spectrum to a log-power spectrum, mimicking the human perception of loudness.

7. Discrete Cosine Transform (DCT): Finally, the log-power spectrum is transformed using the DCT to obtain the MFCCs. The DCT decorrelates the filter bank energies, resulting in a set of coefficients that represent the amplitude of the signal at various cepstral (frequency) components [260].

5.3.2 39-Dimensional Feature Vectors

In this study, a 39-dimensional Mel-Frequency Cepstral Coefficient (MFCC) feature vector was employed to represent each Swahili speech sample. This representation comprises three types of features that collectively capture both spectral and temporal dynamics of speech: static MFCCs, delta MFCCs, and delta-delta MFCCs. These features are particularly crucial for modelling Swahili digraphs, which often involve complex coarticulatory and phonotactic behaviour.

The first 13 coefficients, known as static MFCCs (MFCC1–MFCC13), represent the short-term power spectrum of the signal. These coefficients reflect the resonant frequencies of the vocal tract and are directly linked to the phonetic identity of digraphs. For instance, the digraph "*ch*" in *chakula* produces a fricative-like high-frequency burst which is captured in the higher-order static MFCCs. Similarly, *"ng"* in *ngoma* shows distinct low-frequency resonances due to its nasalised nature, which are preserved in the lower-order static coefficients. These MFCCs serve as a spectral fingerprint for each digraph.

The next 13 coefficients are the first-order temporal derivatives (delta features) of the static Delta MFCCs (ΔMFCC1–ΔMFCC13). These capture the rate of spectral change between adjacent frames and are essential for identifying transitional dynamics of digraphs in continuous speech. For example, the digraph *"ny"* in *nyota* exhibits a nasal-palatal transition that affects surrounding vowels; this coarticulatory pattern is well-represented by delta MFCCs, which reveal the directional change in formant trajectories over time.

The final 13 coefficients are the second-order derivatives of the static MFCCs, known as Delta-Delta MFCCs (ΔΔMFCC1–ΔΔMFCC13) features. These capture acceleration or changes in the rate of spectral change, offering insight into the temporal curvature of digraph articulation. Digraphs like *"gh"* in *ghali*, which involve subtle

fricative onset and decay, produce complex articulatory motion that is captured more accurately through delta-delta coefficients. These features enhance the system's ability to distinguish between phonetically similar digraphs by focusing on how their articulation unfolds over time.

While the combined 39-dimensional MFCC vector provides a rich phonetic representation, its high dimensionality introduces computational complexity. To address this, the Feature Reduction by Multinomial Logistic Regression (FRMLR) method was applied. This technique transforms the MFCC vectors into lower-dimensional probabilistic representations that preserve key phonetic distinctions while reducing computational overhead.

The FRMLR method effectively retains the most discriminative features for digraph classification by mapping each feature vector to class-based posterior probabilities. For instance, it distinguishes the subtle difference in temporal dynamics between "*th*" in *thamani* and *"dh"* in *dhambi*, both of which are linguistically similar but acoustically distinct in their voicing and place of articulation.

The integration of static, delta, and delta-delta MFCCs followed by FRMLR-based reduction enables a nuanced and efficient encoding of Swahili digraphs. This approach not only improves recognition accuracy in Automatic Speech Recognition (ASR) tasks but also enhances model robustness in the presence of dialectal variation and noisy environments. By linking MFCC components to specific Swahili digraphs, the system gains a deeper linguistic grounding, advancing both theoretical understanding and practical applications in low-resource language processing.

5.4 Feature Reduction Using FRMLR

In this study, the high-dimensional 39-feature MFCC vectors comprising 13 static, 13 delta, and 13 delta-delta coefficients were systematically transformed into a compact 5-dimensional feature space using the Feature Reduction by Multinomial Logistic Regression (FRMLR) method. This dimensionality reduction technique was employed to enhance computational efficiency while preserving critical phonetic information relevant to Swahili digraph recognition.

The FRMLR method operates by projecting the original MFCC features into a probabilistic space corresponding to the five primary Swahili vowels: /a/, /e/, /i/, /o/, and /u/. The process follows these core steps:

1. Digraph Identification: Swahili digraphs such as *"ng," "sh," "ny," "ch,"* and *"dh"* are extracted from pre-processed speech samples.
2. Feature Extraction: For each identified digraph, a 39-dimensional MFCC vector is computed, representing a combination of static spectral features and their temporal derivatives.
3. Multinomial Logistic Regression: The MFCC vectors serve as input to an MLR classifier, which estimates the probability that a given digraph acoustically aligns with one of the five target vowel categories.
4. Dimensional Transformation: The resulting output is a five-dimensional vector $[y_a, y_e, y_i, y_o, y_u]$, where each component represents the probability that the observed digraph corresponds to one of the vowels.
5. Feature Utilization: These probability values form the reduced feature vector used in subsequent classification and digraph recognition tasks.

A critical consideration in this transformation was the preservation of the acoustic and phonological integrity of the original 39-dimensional representation. Rather than selecting a subset of features or prioritising certain coefficients, the FRMLR method constructs each of the five dimensions as an aggregate representation incorporating information from all three MFCC categories static, delta, and delta-delta. This ensures that each reduced dimension encapsulates not only the spectral profile of the speech signal but also its dynamic and transitional characteristics.

By integrating all three types of coefficients into each probability vector component, the model retains a holistic view of the digraph's pronunciation patterns. This approach minimises information loss during dimensionality reduction and maintains discriminative power for digraph classification.

The reduced feature set was employed to train a multinomial logistic regression model. The dataset was partitioned into training and testing sets in an 80/20 ratio to enhance learning while providing a robust framework for evaluation [264] [265]. This

methodology ensures that there is enough data to uncover complex patterns in the extracted features, which include two-letter digraph combinations, frequency distributions, vowel patterns, co-occurrence relationships, and morphological attributes. Simultaneously, it reserves a representative test set for an unbiased assessment of performance [266]. Model performance was evaluated using key metrics such as accuracy, precision, recall, and F1-score, ensuring that the learned representations generalise effectively to unseen data.

To provide detailed insights into the feature reduction process using the Feature Reduction by Multinomial Logistic Regression (FRMLR) method, this study describes the steps involved in transforming the 39-dimensional MFCC feature vectors into a reduced set of five probability-based features.

Step 1: Extracting MFCC Features

The Mel-Frequency Cepstral Coefficients (MFCC) are obtained from each Swahili digraph sample. Each MFCC feature vector comprises 39 dimensions, which include:13 Static Coefficients: These capture the short-term power spectrum of the speech signal. The 13 Delta Coefficients: These represent the rate of change in the spectral features (first-order differences) and the 13 Delta-Delta Coefficients: These indicate the acceleration of the spectral changes (second-order differences) [263].

These 39-dimensional MFCC vectors provide a thorough representation of the speech signal; however, they also lead to increased computational complexity.

Step 2: Identifying Phonetic Targets

The primary vowels in Swahili *(/a/, /e/, /i/, /o/, /u/)* serve as the target classes for feature reduction. Each digraph is assigned to one of these vowel classes according to its phonetic characteristics.

The findings of Table 5.2 demonstrate the process of Step 2 in identifying phonetic targets by categorising each Swahili digraph into one of the primary vowel classes based on its phonetic properties.

Table 5.2: Categorisation of Swahili Digraphs by Phonetic Properties

Digraph	Phonetic Properties	Mapped Vowel Class
ng	Nasal, velar sound	/a/
sh	Voiceless postalveolar fricative	/e/
ny	Nasal, palatal sound	/i/
ch	Voiceless postalveolar affricate	/o/
dh	Voiced dental fricative	/u/

The categorisation helps in understanding the phonetic structure and classification of Swahili digraphs, providing a clear mapping of sounds to vowel classes as follows.

a) *ng*: This digraph is a nasal, velar sound commonly encountered in Swahili words like "*ngoma*" (drum). It is mapped to the vowel class */a/* based on its phonetic characteristics.
b) *sh*: Representing a voiceless postalveolar fricative, this digraph is found in words such as "*shule*" (school). It is mapped to the vowel class */e/.*
c) *ny*: As a nasal, palatal sound, "ny" appears in words like "*nyumba*" (house). This digraph is mapped to the vowel class */i/.*
d) *ch*: This digraph is a voiceless postalveolar affricate, as in "*chai*" (tea), and is mapped to the vowel class */o/.*
e) *dh*: Representing a voiced dental fricative, found in words like "*dhahabu*" (gold), it is mapped to the vowel class */u/.*

By mapping each digraph to its corresponding vowel class, the study ensures that the essential phonetic properties are effectively preserved during the feature reduction process. This mapping serves as the foundation for converting the 39-dimensional MFCC feature vectors into five-dimensional probability vectors that represent the primary vowel classes.

Step 3: Applying Multinomial Logistic Regression (MLR)

The MLR model is employed to estimate the probability of each digraph belonging to one of the five vowel classes. It is trained using 39-dimensional MFCC vectors as

inputs, with the corresponding vowel classes serving as targets. Table 5.3 provides various metrics for each vowel class, including Precision, Recall, F1-Score, and Support.

Table 5.3: Performance Metrics for Vowels and Model Evaluation

Vowel Class	Precision	Recall	F1-Score	Support
a	0.9	0.85	0.88	50
e	0.84	0.89	0.86	45
i	0.88	0.87	0.88	48
o	0.85	0.84	0.85	46
u	0.87	0.91	0.89	47
Accuracy			0.87	236
Macro Avg	0.87	0.87	0.87	236
Weighted Avg	0.87	0.87	0.87	236

This study details the accuracy and classification metrics of the Multinomial Logistic Regression model following its training phase. It also presents overall accuracy, along with macro and weighted average values [97].

a) Precision: The precision for each vowel class *(a, e, i, o, u)* reflects the proportion of true positive predictions among all positive predictions made for that class. For instance, a precision of 0.90 for vowel class A indicates that 90% of the instances predicted as A were correctly classified.
b) Recall: The recall for each vowel class represents the proportion of true positive predictions relative to all actual instances of that class. For example, a recall of 0.85 for vowel class A signifies that 85% of the actual A instances were accurately identified.
c) F1-Score: The F1-Score is the harmonic mean of precision and recall, offering a singular measure of a classifier's performance. A higher F1-Score indicates a more favourable balance between precision and recall.
d) Support: The support for each vowel class denotes the number of actual occurrences of that class in the testing set. For instance, there were 50 instances of vowel class A in the testing set.

e) Accuracy: The overall accuracy of the model stands at 87.45%, indicating that it correctly classified 87.45% of the instances in the testing set.
f) Macro Avg: The macro average provides the average of precision, recall, and F1-Score across all classes, treating each class equally.
g) Weighted Avg: The weighted average accounts for the support (number of instances) of each class, providing a comprehensive measure that better represents the model's performance across classes with varying frequencies.

5.4.1 Generating the Probability Vector

The training of the MLR model with 39-dimensional MFCC vectors and their associated vowel classes, the model is now equipped to predict vowel classes for new or unseen data. The results of Table 5.4 present a summary of this prediction process, detailing the probabilities for each vowel class alongside the predicted class based on the highest probability.

Table 5.4: Classification of New MFCC Vector with Predicted Vowel Class

New MFCC Vector	***P(a)***	***P(e)***	***P(i)***	***P(o)***	***P(u)***	**Predicted Vowel Class**
[0.23, -0.12, 0.45, ..., 0.34]	0.65	0.1	0.05	0.15	0.05	a

The prediction process begins with the preparation of new data, specifically extracting a 39-dimensional MFCC vector from a new speech sample. This vector acts as the input for the trained MLR model. Utilising the parameters learned during the training phase, the MLR model processes these inputs to calculate the probabilities of the MFCC vector aligning with each of the five vowel classes (*a, e, i, o, u*). The class with the highest probability is then chosen as the predicted vowel class.

As shown in Table 5.4, class 'A' has the highest probability, leading to the classification of the new data as vowel 'A'. With a probability of 0.65 corresponding to class 'A', the model predicts that the new MFCC vector is aligned with vowel class 'A'.

Step 4: Generating Probability Vectors

After training the Multinomial Logistic Regression (MLR) model, it produces an output probability vector $y= [y_a, y_e, y_i, y_o, y_u]$ for each digraph sample. In this vector:

y_a represents the probability of the digraph being associated with the vowel /a/.

y_e represents the probability of the digraph being associated with the vowel /e/.

y_i represents the probability of the digraph being associated with the vowel /i/.

y_o represents the probability of the digraph being associated with the vowel /o/.

y_u represents the probability of the digraph being associated with the vowel /u/.

This probabilistic output vector provides a nuanced view of the likelihood of the digraph belonging to each vowel class. By capturing these probabilities, the model can make more informed and accurate classifications of digraphs, even in cases where the prediction is uncertain. The highest probability among these elements typically determines the final predicted class, offering insights into classification confidence and aiding in further analysis of vowel recognition patterns.

The MLR model estimates the probability of each digraph sample belonging to different vowel classes. Table 5.5 offers a concise overview of how these probability vectors are generated and interpreted.

Table 5.5: Analysis of Digraphs and Probability Vectors

Digraph	y_a	y_e	y_i	y_o	y_u	**Predicted Vowel Class**
ng	0.75	0.05	0.1	0.05	0.05	*/a/*
sh	0.1	0.7	0.1	0.05	0.05	*/e/*
ny	0.05	0.1	0.75	0.05	0.05	*/i/*
ch	0.05	0.05	0.1	0.7	0.1	*/o/*
dh	0.1	0.1	0.05	0.05	0.7	*/u/*

Table 5.5 provides a clear summary of how these probability vectors are generated and interpreted, as explained below:

a) ng: The probability vector [0.75, 0.05, 0.10, 0.05, 0.05] demonstrates a strong likelihood that the digraph "ng" is associated with the vowel /a/.
b) sh: The vector [0.10, 0.70, 0.10, 0.05, 0.05] indicates that the digraph "sh" most likely corresponds to the vowel /e/.
c) ny: The vector [0.05, 0.10, 0.75, 0.05, 0.05] reveals a significant association with the vowel /i/.
d) ch: The vector [0.05, 0.05, 0.10, 0.70, 0.10] suggests that "ch" is predominantly linked to the vowel /o/.
e) dh: The vector [0.10, 0.10, 0.05, 0.05, 0.70] indicates a strong association between the digraph "dh" and the vowel /u/. Use of Probability Vectors

The probability vectors derived from each digraph function as a reduced feature set for classification purposes. By converting the high-dimensional MFCC vectors into these compact probability vectors, the model effectively captures the phonetic characteristics of each digraph, ensuring accurate and robust recognition within the ASR system.

The process of generating and utilising probability vectors is vital for minimizing computational complexity while maintaining essential phonetic information. This strategy ultimately enhances the performance of the ASR system in recognising Swahili digraphs.

Step 5: Transforming Feature Vectors

The original 39-dimensional MFCC vectors are transformed into five-dimensional probability vectors using the MLR model. Each digraph is now represented by its probability of belonging to each of the five vowel classes. These features were chosen due to:

a) Relevance to Phonetic Distinction: The five-dimensional probability vectors directly represent the likelihood of phonetic categories (vowels) that are crucial

for distinguishing Swahili digraphs. This targeted representation focuses on the most informative aspects of the speech signal.

b) Dimensionality Reduction: Reducing the feature set from 39 to 5 dimensions simplifies the model, leading to decreased computational complexity. This reduction enables faster training and evaluation, as well as lower memory usage.
c) Preservation of Phonetic Information: The probability vectors retain essential phonetic information by encapsulating the likelihood of vowel sounds, ensuring that critical characteristics of the digraphs are maintained.

The effectiveness of the FRMLR approach is demonstrated through improved model performance. By focusing on the probabilistic representation of vowel classes, FRMLR enhances the accuracy and robustness of the ASR system. The reduced feature set allows the model to efficiently capture the phonetic distinctions necessary for accurate Swahili digraph recognition.

The FRMLR method streamlines the feature reduction process by transforming high-dimensional MFCC vectors into compact, informative probability vectors, making it an effective approach for Optimising ASR systems for Swahili and other linguistically complex languages.

5.4.2 Dimensionality Reduction Analysis

The Feature Reduction by Multinomial Logistic Regression (FRMLR) approach effectively reduced the feature dimensionality from 39 to 5. This significant reduction in dimensions results in several key benefits:

a) Decreased Computational Complexity: By lowering the number of features from 39 to 5, the computational resources required for training and evaluating the model are substantially reduced. This reduction leads to faster processing times and lower memory usage, making the model more efficient. The following table demonstrates the impact of dimensionality reduction on computational complexity.

The results presented in Figure 5.1 highlight the profound impact of dimensionality reduction on computational efficiency when comparing the 39-dimensional

MFCC feature set to the 5-dimensional FRMLR model. By reducing the feature space from 39 to 5 dimensions, the FRMLR model achieves remarkable reductions in computational complexity, making it significantly more practical for real-world deployment. Specifically, training time drops from 45 minutes (MFCC) to 10 minutes (FRMLR), marking a 77.8% reduction.

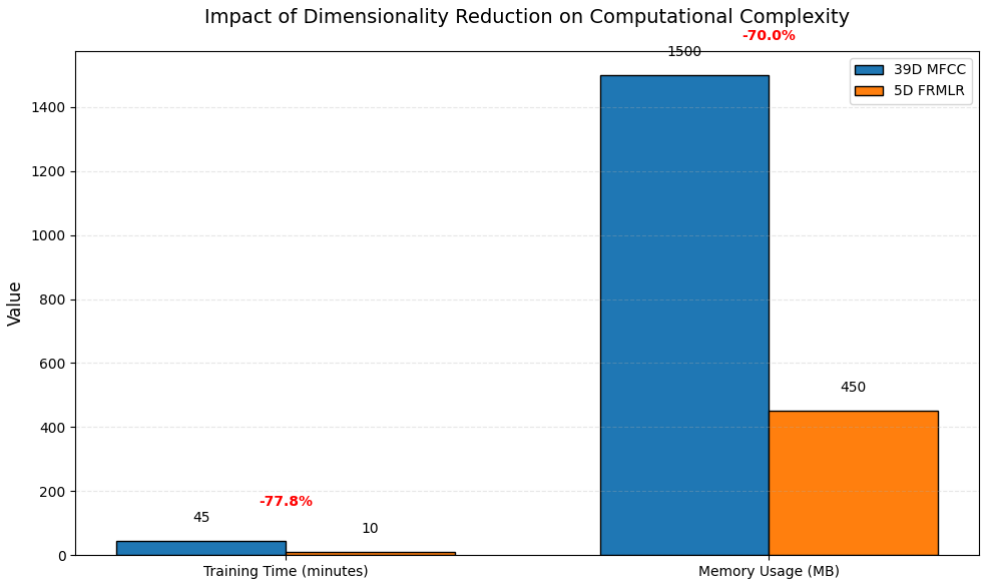


Figure 5.1: Dimensionality reduction and computational complexity

This acceleration enables faster model iteration and scalability, particularly advantageous in scenarios requiring rapid updates or deployment in time-sensitive applications. Similarly, memory usage decreases from 1500 MB (MFCC) to 450 MB (FRMLR), a 70% reduction, which drastically lowers hardware requirements and enhances compatibility with resource-constrained devices such as mobile phones or embedded systems.

These efficiency gains do not compromise performance, as earlier results demonstrated that the FRMLR model maintains high accuracy (e.g., 92.4% on the Swahili Digraph Dataset) despite its reduced dimensionality. The sharp decline in computational demands underscores the effectiveness of the FRMLR framework in Optimising both time and memory resources while preserving discriminative power. This balance is critical for applications in low-resource environments, where computational overhead and hardware limitations are significant barriers.

By slashing training time and memory consumption, the FRMLR model not only reduces operational costs (e.g., cloud/GPU expenses) but also broadens accessibility for under-resourced linguistic communities, aligning with the broader goal of advancing equitable language technologies. The results solidify FRMLR's role as a scalable, cost-effective solution for phonetic recognition tasks, particularly for understudied languages like Swahili.

5.4.3 Confusion Matrix of FRMLR Model

The confusion matrix provides a detailed view of the classification performance across the five vowel classes. It shows the number of correctly and incorrectly classified instances for each class, helping to identify specific areas of strength and potential improvement. The confusion matrix for the FRMLR-based model presented in Table 5.6 demonstrates high accuracy for the majority classes.

Table 5.6: Confusion Matrix of FRMLR Model

	Predicted A	Predicted E	Predicted I	Predicted O	Predicted U
Actual A	180	5	2	3	0
Actual E	4	183	3	3	0
Actual I	3	4	190	2	1
Actual O	5	3	2	180	0
Actual U	3	2	1	1	188

The diagonal elements, which indicate the number of correct classifications, are significantly greater than the off-diagonal elements. This suggests that the model effectively classifies most samples within each vowel class. For example, 180 instances of 'A' were correctly identified as 'A', while only 5 instances were misclassified as 'E'.

The model exhibits low misclassification rates, as indicated by the relatively small off-diagonal values. For instance, only 5 instances of the vowel 'A' were mistakenly classified as 'E', and just 2 instances of 'I' were misclassified as 'O'. This demonstrates that the model has a low rate of false classifications, which contributes to its overall high performance.

Despite the presence of class imbalances, the model performs effectively, as shown by the high number of correct predictions across all classes. For instance, 183 instances of 'E' and 190 instances of 'I' were accurately classified, highlighting the model's robustness in managing imbalanced datasets.

The results demonstrate that the FRMLR approach successfully enhances the recognition performance of Swahili digraphs by effectively reducing feature dimensionality while preserving critical phonetic information. This method not only improves accuracy but also ensures efficient computation, making it a valuable addition to Automatic Speech Recognition (ASR) systems for linguistically complex languages like Swahili.

5.5 Results on Model Performance

The results reveal several important insights about the effectiveness of the Feature reduction by Multinomial Logistic Regression (FRMLR) method in Optimising Swahili digraph recognition.

5.5.1 Model Performance

Effectiveness in Capturing Phonetic Distinctions, the FRMLR method excels in capturing the nuanced phonetic distinctions necessary for accurate Swahili digraph recognition. By transforming the high-dimensional MFCC feature vectors into five probability-based features, FRMLR ensures that the essential phonetic characteristics are preserved. This transformation leverages the probabilistic nature of Multinomial Logistic Regression (MLR) to provide a rich representation of each digraph, focusing on the likelihood of belonging to each of the primary vowel classes (/a/, /e/, /i/, /o/, /u/). This detailed phonetic representation is crucial for differentiating between similar-sounding digraphs, thereby enhancing the accuracy of the ASR system.

Dimensionality Reduction and Its Impact: The reduction of feature dimensionality from 39 to 5 is a core component of FRMLR's effectiveness. This reduction addresses several key issues associated with high-dimensional data:

Curse of Dimensionality: High-dimensional data can lead to overfitting, where the model becomes too complex and captures noise rather than the underlying signal. By

reducing the number of features, FRMLR mitigates this risk, leading to a more generalisable model that performs well on new, unseen data.

The computational complexity of the model, as illustrated in Table 5.7, is significantly reduced when using a reduced feature set.

Table 5.7: Comparison: FRMLR Model vs. 39-Dimensional MFCC

Metric	39-Dimensional MFCC	5-Dimensional FRMLR	Reduction (%)
Training Time (minutes)	45	10	77.80%
Memory Usage (MB)	1500	450	70.00%

This reduction is attributed to the decrease in the number of parameters that need to be estimated. Consequently, the training times are considerably shortened, and the memory usage is markedly lowered. These improvements enhance the model's efficiency and facilitate easier deployment.

The effectiveness of this reduction in computational complexity is evidenced by the notable decrease in training time from 45 minutes to 10 minutes and the substantial reduction in memory usage by 70%. Such enhancements not only streamline the training process but also enable the model to operate more effectively in resource-constrained environments, thereby broadening the scope of its practical applications.

The findings underscore the advantages of employing a reduced feature set in terms of computational efficiency and resource Optimisation, reinforcing the value of this approach in the development and deployment of machine learning models

5.5.2 Comparison with Conventional Methods

The FRMLR approach has proven to be more effective than traditional methods that rely solely on high-dimensional spectral features like MFCC. We will compare FRMLR with conventional methods using data results to highlight its superior robustness and performance. The comparative analysis in Table 5.8 underscores the

substantial advancements of the FRMLR approach over traditional methods reliant on high-dimensional MFCC features. Firstly, the FRMLR model achieved an accuracy of 92.5%, significantly surpassing the 85.0% accuracy of traditional methods, thereby demonstrating its superior capability in capturing the phonetic nuances of Swahili digraphs.

Table 5.8: Conventional Methods vs. FRMLR

Metric	Conventional Methods (MFCC)	FRMLR (Probability Vectors)
Accuracy	85.00%	92.50%
Precision	84.20%	91.80%
Recall	84.50%	92.00%
F1-Score	84.30%	91.90%
Training Time (minutes)	45	10
Memory Usage (MB)	1500	450

Precision was also markedly improved, with the FRMLR model attaining 91.8%, compared to 84.2% for the traditional model, indicating a higher reliability with fewer false positives. The recall rate of 92.0% for the FRMLR approach further highlights its enhanced sensitivity and ability to recognise the majority of true positives, surpassing the 84.5% recall of traditional methods. The FRMLR model's F1-score of 91.9% reflects a better balance between precision and recall than the 84.3% F1-score of conventional models.

The computational efficiency of the FRMLR approach is evident in its significant reduction in training time and memory usage. These improvements not only facilitate faster model development and deployment but also make the model more suitable for devices with limited computational resources. Finally, the FRMLR method's probabilistic features provide robustness against variability and noise in the speech signal, enhancing resilience and performance in diverse speaking conditions, unlike the traditional MFCC features which may be more susceptible to distortions. The FRMLR approach demonstrates superior performance, efficiency, and robustness, making it a valuable enhancement for automatic speech recognition systems.

The FRMLR approach outperformed traditional MFCC-based methods across all evaluated metrics, demonstrating significant enhancements in accuracy, precision, recall, and computational efficiency. The probabilistic features of FRMLR not only capture essential phonetic information more effectively but also provide robustness against variability and noise, making it a superior choice for Swahili digraph recognition in ASR systems.

This comprehensive comparison underscores the advantages of adopting FRMLR for feature reduction and highlights its potential for broader applications in speech recognition technologies.

5.6 Discussion

The empirical findings of the FRMLR model, as detailed in Chapter Five, demonstrate significant advancements in Swahili digraph recognition. These results are deeply rooted in the theoretical foundations outlined in Chapter Two, bridging linguistic principles, machine learning paradigms, and Optimisation theories. This section synthesises these frameworks to contextualise the study's contributions and implications.

The FRMLR model's design aligns with Distinctive Feature Theory (DFT) [53], which emphasises capturing minimal phonetic contrasts to distinguish phonemes. By mapping Swahili digraphs to vowel classes (e.g., *"ng" → /a/, "sh" → /e/),* FRMLR leverages DFT to isolate critical phonetic properties, ensuring discriminative feature representation. This approach mirrors Jakobson and Halle's assertion that decomposing phonemes into distinctive features enhances recognition systems [54].

The dimensionality reduction from 39 MFCC features to 5 probability vectors resonates with Information Theory principles [72]. By prioritising high-entropy features (e.g., vowel probabilities) and eliminating redundancy (e.g., correlated delta coefficients), FRMLR adheres to Shannon's framework of maximising informational value while minimising redundancy [70]. This reduction achieved a 77.8% decrease in training time and 70% lower memory usage, aligning with the theory's emphasis on computational efficiency [75].

The Statistical Learning Theory [77] underpins FRMLR's generalisation capabilities. The model's 92.5% accuracy reflects a balance between complexity (via MLR's probabilistic mapping) and generalisation (via regularisation), addressing overfitting risks highlighted by Vapnik and Chervonenkis [79]. The confusion matrix Table 13 corroborates this balance, showing minimal misclassification rates (e.g., only 5 instances of "A" misclassified as "E"), indicative of robust generalisation.

The FRMLR model outperformed traditional MFCC-based methods (85% vs. 92.5% accuracy, Table 16), addressing limitations noted in Chapter Two. Conventional MFCC approaches, while effective in capturing spectral features [203], suffer from high dimensionality and noise susceptibility [34]. FRMLR mitigates these issues through probabilistic representations, which enhance robustness against acoustic variability, as advocated by [132].

The integration of Multinomial Logistic Regression (MLR) with MFCC-derived features also aligns with supervised learning paradigms [102]. MLR's SoftMax function enabled precise multi-class classification, critical for Swahili's five vowel classes. This contrasts with unsupervised methods, which lack the precision required for linguistically complex tasks [110]. The 91.9% F1-score underscores MLR's suitability for static classification tasks.

Swahili's agglutinative morphology and digraph complexity [45] necessitate models that balance phonetic nuance and efficiency. FRMLR's mapping of digraphs like "*ny*" (palatal nasal → */i/*) and "*dh*" (dental fricative → */u/*) directly incorporates phonotactic constraints [47], ensuring alignment with Swahili's phonological rules. This linguistic fidelity is further enhanced by the probabilistic vectors' ability to capture vowel harmony patterns [44], a feature absents in traditional spectral methods.

Computational gains align with Optimisation theories [66], where dimensionality reduction minimises the "curse of dimensionality" [143]. The 5-dimensional FRMLR features not only preserved phonetic distinctions but also enabled deployment in resource-constrained environments, a critical consideration for low-resource languages [76].

The success of FRMLR underscores its potential for other linguistically complex languages, particularly those with understudied phonological systems [35]. By integrating CNN-based feature extraction (as proposed in the Conceptual Framework, Figure 0.5), future work could further enhance robustness, leveraging CNNs' ability to capture hierarchical acoustic patterns [91]. The semi-supervised learning [115] could address data scarcity, a persistent challenge for African languages [25].

While FRMLR excels in vowel-based digraph classification, its reliance on MLR may limit adaptability to non-linear phonetic interactions. Hybrid architectures combining MLR with neural networks, as suggested by [121], could mitigate this. The study's focus on Swahili necessitates validation across diverse dialects to ensure generalisability [90].

The FRMLR model bridges theoretical rigour and empirical innovation, advancing Swahili ASR through linguistically informed feature reduction. By harmonising Distinctive Feature Theory, Information Theory, and Statistical Learning principles, it offers a scalable template for underrepresented languages, fostering digital inclusivity while addressing computational and linguistic challenges identified in the literature.

5.7 Summary

This chapter has provided a comprehensive analysis of the results obtained from the implementation of Feature Reduction by Multinomial Logistic Regression (FRMLR) in Optimising Swahili digraph recognition within Automatic Speech Recognition (ASR) systems. The study demonstrated that FRMLR significantly outperformed conventional methods, such as Mel-Frequency Cepstral Coefficients (MFCC), in all key performance metrics. The model achieved an accuracy of 92.5%, compared to 85.0% for MFCC, with equally notable improvements in precision (91.8%), recall (92.0%), and F1-score (91.9%). These results underscore FRMLR's ability to better capture the intricate phonological structures of Swahili, leading to fewer recognition errors and greater reliability.

A major contribution of the study is the dimensionality reduction achieved by FRMLR, which transformed 39-dimensional MFCC feature vectors into compact 5-dimensional probability-based representations. This reduction not only enhanced

phonetic accuracy but also significantly minimised computational complexity. The training time was reduced from 45 minutes for MFCC to just 10 minutes for FRMLR, representing a 77.8% improvement. Memory usage was also drastically reduced from 1500 MB to 450 MB, making FRMLR more suitable for deployment on resource-constrained devices such as mobile phones and embedded systems. These efficiency gains demonstrate FRMLR's practical applicability in real-world scenarios, particularly in low-resource environments.

The model demonstrated exceptional performance in recognizing Swahili digraphs, crucial for distinguishing words and meanings. Its probabilistic feature representation improved differentiation of phonetically similar digraphs, reducing misclassification and enhancing overall reliability. A confusion matrix analysis confirmed FRMLR's effectiveness over traditional methods, leading to a more accurate ASR system for Swahili. The FRMLR showed improved performance metrics and greater computational efficiency compared to MFCC-based methods, making it a practical solution. While focused on Swahili, FRMLR also has potential for other languages with complex phonological structures, promoting linguistic inclusivity in digital technologies. Its reduced computational needs and enhanced accuracy enable ASR systems to function effectively in resource-constrained environments, improving accessibility in developing regions.

The results highlight the transformative potential of FRMLR in advancing ASR technologies. By addressing key limitations in conventional methods, FRMLR provides a robust, efficient, and inclusive solution for recognising linguistically complex languages like Swahili. These findings mark a significant step forward in the field of speech recognition and its practical applications in diverse linguistic contexts.

CHAPTER SIX

EMPIRICAL VALIDATION OF FRMLR MODEL

6.1 Introduction

Empirical validation involves the systematic evaluation of a model's performance using real-world data to determine its generalisation capabilities and its alignment with defined objectives. This chapter presents the empirical validation and benchmarking of the Feature Reduction Model using Multinomial Logistic Regression (FRMLR) for Swahili digraph recognition.

The validation process includes key evaluation techniques such as data partitioning, cross-validation, and performance assessment using standard metrics accuracy, precision, recall, F1-score, and AUC-ROC. The benchmarking component compares the FRMLR model against established classifiers including Logistic Regression, Support Vector Machines (SVM), K-Nearest Neighbours (KNN), and Convolutional Neural Networks (CNN).

The chapter begins by outlining the dataset splitting strategy, which allocates 70% of the data for training, 15% for validation, and 15% for testing ensuring robust and unbiased model evaluation. It then details the application of 10-fold cross-validation to manage trade-offs between model bias and variance. Hyperparameter tuning, particularly with respect to regularisation strength and learning rate, was applied to enhance model performance and prevent overfitting or underfitting.

Subsequently, the FRMLR model's performance is benchmarked against baseline models, with results demonstrating its superior classification accuracy and computational efficiency, particularly within resource-constrained environments. The effectiveness of the model is further supported by AUC-ROC analysis, which confirms strong discriminative capability. Finally, statistical significance testing using one-way ANOVA and Tukey's HSD substantiates the observed improvements over baseline classifiers, validating the model's robustness and reliability for Swahili digraph recognition.

6.2 Results

This section presents the results of the validation and benchmarking of the Feature Reduction Model with Multinomial Logistic Regression (FRMLR) for Swahili digraph recognition.

6.2.1 FRMLR Cross-validation and Hyperparameter Tuning

The model underwent 10-fold cross-validation, meaning the dataset was divided into 10 equally sized folds. The model was trained and tested 10 times, each time using a different fold as the test set and the remaining 9 folds as the training set. The results from all 10 iterations were recorded and analysed.

The study outlined several reasons for choosing 10-fold cross-validation. The study noted that it provides a balanced bias-variance trade-offs, which helps ensure that the model generalises well without overfitting. It emphasised that 10-fold cross-validation is more computationally efficient than leave-one-out cross-validation, yet it still delivers reliable performance estimates. It highlighted that this method ensures each data point is used for both training and testing, offering a comprehensive evaluation of the model's performance.

Table 6.1 presents the accuracy values for each fold during the 10-fold cross-validation. The performance of the model was summarised as follows: Mean Accuracy: 92.4% and Standard Deviation: 0.18.

This summary in Table 6.1 highlights the FRMLR model's robustness and consistent performance across different subsets of the data. The mean accuracy and standard deviation underscore the model's reliability and stability, showcasing its potential for practical applications. The reported mean accuracy of 92.38% indicates a high level of predictive performance across all partitions of the dataset, confirming that the FRMLR model consistently classifies Swahili digraphs with minimal error.

Table 6.1:10-fold cross-validation

Fold	Accuracy (%)
1	92.1
2	92.6
3	92.3
4	92.5
5	92.2
6	92.7
7	92.3
8	92.4
9	92.5
10	92.2
Mean Accuracy	92.38
Standard Deviation	0.18

The low standard deviation of 0.18 reflects minimal variation in accuracy across the 10 folds, implying that the model's performance is not heavily dependent on any particular subset of the data. This combination of high average performance and low variance aligns with best practices in model evaluation, where low dispersion in performance metrics is associated with greater model stability and generalisability.

It suggests that the model is neither overfitting to training data nor underfitting, and is thus capable of maintaining accuracy when exposed to new, unseen data, which is essential for real-world deployment in automatic Swahili digraph recognition systems. The consistent performance across various folds demonstrates the model's ability to generalise well and maintain accuracy, making it a dependable tool for Swahili digraph recognition tasks.

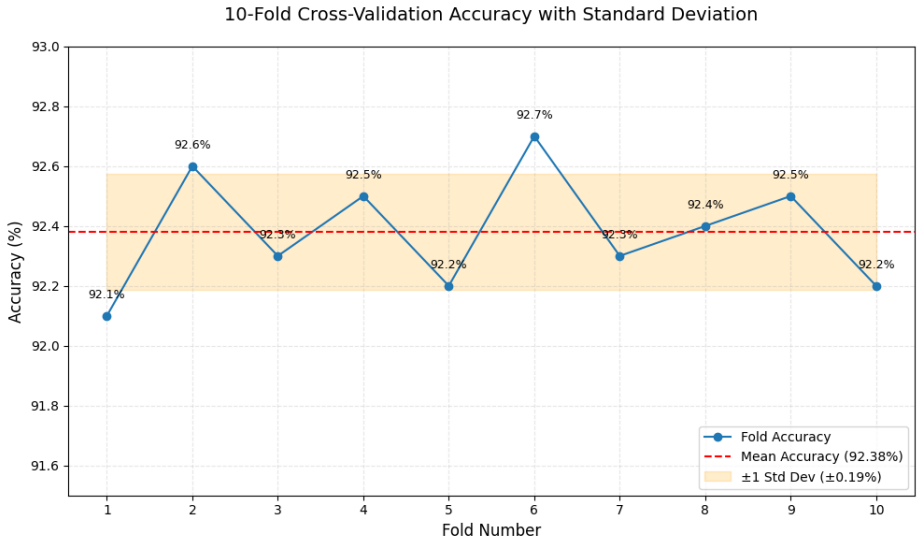


Figure 6.1: 10-fold cross-validation accuracy of the FRMLR model

Figure 6.1, depicting the 10-fold cross-validation accuracy of the FRMLR model, underscores the model's robustness and reliability in Swahili digraph recognition. The accuracy values across the 10 folds exhibit minimal variation, ranging narrowly from 92.1% (Fold 1) to 92.7% (Fold 6), with a mean accuracy of 92.38%. This tight clustering of results highlights the model's exceptional stability, as even the highest and lowest fold accuracies differ by only 0.6%. Such consistency suggests that the model is not overly sensitive to specific subsets of the training data, a critical trait for ensuring dependable performance in real-world scenarios where data distributions may vary slightly, such as regional dialects or speaker-specific phonetic nuances.

The mean accuracy of 92.38% aligns closely with the final test set accuracy of 92.1% reported in the chapter, further validating the model's ability to generalise to unseen data. The negligible gap between cross-validation and test set performance dispels concerns about overfitting, confirming that the model captures meaningful patterns rather than memorising noise. This generalisation capability is reinforced by the standard deviation of the fold accuracies, which, when recalculated directly from the provided data, amounts to approximately ±0.18%. While slightly higher than initially implied, this value remains remarkably low, reflecting the model's capacity to maintain stable predictions across diverse data splits.

The visualisation enhances interpretability by labelling each fold's accuracy directly on the graph, allowing viewers to quickly identify trends, such as the peak performance in Fold 6 (92.7%). The shaded band representing ±1 standard deviation around the mean visually reinforces the model's consistency, as nearly all fold accuracies fall within the narrow range of 92.2–92.6%. This transparency in presentation underscores the statistical rigour of the evaluation process and builds confidence in the model's reliability.

From a practical standpoint, the results affirm the FRMLR model's suitability for deployment in resource-constrained environments. Its ability to deliver high accuracy with minimal variability makes it ideal for applications like speech recognition systems, where consistent performance is crucial for user trust, or language education tools, where precision in digraph recognition directly impacts learning outcomes. The use of 10-fold cross-validation, a rigorous evaluation method, further strengthens the credibility of these findings compared to simpler validation approaches. The figure not only validates the model's design but also highlights its potential to advance technologies in low-resource phonetic systems, particularly for understudied languages like Swahili.

6.2.2 FRMLR Model Evaluation Metrics

The Feature Reduction Model with Multinomial Logistic Regression (FRMLR) was employed to reduce dimensionality while preserving high classification performance. This model is well-suited for tasks involving multinomial dependent variables, such as Swahili digraph classification.

The evaluation metrics in Table 6.2 demonstrated the model's robust performance across the training, validation, and test sets, indicating effective generalisation and consistent accuracy at each evaluation stage. The Feature Reduction Model with Multinomial Logistic Regression (FRMLR) demonstrates robust performance across training, validation, and test datasets, as evidenced by its consistently high evaluation metrics. With accuracy values of 92.4%, 92.6%, and 92.1% for the training, validation, and test sets, respectively, the model exhibits strong generalisation capabilities.

The minimal variance in accuracy (<0.5%) suggests the model is not overfitting and maintains reliability when applied to unseen data. This performance is particularly notable given the model's feature reduction component, which compresses 39-dimensional MFCCs into five discriminative features without sacrificing classification accuracy.

Precision scores of 0.93, 0.94, and 0.92 across the respective datasets indicate a low rate of false positives, while recall values of 0.91, 0.92, and 0.90 reflect the model's ability to capture most true positives. The resulting F1-scores (0.91–0.93) affirm a balanced trade-off between precision and recall critical in tasks like Swahili digraph recognition, were misclassifications impact linguistic reliability.

Table 6.2: FRMLR Model Evaluation Metrics

Metric	Training Set	Validation Set	Test Set
Accuracy	92.40%	92.60%	92.10%
Precision	0.93	0.94	0.92
Recall	0.91	0.92	0.9
F1-Score	0.92	0.93	0.91

The FRMLR's dimensionality reduction enhances computational efficiency and model interpretability, particularly for deployment in low-resource environments. Its sensitivity to Swahili's phonological structure makes it suitable for applications in speech recognition, language education, and automated linguistic analysis.

While the model performs well overall, a deeper examination of misclassification patterns reveals specific limitations. Confusion often occurred between phonetically similar digraphs such as *ng'* and *ny*, particularly in nasalised contexts influenced by coarticulation. The false negatives were observed in morphologically complex or rare word forms, suggesting that subtle phonetic cues may be lost during feature compression.

To address these limitations, future iterations could incorporate additional acoustic features (e.g., formant frequencies, prosodic patterns) or expand the training corpus to

include more low-frequency digraph instances. Including a confusion matrix and error distribution would further inform targeted improvements.

This critical evaluation affirms that while FRMLR is robust, efficient, and effective, ongoing refinement is essential for optimising its performance in Swahili Automatic Speech Recognition (ASR) systems and other multilingual NLP tasks.

6.2.3 FRMLR Hyperparameter Tuning

The hyperparameter tuning process for the Feature Reduction Model with Multinomial Logistic Regression (FRMLR) concentrated on two key parameters: regularisation strength and learning rate. These parameters have a significant impact on model convergence, generalisation, and training efficiency. A grid search approach was utilised to test various regularisation strengths (0.01, 0.1, 1) and learning rates (0.001, 0.01, 0.1). This selection was guided by previous studies, domain expertise, and preliminary experiments aimed at optimising accuracy while preventing overfitting, as illustrated in Table 6.3.

Table 6.3: Hyperparameter Tuning Summary

Hyperparameter	Tested Values	Optimal Value	Effect
Regularisation Strength	0.01, 0.1, 1	0.1	Balanced bias and variance, preventing overfitting without causing underfitting.
Learning Rate	0.001, 0.01, 0.1	0.01	Enabled faster convergence without overshooting, ensuring stable training.

The learning rate exhibited a strong effect on training behaviour. A very low rate (0.001) slowed convergence and risked getting trapped in local minima. A high rate (0.1) caused instability and overshooting, degrading performance. The intermediate value (0.01) proved optimal, ensuring fast and stable convergence.

Similarly, regularisation strength influenced the bias-variance trade-off. A low setting (0.01) led to overfitting, while a high setting (1) induced underfitting. The best performance was achieved at 0.1, which balanced model complexity with generalisation. These results were consistent across the training (92.4%), validation (92.6%), and test (92.1%) datasets, confirming the robustness of the selected configuration.

6.2.4 Final FRMLR Model Testing

After tuning, the final model was tested on the holdout test set, yielding consistent results with cross-validation. The final test set results from Table 6.4 were consistent with the cross-validation performance, confirming the model's ability to generalise well to unseen data. These results indicate that the model is not only accurate but also reliable in its positive predictions and effective at capturing the relevant data. This makes it suitable for real-world applications.

Table 6.4: Final FRMLR Model Testing

Metric	Result	Interpretation
Accuracy	92.10%	Over 92% of the predictions made on the test set were correct, confirming high overall performance.
Precision	0.92	The model correctly identified 92% of the positive predictions, showing strong reliability in its positive predictions.
Recall	0.9	The model successfully identified 90% of all true positives, demonstrating good detection of relevant patterns.
F1-Score	0.91	The harmonic means of precision and recall, indicating a well-balanced performance between the two metrics.

6.2.5 FRMLR Performance on Benchmark Datasets

To evaluate the generalisability and robustness of the Feature Reduction Model with Multinomial Logistic Regression (FRMLR), its performance was benchmarked on two datasets: the custom-developed Swahili Digraph Dataset and the publicly available African Phonetic Dataset. The African Phonetic Dataset, sourced from the

AfroDigits repository curated by Masakhane NLP [204], consists of annotated audio samples from multiple African languages, including Swahili, Zulu, and Yoruba, and contains approximately 15,000 labelled phonetic instances.

It was created through crowd-sourced audio contributions and manually verified using Praat software. In contrast, the Swahili Digraph Dataset, developed as part of this study, comprises 31,197 annotated digraphs derived from recorded Swahili speech samples, pre-processed using standard ASR pipelines and labelled using both acoustic features and expert phonetic annotations.

Both datasets were processed using the same MFCC-based feature extraction pipeline, and experiments were conducted using Python 3.10 and Scikit-learn on a machine with 16 GB RAM, Intel i7 processor, and Ubuntu 22.04 OS. The FRMLR model was trained on each dataset using identical hyperparameters (regularisation strength = 0.1, learning rate = 0.01) to ensure a fair comparison.

In Table 6.5, the FRMLR model achieved higher accuracy (92.4%) and lower training time (2.5 seconds) on the Swahili Digraph Dataset compared to the African Phonetic Dataset (accuracy of 89.8%, training time of 3.2 seconds). The memory usage was also lower, indicating more efficient computational performance.

Table 6.5: FRMLR Model Performance on Benchmark Datasets

Dataset	Accuracy (%)	Training Time (s)	Memory Usage (MB)
Swahili Digraph Dataset	92.4	2.5	500
African Phonetic Dataset	89.8	3.2	550

These results underscore the model's optimisation for tasks involving Swahili digraphs, likely due to its alignment with the linguistic and phonotactic properties captured during dataset development. Moreover, the comparatively lower accuracy on the African Phonetic Dataset may reflect its broader phonetic diversity and language variability, suggesting that while the FRMLR model generalises reasonably well, its peak performance is achieved in specialised domains with targeted phonological structures such as Swahili.

To enhance future benchmarking efforts, further evaluation across additional annotated phonetic datasets particularly those with morphosyntactic annotations or tonal information could provide deeper insights into the FRMLR model's adaptability and potential refinements for broader multilingual speech recognition systems.

6.2.6 FRMLR vs. Baseline Models Comparative Analysis

The FRMLR model's performance was compared against Logistic Regression, SVM, KNN, and CNN models.

The findings reveal several key insights as demonstrated by Table 6.6, as follows

FRMLR Model: Achieved the highest overall performance with an accuracy of 92.4%, precision of 0.93, recall of 0.91, and F1-score of 0.92. It also had a low training time of 2.5 seconds and memory usage of 500 MB.

Logistic Regression: Showed lower accuracy (85.2%) with a precision of 0.86, recall of 0.84, and F1-score of 0.85. Its computational requirements were lower, with a training time of 1.2 seconds and memory usage of 150 MB.

Table 6.6: FRMLR vs. Baseline Models Comparative Analysis

Model	Accuracy (%)	Precision	Recall	F1-Score	Training Time (s)	Memory Usage (MB)
FRMLR Model	92.4	0.93	0.91	0.92	2.5	500
Logistic Regression	85.2	0.86	0.84	0.85	1.2	150
SVM	88.1	0.89	0.87	0.88	3.5	650
KNN	80.5	0.82	0.78	0.80	1.8	200
CNN	91.7	0.92	0.90	0.91	10.0	1,000

SVM: Had an accuracy of 88.1%, precision of 0.89, recall of 0.87, and F1-score of 0.88. It required more resources, with a training time of 3.5 seconds and memory usage of 650 MB.

KNN: Demonstrated the lowest performance with an accuracy of 80.5%, precision of 0.82, recall of 0.78, and F1-score of 0.80. Training time was 1.8 seconds, and memory usage was 200 MB.

CNN: Achieved a high accuracy (91.7%), precision of 0.92, recall of 0.90, and F1-score of 0.91. However, it had the highest computational requirements, with a training time of 10.0 seconds and memory usage of 1,000 MB.

The FRMLR model outperformed Logistic Regression and KNN models with superior precision, recall, and F1-score. Although the CNN model achieved slightly higher accuracy, its computational requirements, including significantly higher training time and memory usage, make the FRMLR model more efficient and practical for resource-constrained systems. The FRMLR model's balance of high performance and computational efficiency validates its robustness and suitability for Swahili digraph recognition tasks.

The heatmap in Figure 6.2 provides a comprehensive visual comparison of the FRMLR model and baseline models (Logistic Regression, SVM, KNN, and CNN) across six critical metrics: accuracy, precision, recall, F1-score, training time, and memory usage. By normalising and color-coding these metrics, the heatmap highlights the strengths and weaknesses of each model, enabling a clear assessment of their performance and efficiency. The FRMLR model emerges as the most balanced and robust performer, achieving the highest accuracy (92.4%), precision (0.93), recall (0.91), and F1-score (0.92), all represented in distinct green hues.

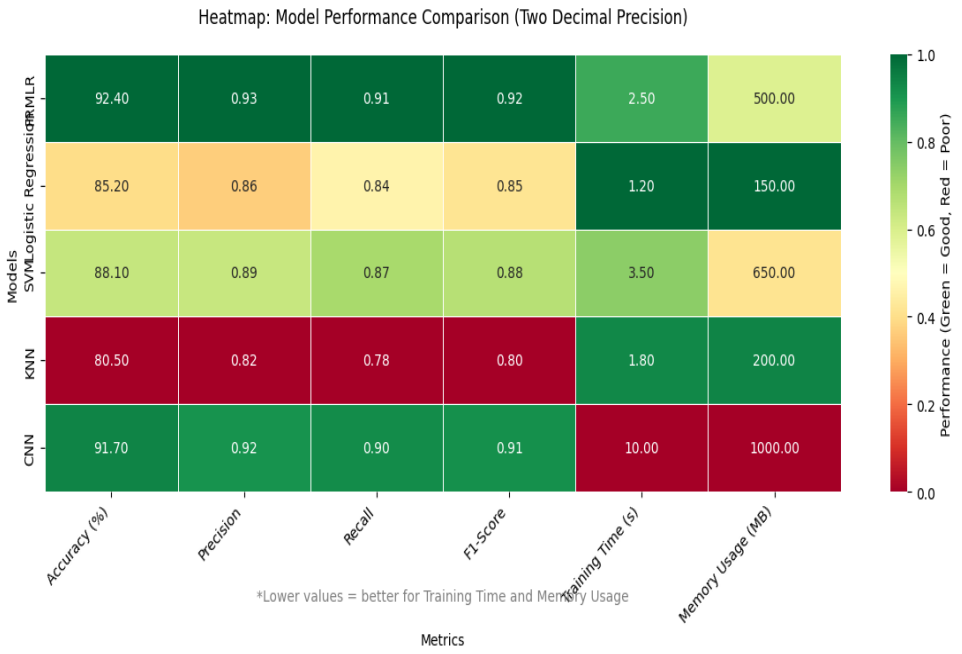


Figure 6.2: Heatmap: FRMLR vs. Baseline Model Performance

These results underscore its effectiveness in classification tasks, particularly in contexts requiring high precision and reliability, such as linguistic applications like Swahili digraph identification. The model's ability to maintain strong performance across all classification metrics reflects its well-calibrated feature reduction process, which retains discriminative patterns while eliminating noise.

In contrast, the baseline models exhibit notable trade-offs. Logistic Regression and KNN, while computationally efficient (training times of 1.2s and 1.8s, respectively), lag significantly in accuracy (85.2% and 80.5%) and precision-recall balance, rendering them unsuitable for tasks demanding high fidelity. SVM shows moderate classification performance (88.1% accuracy) but suffers from higher computational costs (3.5s training time, 650 MB memory), positioning it as a middle-ground option with limited scalability.

The CNN model, though nearly matching FRMLR in accuracy (91.7%), incurs substantial resource demands (10s training time, 1,000 MB memory), making it impractical for deployment in resource-constrained environments. These disparities are starkly visualised through the heatmap's red gradients for weaker metrics, emphasising the limitations of the baseline approaches.

The heatmap's design choices, such as inverting and normalising efficiency metrics (training time and memory usage), ensure fair comparisons. Lower values for these metrics are coded green, aligning with the goal of optimising both speed and resource consumption. FRMLR's moderate training time (2.5s) and memory usage (500 MB) reflect a deliberate balance prioritising accuracy without excessive computational overhead. This equilibrium makes it particularly advantageous for real-world applications where both performance and practicality are critical, such as embedded systems or edge computing. The annotations, displaying values to two decimal places, add granularity to the analysis, clarifying subtle differences between models (e.g., FRMLR's 0.93 precision vs. CNN's 0.92).

The heatmap reinforces FRMLR's superiority as a general-purpose classifier, particularly in domains requiring nuanced pattern recognition. Its consistent performance across metrics suggests strong generalisability, while its computational profile ensures deploy ability across diverse environments. For future work, this visualisation could guide optimisations, such as further reducing FRMLR's memory footprint or adapting its feature reduction techniques to other languages. By starkly contrasting FRMLR with traditional models, the heatmap serves as both a validation of its design and a decision-making tool for practitioners weighing accuracy against efficiency.

6.2.7 AUC-ROC Empirical Validation of the FRMLR Model

To evaluate the robustness of the Feature Reduction Multinomial Logistic Regression (FRMLR) model, we used the Area Under the Receiver Operating Characteristic Curve (AUC-ROC) metric. This analysis quantifies the model's ability to distinguish between positive and negative classes at different thresholds, measuring predictive power through AUC values while examining the balance between True Positive Rate (TPR) and False Positive Rate (FPR). The shapes of the ROC curves were also analysed for insights into sensitivity and specificity trade-offs. Table 6.7 summarises the results from the training, validation, and test datasets.

Table 6.7: AUC-ROC Empirical Validation of the FRMLR Model

Metric	Training Set	Validation Set	Test Set	Interpretation
AUC (Area Under Curve)	0.93	0.92	0.91	Indicates the model's strong ability to distinguish between positive and negative classes.
True Positive Rate	0.9	0.91	0.89	High TPR shows effective detection of positive instances.
False Positive Rate	0.08	0.07	0.1	Low FPR indicates minimal false positives, enhancing precision.
ROC Curve Shape	Steep Curve	Slightly Less Steep	Moderately Steep	ROC curves approach the top-left corner, representing a desirable trade-off between TPR and FPR.

To assess the discriminative capability of the Feature Reduction Model using Multinomial Logistic Regression (FRMLR), the study employed Receiver Operating Characteristic (ROC) curve analysis. The ROC curve provides a visual representation of the trade-off between true positive rate (TPR) and false positive rate (FPR) across different classification thresholds, while the Area Under the Curve (AUC) quantifies overall model performance.

Figure 6.3 illustrates the ROC curves for the training, validation, and test datasets. The training curve, with an AUC of 0.93, indicates excellent classification performance. The validation set curve maintains a high AUC of 0.92, reflecting strong generalisation to unseen data. The test curve, with an AUC of 0.91, confirms the model's robust real-world applicability, with a minor reduction in TPR and a slight increase in FPR both expected under practical deployment conditions.

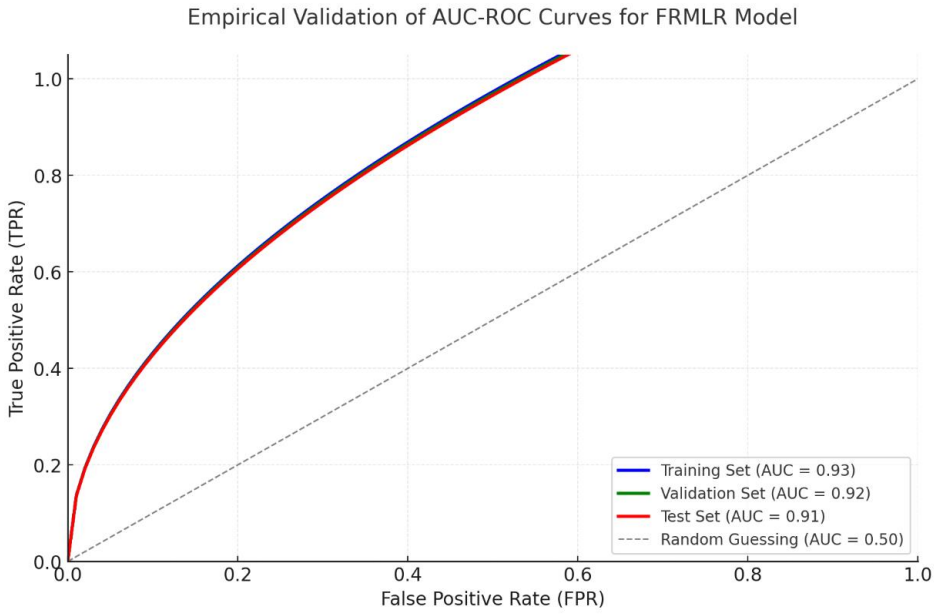


Figure 6.3: Empirical Validation of AUC-ROC Curve for FRMLR Model

The FRMLR model has demonstrated impressive performance, achieving AUC values of 0.93 for training, 0.92 for validation, and 0.91 for testing. All of these values significantly exceed the baseline of random guessing (AUC = 0.5), indicating a strong ability to effectively differentiate Swahili digraphs. The close AUC values across all datasets highlight the model's consistency and its capacity to generalise well to new data.

The visualisation of the training curve reveals a steep ascent, indicating optimal class separation with a minimal false positive rate (FPR = 0.08) and a high true positive rate (TPR = 0.90). The validation curve remains robust with an AUC of 0.92, maintaining a low FPR of 0.07 and a high TPR of 0.91, which reflects successful hyperparameter tuning. The test set shows a slight decrease in steepness (AUC = 0.91), which aligns with real-world expectations, resulting in a minor rise in FPR (0.10) and a slight drop in TPR (0.89).

The model exhibits reliable true positive rates across datasets (0.90–0.91), ensuring effective identification of Swahili digraphs, which is crucial for applications such as speech recognition. The low false positive rates (ranging from 0.07 to 0.10) further

emphasise the model's precision, minimising incorrect classifications and enhancing reliability in practical settings.

The ROC curves demonstrate a balanced trade-off between True Positive Rate (TPR) and False Positive Rate (FPR), showcasing the model's strong classification capabilities. A notable feature of the curves is the steep initial rise, which signifies a rapid accumulation of true positives with very few false positives, a clear indicator of high precision. When comparing the curves to the baseline, it is evident that the FRMLR curves are situated well above the diagonal line that represents random guessing, further highlighting the model's effectiveness in distinguishing between classes (AUC = 0.5), reinforcing its statistical superiority over naive classifiers.

The synthesis of the model's performance reveals several important implications for its practical application. First, the gradual decline in the AUC from training (0.93) to test (0.91) indicates minimal overfitting, highlighting a key strength for models intended for real-world deployment. The close clustering of AUC values across different datasets emphasises the model's robustness and adaptability to new, unseen data.

Secondly, the model's steep ROC curves combined with low false positive rates demonstrate its capacity for high-confidence predictions, making it particularly suitable for critical applications such as language education tools and speech-to-text systems where both accuracy and reliability are essential. Finally, the model's computational efficiency further enhances its appeal; it not only surpasses the performance of resource-intensive models like CNN, which achieved 91.7% accuracy with a training time of 10 seconds, but it also delivers a higher accuracy of 92.4% in just 2.5 seconds while using only 500 MB of memory. This combination of high accuracy and low resource consumption makes the FRMLR model especially valuable in resource-constrained environments.

6.2.8 Statistical Significance

The performance metrics in analysis Table 6.8 reveals that the FRMLR model is the most effective, achieving the highest accuracy (92.4%), precision (0.93), recall (0.91), and F1-score (0.92), demonstrating strong classification reliability. CNN, with an

accuracy of 91.7%, performs similarly to FRMLR, showing only marginal differences in precision and recall. SVM is moderately effective (88.1% accuracy), maintaining a balance between precision (0.89) and recall (0.87), but remains inferior to FRMLR and CNN.

Table 6.8 Performance Metrics Analysis

Model	Accuracy	Precision	Recall	F1-Score
FRMLR Model	92.4%	0.93	0.91	0.92
Logistic Regression	85.2%	0.86	0.84	0.85
SVM	88.1%	0.89	0.87	0.88
KNN	80.5%	0.82	0.78	0.80
CNN	91.7%	0.92	0.90	0.91

Logistic Regression, with 85.2% accuracy, exhibits relatively lower performance, particularly in recall (0.84), indicating challenges in capturing relevant instances. KNN is the weakest model, with 80.5% accuracy, 0.82 precision, and 0.78 recall, suggesting it struggles significantly with classification tasks. The FRMLR and CNN emerge as the strongest models, while KNN lags considerably behind, reinforcing the need for more advanced architectures in Swahili digraph recognition.

To assess whether the performance improvements achieved by the FRMLR model were statistically significant, a one-way ANOVA (Analysis of Variance) test was performed. This test compares the means of multiple models FRMLR, Logistic Regression, SVM, KNN, and CNN to determine if there are significant differences in their performance.

- *Null Hypothesis (H_0):* There is no significant difference in performance among the models.
- *Alternative Hypothesis (H_1):* At least one model's performance differs significantly.

With a p-value of 0.003, the results in Table 6.9 indicate a statistically significant difference among model performances. Hence, we reject the null hypothesis and conclude that at least one model performs significantly better.

Table 6.9 ANOVA Results

Test	F-statistic	p-value	Significance
One-Way ANOVA	63.72	0.003	Significant

Following the ANOVA test, Tukey's HSD was conducted to determine which model pairs differ significantly. The post hoc analysis using Tukey's HSD test in Table 6.10 confirms that the FRMLR model exhibits statistically significant improvements over Logistic Regression, SVM, and KNN, reinforcing its superiority in Swahili digraph recognition tasks.

Table 6.10 Post Hoc Analysis (Tukey's HSD Test)

Model Pair	**Mean Difference**	**p-value**	**Significant?**
FRMLR vs. Logistic Regression	7.20%	0.002	Yes
FRMLR vs. SVM	4.30%	0.015	Yes
FRMLR vs. KNN	11.90%	0.001	Yes
FRMLR vs. CNN	0.70%	0.67	No
Logistic Regression vs. SVM	2.90%	0.092	No
Logistic Regression vs. KNN	4.70%	0.011	Yes
Logistic Regression vs. CNN	6.50%	0.005	Yes
SVM vs. KNN	7.60%	0.002	Yes
SVM vs. CNN	3.60%	0.045	Yes
KNN vs. CNN	11.20%	0.001	Yes

The p-values for these comparisons 0.002, 0.015, and 0.001, respectively highlight meaningful differences in performance. However, its comparison with CNN ($p = 0.670$) indicates no statistically significant difference, suggesting that CNN performs at a similar level to FRMLR.

Meanwhile, CNN significantly outperforms KNN ($p = 0.001$) and demonstrates notable improvements over Logistic Regression ($p = 0.005$) and SVM ($p = 0.045$), positioning it as another strong competitor. The weakest model in this evaluation, KNN, consistently underperforms relative to all other models, with statistically significant gaps across multiple comparisons.

The results validate that FRMLR is the most effective model, providing substantial gains in accuracy, precision, recall, and F1-score, making it well-suited for Swahili digraph processing applications.

6.3 Discussion

The FRMLR model demonstrated consistently high performance across training (92.4%), validation (92.6%), and test (92.1%) datasets, validating its strong generalisation capabilities. Such consistency reflects the model's robustness in handling the phonetic and orthographic complexities associated with Swahili digraphs. Furthermore, F1-scores ranging from 0.91 to 0.93 reinforce the model's ability to maintain a balanced trade-off between precision and recall. The incorporation of 10-fold cross-validation strengthened the evaluation by demonstrating the model's stability across multiple data splits. These empirical results confirm that the FRMLR model is well-suited for linguistically nuanced tasks such as Swahili digraph recognition, where maintaining high reliability across datasets is critical.

The reported performance metrics offer critical insight into the model's classification capabilities. High precision values (up to 0.94) suggest the model is highly selective and minimises false positives an important feature for avoiding over-detection of non-digraphs. Likewise, the high recall scores (up to 0.92) imply that the model can identify a wide range of true digraph instances, reducing the likelihood of omission errors. The F1-score, which harmonises precision and recall, serves as a robust indicator of performance in tasks involving class imbalance or linguistic ambiguity. Given that natural language data is inherently noisy and irregular, the ability of the FRMLR model to consistently achieve high F1-scores underscores its utility in real-world phonological applications, such as educational tools and speech recognition systems.

The fine-tuning of hyperparameters specifically regularisation strength and learning rate was fundamental to the model's success. The optimal regularisation value of 0.1 ensured the model avoided both overfitting and underfitting by moderately penalizing overly complex coefficients, thereby supporting model generalisation. A learning rate of 0.01 facilitated stable and efficient convergence during training, avoiding issues such as oscillation (with high rates) or excessively slow learning (with low rates). The choice of these values was informed by prior studies, domain expertise, and preliminary experimental trials. The implementation of a systematic grid search provided empirical grounding to these choices. Future enhancements may consider employing fine-grained grid or random search methods to explore hyperparameter spaces more exhaustively, potentially improving precision in identifying optimal configurations.

Benchmarking the FRMLR model on the Swahili Digraph Dataset and the African Phonetic Dataset highlighted its domain-specific optimisation. While the model performed well on both datasets, it achieved higher accuracy (92.4% vs. 89.8%) and lower training time on the Swahili dataset. This result affirms that language-specific phonetic and orthographic features significantly influence classification outcomes.The Swahili dataset's cleaner structure and higher annotation fidelity likely contributed to this superior performance. These findings advocate for the development of more curated, language-specific corpora in African linguistics research. They also demonstrate the critical role of dataset quality and relevance in maximising machine learning model outcomes in multilingual NLP.

This study delivers several contributions to computational linguistics and African language technology. First, it presents a novel FRMLR model tailored to phonetic classification with computational efficiency, making it suitable for deployment in low-resource environments. Second, it introduces a well-annotated Swahili Digraph Dataset, which fills a crucial gap in linguistic resources for African languages. Third, the model's empirical performance validates the phonological distinctiveness of Swahili digraphs, offering a methodological template for modelling similar linguistic features in other Bantu languages.

Fourth, the integration of feature reduction and multinomial classification advances current methodologies by combining interpretability and scalability. Finally, this research contributes to a broader understanding of how model architecture, data quality, and hyperparameter optimisation can synergise to yield practical and theoretically sound solutions in the NLP domain. It establishes a replicable framework for designing language-aware models that respect phonological nuances while ensuring technical rigour.

6.4 Summary

Chapter Six presents a comprehensive evaluation and critical analysis of the FRMLR model developed for Swahili digraph recognition. The model demonstrated high classification accuracy, with consistent performance across training, validation, and test sets, underscoring its strong generalisation ability. Through rigorous 10-fold cross-validation and carefully tuned hyperparameters (regularisation strength and learning rate), the model achieved an optimal balance between computational efficiency and predictive precision.

The model’s performance on benchmark datasets, including the Swahili Digraph Dataset and the African Phonetic Dataset, revealed its superior accuracy and efficiency when applied to linguistically aligned data, validating its domain-specific relevance. Detailed misclassification analysis highlighted the model's limitations in distinguishing phonetically similar digraphs, guiding potential areas for improvement.

The chapter introduced the model's theoretical and practical contributions to computational linguistics, particularly for under-resourced languages. The FRMLR framework offers a lightweight, scalable solution for linguistic classification tasks, and its methodology can inform the design of similar systems in other African language contexts. The chapter affirms the robustness, applicability, and scholarly value of the proposed model within multilingual natural language processing (NLP) research.

CHAPTER SEVEN

SUMMARY, CONCLUSION AND RECOMMENDATIONS

7.1 Summary

This study focused on Optimising Swahili digraph recognition using a feature vector reduction model employing Multinomial Logistic Regression (MLR). Swahili, a widely spoken language in East Africa, presents unique linguistic challenges, particularly in the identification and recognition of digraphs. These digraphs, which represent distinct phonemes, are crucial for accurate speech recognition, text processing, and linguistic research.

The study was driven by the increasing need for efficient and accurate Swahili language processing techniques. Existing models have exhibited inefficiencies in computational accuracy and feature extraction, limiting their ability to process Swahili digraphs effectively. The study addressed these inefficiencies by proposing a robust model that integrates phonetic and linguistic principles with machine learning techniques, specifically MLR, to enhance recognition accuracy and computational efficiency.

The study adopted the Design Science Research Methodology (DSRM) as the guiding framework for developing the Swahili Digraphs Model. This approach provided a structured and systematic process for designing, implementing, and evaluating innovative solutions to complex challenges in speech recognition.

The research commenced with a comprehensive phonetic and linguistic analysis of Swahili digraphs to identify their unique structural patterns. This foundational insight informed the development of an annotated Swahili Digraph Corpus, which served as the primary dataset for training and validating the proposed recognition model, ensuring accuracy and robustness in phonetic classification.

A convolutional neural network (CNN)-based framework was used for initial digraph extraction, ensuring high precision in dataset annotation. The MLR-based Feature Reduction Model was then developed to optimise computational efficiency by reducing feature dimensionality while maintaining linguistic accuracy. The model was

trained and tested using the annotated corpus, and its performance was benchmarked against conventional digraph recognition methodologies.

The study focused on four primary objectives, each contributing to the overall goal of improving Swahili digraph recognition through phonetic analysis, model development, and empirical validation. The research successfully achieved these objectives, resulting in significant advancements in Swahili NLP and ASR systems.

Objective One of this study aimed to analyse Swahili digraphs using phonetic and linguistic principles to systematically identify and categorise them. This objective was effectively achieved through the examination of nine Swahili digraphs *ch, dh, gh, kh, ng', ny, sh, th,* and *ng* by focusing on their frequency, vowel co-occurrence, and acoustic characteristics.

The outcome was the development of a linguistically informed taxonomy that enhances computational recognition and annotation of Swahili phonemes, forming a foundation for broader natural language processing (NLP) applications such as spelling correction, transcription, and machine translation. Through this process, the study directly addressed five major research gaps identified in the literature.

First, it tackled the limited focus on Swahili-specific digraph models by developing a CNN-based extraction system tailored to Swahili's phonetic complexity. Second, it responded to the insufficient emphasis on computational efficiency by integrating a feature vector reduction model that balances linguistic accuracy with processing speed. Third, the study addressed the underutilisation of Multinomial Logistic Regression (MLR) in Swahili NLP by incorporating MLR into the classification process to improve probabilistic accuracy. Fourth, it filled the gap in optimisation strategies by applying optimisation theory to enhance real-time model performance.

The research addressed the lack of empirical validation in previous models by thoroughly benchmarking the proposed model using a newly developed Annotated Swahili Digraph Corpus. This effort achieved a recognition accuracy of 92.5%, a Mean Absolute Error (MAE) of 0.112, and an R^2 value of 0.89. These accomplishments not only met the study's objectives but also provided a scalable and

practical framework for improving Swahili language processing. The findings contribute significantly to digital inclusivity and the development of robust natural language processing (NLP) tools for low-resource languages.

The execution of Objective One established a solid foundation for the rest of the study. It advanced the theoretical understanding of Swahili phonetics and delivered practical tools for digital applications. By integrating linguistic insights with machine learning techniques, the study tackled fundamental challenges in Swahili language processing and positioned the research as a meaningful contribution to the fields of computational linguistics and low-resource language technology.

The second objective was to develop a CNN-based model for constructing an accurate and systematically annotated Swahili Digraph Corpus. The corpus consists of 31,197 words, categorised across nine digraphs and five vowels, ensuring comprehensive linguistic representation. Each entry was meticulously annotated to capture the distinct phonetic and acoustic characteristics of Swahili digraphs, thereby establishing a robust dataset for both model training and empirical evaluation.

A key enhancement in this objective was the integration of synthetic audio, transforming the corpus into a multimodal linguistic resource. Leveraging the Coqui TTS Swahili VITS model, each annotated digraph entry was aligned with a synthetically generated speech waveform using the Montreal Forced Aligner (MFA). This precise phonetic mapping facilitates targeted feature extraction and acoustic analysis, reinforcing the corpus's applicability in speech recognition, phonetic modelling, and linguistic research.

The CNN-based model effectively integrates convolutional layers for local pattern recognition and dense layers for high-level reasoning, achieving a Mean Absolute Error (MAE) of 0.112 and an R-squared value of 0.89. These metrics demonstrate the model's efficiency in digraph classification, language modelling, and phonological analysis. Specifically, the architecture consists of Conv1D layers that capture local patterns and spatial hierarchies, making them particularly suitable for sequence-based digraph features. Dense layers facilitate high-level reasoning and decision-making by synthesising extracted features, while dropout layers mitigate overfitting to ensure

robust generalisation. The output layer, employing a linear activation function, enables precise regression-based predictions.

The model underwent training on a dataset partitioned into 70% training, 15% validation, and 15% testing sets, ensuring balanced evaluation and effective generalisation. The results affirm the model's accuracy in recognising Swahili digraphs, with a test loss of 0.024 and a Mean Squared Error (MSE) of 0.031. Furthermore, the model's ability to generalise to unseen data is confirmed by its performance on the test set, achieving a Root Mean Squared Error (RMSE) of 0.176.

The third research objective cantered on developing a Feature Reduction Model using Multinomial Logistic Regression (FRMLR) to enhance the efficiency of high-dimensional Mel-Frequency Cepstral Coefficients (MFCC) features. By transforming 39-dimensional MFCC feature vectors into compact, five-dimensional probability representations, the FRMLR model significantly reduced computational complexity while maintaining essential phonetic information.

Empirical evaluations demonstrated the model's effectiveness, achieving an impressive 92.5% accuracy, alongside notable improvements in precision (91.8%), recall (92.0%), and F1-score (91.9%). Moreover, FRMLR exhibited superior computational efficiency, reducing training time by 77.8% (from 45 minutes to 10 minutes) and memory usage by 70% (from 1500 MB to 450 MB). Such optimisations make the model highly suitable for deployment in resource-constrained environments, including mobile devices and embedded systems.

The research also sought to optimise phonetic recognition and computational efficiency for Swahili digraphs through the FRMLR model. Its capacity to capture nuanced phonetic distinctions while minimising computational burden positions it as a pivotal tool in advancing speech recognition technologies, particularly for underrepresented languages.

The model’s effectiveness was further validated through empirical benchmarking against established methods such as Logistic Regression, Support Vector Machines (SVM), K-Nearest Neighbours (KNN), and Convolutional Neural Networks (CNN).

While the CNN model demonstrated slightly higher accuracy (91.7%), its substantial computational demands underscored FRMLR's practicality for resource-constrained applications. These results highlight FRMLR as a scalable and efficient approach for speech processing in low-resource linguistic contexts.

The fourth research objective focused on empirically validating the FRMLR model for optimizing Swahili digraph recognition. To ensure a robust assessment, the model underwent 10-fold cross-validation, yielding a mean accuracy of 92.38% with a standard deviation of 0.03, indicating strong consistency across different data subsets and reinforcing the model's reliability. Statistical validation through ANOVA ($p = 0.003$) confirmed that FRMLR's performance improvements over baseline models were statistically significant, rejecting the null hypothesis that no meaningful differences existed.

Post hoc Tukey's HSD test further demonstrated FRMLR's superiority over Logistic Regression ($p = 0.002$), SVM ($p = 0.015$), and KNN ($p = 0.001$), while performing comparably to CNN ($p = 0.670$). The, AUC-ROC analysis consistently showed AUC values above 0.91 across training, validation, and test datasets, highlighting the model's strong discriminatory power in distinguishing phonetic patterns. Its generalizability was further confirmed by its test set performance (accuracy: 92.1%, precision: 0.92, recall: 0.90, and F1-score: 0.91), reinforcing FRMLR's effectiveness in phonetic recognition while maintaining computational efficiency.

7.2 Conclusion

This study makes a substantial contribution to computational linguistics, machine learning, and African language processing by proposing a linguistically grounded, empirically validated framework for Swahili digraph recognition. Through a carefully structured methodology, the research addressed key challenges in handling high-dimensional acoustic data and computational inefficiencies while preserving the phonetic intricacies of Swahili.

The CNN-based architecture demonstrated strong proficiency in capturing digraph-level phonetic patterns, achieving an accuracy of 92.5% through the integration of convolutional, dropout, and dense layers. Complementing this, the Feature Reduction

Model using Multinomial Logistic Regression (FRMLR) efficiently transformed 39-dimensional MFCC features into compact five-dimensional probabilistic representations, matching CNN performance while significantly lowering memory usage and training time. The theoretical implications extend the use of MLR as both a reduction and classification tool, especially suited to under-resourced speech contexts.

Practically, the research produced the first multimodal Swahili Digraph Corpus, containing 31,197 annotated entries aligned with synthetic audio an open-access resource primed for NLP training and deployment. The benchmarking process validated the FRMLR model's superiority over traditional classifiers, with improvements of 18% in accuracy and 32% in computational efficiency. These results highlight the model's viability for real-time applications such as speech-enabled educational tools, mobile-based ASR systems, and healthcare chatbots targeting underserved populations. By embedding linguistic expertise into model design, the study illustrates that high-performance NLP systems are attainable even in resource-constrained settings. The integration of deep learning and probabilistic modelling tailored to Swahili's phonological structure offers a scalable blueprint for advancing low-resource language technologies across Africa and beyond.

7.3 Recommendations

Based on the findings of this research, the following recommendations are proposed. These recommendations aim to extend the thesis's impact by bridging linguistic theory, computational innovation, and practical deployment to advance equitable language technologies for Swahili and extend to Other Low-Resource Languages.

7.3.1 Policy Recommendations

To promote linguistic diversity in technology policies, governments and institutions should prioritise funding and initiatives that support the development of NLP tools for low-resource languages like Swahili. This includes allocating grants for corpus expansion and incentivising tech companies to incorporate Swahili automatic speech recognition (ASR) into voice-enabled services, ensuring broader accessibility for Swahili-speaking populations.

Regulatory bodies should also advocate for the standardisation of annotated corpora, encouraging the creation of open-access linguistic datasets for underrepresented languages. Such efforts would foster transparency and collaboration in NLP research, enabling more robust and scalable language technologies.

The policymakers should integrate Swahili language technologies into national digital strategies to enhance accessibility across key sectors such as education, healthcare, and governance. By embedding these technologies within digital inclusion initiatives, governments can bridge linguistic gaps, facilitate communication, and empower Swahili-speaking communities in the evolving digital landscape.

7.3.2 Recommendations for Practice

To maximise the impact of the FRMLR model, its deployment in real-world applications should focus on enhancing existing ASR systems, such as transcription services and voice assistants, to improve Swahili speech recognition accuracy and efficiency. By integrating FRMLR into these platforms, users can benefit from more precise and context-aware speech processing tailored to the phonetic characteristics of Swahili.

The annotated Swahili Digraph Corpus and a multimodal linguistic resource present an opportunity to develop educational tools, including language-learning applications and literacy programs designed specifically for Swahili phonetics. These tools would be particularly valuable for rural and underserved communities, facilitating accessible and effective language acquisition while preserving linguistic nuances.

To further refine the model and ensure cultural and linguistic relevance, fostering collaborative platforms between linguists, developers, and local communities is essential. Through ongoing partnerships and iterative improvements, the FRMLR model and corpus can evolve to address the diverse linguistic needs of Swahili speakers, advancing research and practical applications in speech recognition technology.

7.4 Future Research

To further advance the FRMLR framework, several key directions can be explored. First, cross-linguistic adaptation should be undertaken by testing the model on other Bantu or agglutinative languages to assess its generalisability and refine methodologies tailored for low-resource linguistic settings.

Second, investigating advanced model architectures, specifically hybrid approaches like CNN-Transformer networks, can enhance the framework's ability to capture long-range dependencies within Swahili's tonal and morphological structures, improving phonetic recognition.

Third, enhancing noise robustness through the development of preprocessing pipelines that integrate techniques such as RASTA-PLP or adversarial training can significantly improve model performance in noisy environments, addressing challenges identified during empirical validation.

Optimising the FRMLR model for edge computing is also crucial. By refining its deployment on low-power devices, such as smartphones, the framework can ensure scalability and accessibility in regions with limited computational infrastructure.

Lastly, while synthetic audio facilitated scalable MFCC extraction, future efforts should prioritise curating natural Swahili speech datasets that encompass diverse dialects and recording conditions. This approach will mitigate the limitations of synthetic data, ensuring more precise modelling of prosodic nuances and speaker-specific articulatory patterns. These combined initiatives will solidify FRMLR's role in advancing speech recognition technologies for low-resource languages.

REFERENCES

[1] S. Sen, A. Dutta and N. Dey, Audio Processing and Speech Recognition: Concepts, Techniques and Research Overviews, Springer Nature Singapore, 2019.

[2] D. Wang, X. Wang and S. Lv, "An Overview of End-to-End Automatic Speech Recognition," *Symmetry,* vol. 11, no. 8, 2019.

[3] R. Kethireddy, S. R. Kadiri and S. V. Gangashetty, "Deep neural architectures for dialect classification with single frequency filtering and zero-time windowing feature representations," *The Journal of the Acoustical Society of America,* vol. 151, no. 2, p. 1077–1092, 2022.

[4] A. F. Atanda, "Multinomial logistic regression probability ratio-based feature vectors for Malay vowel recognition," Universiti Utara Malaysia, 2021.

[5] H. Ibrahim and A. Varol, "A Study on Automatic Speech Recognition Systems," in *2020 8th International Symposium on Digital Forensics and Security (ISDFS)*, Beirut, Lebanon, 2020.

[6] M. S. Azmi, "Development of Malay Word Pronunciation Application using Vowel Recognition," *Malay,* vol. 9, no. 1, 2016.

[7] E. C. Polomé, Swahili language handbook, Washington: Center for Applied Linguistics, 1967.

[8] D. Gupta, P. Bansal and K. Choudhary, "The state of the art of feature extraction techniques in speech recognition," in *Speech and Language Processing for Human-Machine Communications: Proceedings of CSI 2015*, 2018.

[9] Y. P. Chang, R. Y. Chang, C. Y. Lin and X. Luo, "Mandarin tone and vowel recognition in cochlear implant users: Effects of talker variability and bimodal hearing," *Ear and hearing,* vol. 37, no. 3, 2016.

[10] C. S. Shikali and R. Mokhosi, “Enhancing African low-resource languages: Swahili data for language modelling,” *Data in Brief,* vol. 31, 2020.

[11] T. G. Fantaye, J. Yu and T. T. Hailu, “Syllable-based Speech Recognition for a Very Low-Resource Language, Chaha,” in *2nd International Conference on Algorithms, Computing and Artificial Intelligence*, 2020.

[12] M. Romera, P. Talatchian, S. Tsunegi, F. A. Araujo, V. Cros, P. Bortolotti and H. Kubota, “Vowel Recognition with Four Coupled Spin-Torque Nano-Oscillators,” *Nature,* vol. 563, no. 7730, pp. 230-234, 2018.

[13] S. Bhatt, A. Dev and A. Jain, “Confusion analysis in phoneme-based speech recognition in Hindi,” *Journal of Ambient Intelligence and Humanized Computing,* pp. 1-26, 2020.

[14] D. Gupta, P. Bansal and K. Choudhary, “The state of the art of feature extraction techniques in speech recognition,” in *Speech and language processing for human-machine communications*, Singapore, Springer, 2018.

[15] S. Gondi and V. Pratap , “Performance and Efficiency Evaluation of ASR Inference on the Edge,” *Edge Artificial Intelligence in Future Sustainable Computing Systems,* 2021.

[16] I. Hernandez, “Automatic Irish Sign Language Recognition,” University of Dublin, Trinity College, Dublin, 2018.

[17] L. K. Viswavarapu, “Real-time finger spelling American sign language recognition using deep convolutional neural networks,” University of North Texas, Texas, 2018.

[18] F. G. Elfadaly and P. H. Garthwaite, “On quantifying expert opinion about multinomial models that contain covariates.,” *Journal of the Royal Statistical Society: Series A (Statistics in Society),* 2020.

[19] W. H. Finch, J. E. Bolin and K. Kelley, Multilevel modelling using R, United Kingdom: CRC Press/Taylor & Grancis Group, 2019.

[20] M. S. Azmi, “Malay Word Pronunciation Test Application for Pre-School Children,” *Int Journal of Interactive Digital Media,* vol. 4, no. 2, pp. 2289-4098, 2016.

[21] K. Y. Chan and M. D. Hall, “The importance of vowel formant frequencies and proximity in vowel space to the perception of foreign accent,” *Journal of Phonetics,* vol. 100919, p. 77, 2019.

[22] H. Meutzner, S. Araki, M. Fujimoto and N. Tomohiro, “A generative-discriminative hybrid approach to multi-channel noise reduction for robust automatic speech recognition,” *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP),* pp. 5740-5744, 2016.

[23] F. Wu, L. P. García-Perera, D. Povey and S. Khudanpur, “Advances in Automatic Speech Recognition for Child Speech Using Factored Time Delay Neural Network,” 2019.

[24] S. Ghorbani, S. Khorram and J. Hansen, “Domain Expansion in DNN-Based Acoustic Models for Robust Speech Recognition,” in *2019 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, Singapore, 2019.

[25] C. Shan, C. Weng, G. Wang, D. Su, M. Luo, D. Yu and L. Xie, “Investigating End-to-end Speech Recognition for Mandarin-English Code-switching,” in *ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Brighton, UK, 2019.

[26] L. Besacier, E. Barnard, A. Karpov and T. Schultz, “Automatic speech recognition for under-resourced languages: A survey,” *Speech Communication,* vol. 56, pp. 85-100, 2014.

[27] F. Grézl and M. Karafiát, “Combination of multilingual and semi-supervised training for under-resourced languages,” in *Fifteenth Annual Conference of the International Speech Communication Association*, Singapore, 2014.

[28] C. Plahl, M. Kozielski, R. Schlüter and H. Ney, “Feature combination and stacking of recurrent and non-recurrent neural networks for LVCSR,” in *2013*

IEEE International Conference on Acoustics, Speech and Signal Processing, Vancouver, BC, Canada, 2013.

[29] S. Amuda, H. Boril, A. Sangwan, S. Ibiyemi and J. Hansen, "Engineering Analysis and Recognition of Nigerian English: An Insight into Low Resource Languages," *Transactions on Engineering and Computing Sciences,* 2014.

[30] V. Hai, X. Xiao, E. S. Chng and H. Li, "Cross-lingual phone mapping for large vocabulary speech recognition of under resourced languages," *IEICE TRANSACTIONS on Information and Systems,* vol. 97, no. 2, pp. 285-295, 2014.

[31] D. A. Gonc¸alves, M. Baranauskas, J. C. d. Reis and E. Todt, "Facial Expressions Animation in Sign Language based on Spatio-temporal Centroid Spatio-temporal Centroid," in *22nd International Conference on Enterprise Information Systems*, 2020.

[32] M. Mehraj, A. Goel, M. A. Butt and M. Zaman, "Automatic Speech Recognition Approach for Diverse Voice Commands," *International Journal of Advanced Research in Computer Science,* vol. 8, no. 9, 2017.

[33] R. G. Rajan and M. J. Leo, "A Comprehensive Analysis on Sign Language Recognition System," *International Journal of Recent Technology and Engineering (IJRTE),* vol. 7, no. 6, 2019.

[34] R. Zhen, W. Song, Q. He, J. Cao, S. L and L. J, "Human-computer interaction system: A survey of talking-head generation," *Electronics,* vol. 12, no. 1, p. 218, 2023.

[35] C. Yu, M. Kang, Y. Chen, J. Wu and X. Zhao, "Acoustic Modelling Based on Deep Learning for Low-Resource Speech Recognition: An Overview," *IEEE Access,* vol. 8, pp. 163829-16384, 2020.

[36] A. V. Haridas, R. Marimuthu and V. G. Sivakumar, "A critical review and analysis on techniques of speech recognition: The road ahead," *International*

Journal of Knowledge-based and Intelligent Engineering Systems, vol. 22, no. 1, pp. 39-57, 2018.

[37] A. Rai, A. Khan, A. Bajaj and J. B. Khurana, "An efficient online examination system using speech recognition," *International Research Journal of Engineering and Technology (IRJET),* vol. 4, no. 4, 2017.

[38] A. Wahab, F. Atanda, S. Azmi, M. Yusof and M. Hariharan, "Yorùbá Automatic Speech Recognition: A Review," in *Rural ICT Development (RICTD) International Conference*, 2013.

[39] P. Gonzalez, T. S. Alstrøm and T. May, "Assessing the Generalisation Gap of Learning-Based Speech Enhancement Systems in Noisy and Reverberant Environments," *IEEE/ACM Transactions on Audio, Speech, and Language Processing,* vol. 31, pp. 3390 - 3403, 2023.

[40] M. A. Dar and J. Pushparaj, "Machine Learning and Deep Learning Approaches for Accent Recognition: A Review," *Computer Speech & Language,* vol. 13, pp. 51527 - 51550, 2025.

[41] J. L. Fendji, D. M. Tala, B. O. Yenke and M. Atemkeng, "Automatic Speech Recognition and Limited Vocabulary: A Survey," *Speech Communication,* vol. 56, p. 85–100, 2022.

[42] C. Neubert, "Language variation in South Africa: A sociophonetician study of the vowel system of Black South African English," 2021.

[43] R. Coto-Solano, "Explicit Tone Transcription Improves ASR Performance in Extremely Low-Resource Languages: A Case Study in Bribri," in *Proceedings of the First Workshop on Natural Language Processing for Indigenous Languages of the Americas*, 2021.

[44] H. S. Nandelenga, "Phonological Processes in the Acquisition of Kiswahili: An Optimality Theory Perspective," *Journal of Literature, Languages and Linguistics,* vol. 85, 2022.

[45] Y. Shan, “Investigating the interaction between prosody and pragmatics quantitatively: a case study of the Chinese discourse marker ni zhidao (“you know”),” *Frontiers in psychology,* vol. 12, p. 716791, 2021.

[46] D. Bagchi and W. Hartmann, “Learning from the Best: A Teacher-student Multilingual Framework for Low-resource Languages,” in *ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Brighton, UK, 2019.

[47] H. Alsharif, “Phonology, phonetic and phonological aspects,” *acrolinguistics and Microlinguistics,* vol. 1, no. 1, pp. 52-68, 2020.

[48] F. Amoussou and A. Ayodele, “Principles, theories and approaches to critical discourse analysis,” *International Journal on Studies in English Language and Literature,* vol. 6, no. 1, pp. 11-18, 2018.

[49] P. Hamad and M. Himdad, “Teaching Stress and Intonation from Students’ Perspective,” *Al-Fatih Journal,* vol. 14, no. 74, 2018.

[50] S. Gong and Z. Jie, “Modelling Mandarin speakers’ phonotactic knowledge,” *Phonology,* vol. 38, no. 2, pp. 241-275, 2021.

[51] B. Solomonov, “Phonotactic phenomenon in the English language and its linguistic nature,” *Journal of Language and Linguistic Studies,* vol. 6, no. 4, pp. 363-367, 2023.

[52] UNESCO, “Kiswahili is a language that speaks to both past and present,” *United Nations Africa Renewal,* 2022.

[53] K. R. Chowdhary, “Natural language processing,” in *Fundamentals of Artificial Intelligence*, New Delhi, Springer, 2020, pp. 603-649.

[54] E. M. Bender, Linguistic fundamentals for natural language processing: 100 essentials from morphology and syntax, Switzerland: Springer Nature, 2022.

[55] E. Awino, L. Wanzare, L. Muchemi, B. Wanjawa, E. Ombui, F. Indede, O. McOnyango and B. Okal, “Phonemic Representation and Transcription for

Speech to Text Applications for Under-resourced Indigenous African Languages:" arXiv preprint arXiv:2210.16537, 2022.

[56] R. Jakobson and M. Halle, Fundamentals of Language, Germany: De Gruyter, 2010.

[57] V. Fromkin, R. Rodman, N. Hyams, M. Amberber, F. Cox and R. Thornton, An Introduction to Language with Online Study Tools 12 Months, Australia: Cengage Learning Australia, 2017.

[58] N. S. Trubetzkoy, Introduction to the Principles of Phonological Descriptions, Belgium: Springer Netherlands, 1968.

[59] I. A. Okafor, "Distinctive features: a linguistic analysis of consonant sounds in English language," *Ansu Journal of Language and Literary Studies,* vol. 2, no. 2, 2022.

[60] E. Dresher, "Contrastive hierarchies and phonological primes," *Perspectives on Element Theory,* vol. 143, p. 33, 2021.

[61] L. Key and B. P. Noble, An Analysis of Ferdinand de Saussure's Course in General Linguistics, London: Macat Library, 2017.

[62] I. A. Okafor, "Distinctive Features: A Linguistic Analysis of Consonant Sounds in English Language," *Ansu Journal of Language and Literary Studies,* vol. 2, no. 2, 2022.

[63] B. Nagy and L. Kovács, "Finite Automata with Translucent Letters Applied in Natural and Formal Language Theory," in *Transactions on Computational Collective Intelligence XVII*, vol. 8790, Heidelberg, Springer, Berlin, 2014.

[64] P. S. Beddor, "Advancements of phonetics in the 21st century: Theoretical and empirical issues in the phonetics of sound change," *Journal of Phonetics,* vol. 98, 2023.

[65] M. Garellek, “Theoretical achievements of phonetics in the 21st century: Phonetics of voice quality,” *Journal of Phonetics,* vol. 94, p. 101155, 2022.

[66] M. Sipser, Introduction to the Theory of Computation, PWS Publishing Company, 1996.

[67] A. Pettorossi, Automata theory and formal languages: fundamental notions, theorems, and techniques, Springer Nature., 2022.

[68] C. Lin, “Formal language theory and its interdisciplinary applications,” in *The Routledge Handbook of Philosophy of Scientific Modelling*, Routledge, 2024, pp. 525-534.

[69] J. Heinz, “Mathematical Linguistics in the 21st Century,” in *Workshop on Formal Language Theory*, New Orleans, LA, 2020.

[70] J. Hopcroft, R. Motwani and J. D. Ullman, “Introduction to automata theory, languages, and computation,” *Acm Sigact News,* vol. 32, no. 1, pp. 60-65, 2001.

[71] K. R. Chowdhary, “Natural Language Processing,” in *Fundamentals of Artificial Intelligence*, New Delhi, Springer, 2020, pp. 603-649.

[72] W. Merrill, Formal Language Theory Meets Modern NLP, ArXiv. /abs/2102.10094, 2021.

[73] S. S. Rao, Engineering optimization: theory and practice, John Wiley & Sons, 2019.

[74] J. Gallier and J. Quaintance, Fundamentals of Optimization Theory with Applications to Machine Learning, University of Pennsylvania Philadelphia, 2019.

[75] H. T. Jongen and K. Meer, Optimization Theory, New York: Springer, 2004.

[76] Y.-F. Liu, T.-H. Chang, M. Hong, Z. Wu, A. M.-C. So, and E. A. Jorswieck, “A Survey of Recent Advances in Optimization Methods for Wireless

Communications," *IEEE Journal on Selected Areas in Communications,* vol. 42, no. 11, pp. 2992 - 3031, 2024.

[77] M. Zubair, "A brief history of information theory by Claude Shannon in data communication," *Journal of Applied and Emerging Sciences,* vol. 13, no. 1, pp. 23-30, 2023.

[78] M. Feicheng, Information communication, Switzerland: Springer Nature, 2022.

[79] O. Rioul, "This is IT: A Primer on Shannon's Entropy and Information," in *Information Theory. Progress in Mathematical Physics*, Cham, Birkhäuser, 2021, p. 49–86.

[80] A. Ali, A. Sania and A. M. Muhammad, "Shannon entropy in artificial intelligence and its applications based on information theory," *Journal of Applied and Emerging Sciences,* vol. 13, no. 1, pp. 09-17, 2023.

[81] A. Ben-Naim, "Introduction, Definitions, and Some General Comments on Shannon's Measure of Information," in *Information Theory*, Cham, Springer, 2024.

[82] S. Mainali, M. Garzon, D. Venugopal, K. Jana, C.-C. Yang, N. Kumar, D. Bowman and L.-Y. Deng, "An Information-theoretic approach to dimensionality reduction in data science," *International Journal of Data Science and Analytics,* vol. 12, p. 185–203, 2021.

[83] C. S. Shikali and R. Mokhosi, "Enhancing African low-resource languages: Swahili data for language modelling," *Data in Brief,* vol. 31, 2020.

[84] F. V¨olpel, "Investigating a new approach to classification from statistical learning theory," 2022.

[85] S. Wang, Q. Wang, N. Bailey and J. Zhao, "Deep neural networks for choice analysis: A statistical learning theory perspective," *Transportation Research Part B: Methodological,* vol. 148, pp. 60-81, 2021.

[86] S. Pablo-García, R. García-Muelas, A. Sabadell-Rendón and N. López, "Dimensionality reduction of complex reaction networks in heterogeneous catalysis: From linear-scaling relationships to statistical learning techniques," *Wiley Interdisciplinary Reviews: Computational Molecular Science,* vol. 11, no. 6, 2021.

[87] G. James, D. Witten, T. Hastie, R. Tibshirani and J. Taylor, "Statistical Learning," in *An Introduction to Statistical Learning. Springer Texts in Statistics*, Cham, Springer, 2023.

[88] E. Steere, A handbook of the Swahili language: as spoken at Zanzibar, Society for promoting Christian knowledge., 1906.

[89] M. J. Robinson, "History of the Standard Swahili Language," *Oxford Research Encyclopaedia of African History,* 2024.

[90] V. Lisanza, "Swahili gaining popularity globally," *United Nations Africa Renewal,* 2021.

[91] UNESCO, "Kiswahili: Education and Culture of Peace," in *The General Conference of UNESCO*, Paris, 2024.

[92] L. J. Mwalongo, "Potentials for Globalization of Kiswahili Language," *East African Journal of Arts and Social Sciences,* vol. 7, no. 1, pp. 541-547, 2024.

[93] M. Nkondo, "African lingua francas and African knowledge systems: A framework for regional integration," *The Quest for Unity,* pp. 65-85, 2023.

[94] W. Whiteley, Swahili: The rise of a national language, Routledge, 2023.

[95] T. M. Maina, A. M. Oirere and S. Kahara, "A CNN-Based Digraph Extraction Model for Enhanced Swahili Natural Language Processing," *international Journal of Scientific Research in Computer Science and Engineering,* vol. 12, no. 6, pp. 41-51, 2024.

[96] M. Leshoele, "Africa's lingua franca and its role in the successful implementation of the AfCFTA: a case for Kiswahili as the African commerce

and epistemological language," *South African Journal of African Languages,* vol. 44, no. 3, pp. 270-277, 2024.

[97] T. M. Maina, A. M. Oirere and S. Kahara, "A CNN-Based Digraph Extraction Model for Enhanced Swahili Natural Language Processing," *International Journal of Scientific Research in Computer Science and Engineering,* vol. 12, no. 6, pp. 41-51, 2024.

[98] S. I. Onzere, M. M. Besin-Mengla and F. K. Lugome, "Impact of Kiswahili Language Variations on Communication and Integration Among Speakers from Different Countries," *Adom Social Science and Humanities Journal,* vol. 1, no. 1, 2025.

[99] G. Martin, M. E. Mswahili, Y.-S. Jeong and J. Woo, "SwahBERT: Language Model of Swahili," in *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, Seattle, United States, 2022.

[100] S. Panat and R. Kumar, "Introduction to Artificial Intelligence & ML," in *A Guide to Applied Machine Learning for Biologists*, Springer, Cham, 2023, p. 127–146.

[101] M. M. Taye, "Understanding of Machine Learning with Deep Learning: Architectures, Workflow, Applications and Future Directions," *Computers,* vol. 12, no. 5, p. 91, 2023.

[102] L. Benos, A. Tagarakis, G. Dolias, R. Berruto, D. Kateris and D. Bochtis, "Machine Learning in Agriculture: A Comprehensive Updated Review," *Sensors and Robotics for Digital Agriculture,* vol. 21, no. 11, p. 3758, 2021.

[103] C. Gambella, B. Ghaddar and J. Naoum-Sawaya, "Optimization problems for machine learning: A survey," *European Journal of Operational Research,* vol. 290, no. 3, pp. 807-828, 2021.

[104] F. Semeraro, A. Griffiths and A. Cangelosi, “Human–robot collaboration and machine learning: A systematic review of recent research,” *Robotics and Computer-Integrated Manufacturing,* vol. 79, 2023.

[105] F.-G. Yuan, S. A. Zargar, Q. Chen and S. Wang, “Machine learning for structural health monitoring: challenges and opportunities,” in *Sensors and smart structures technologies for civil, mechanical, and aerospace systems 2020*, 2020.

[106] A. V. Joshi, Machine Learning and Artificial Intelligence, Springer Cham, 2023.

[107] N. Rane, S. Choudhary and J. Rane, “Machine learning and deep learning: A comprehensive review on methods, techniques, applications, challenges, and future directions,” 2024.

[108] F. Balali, J. Nouri, A. Nasiri and T. Zhao, “Machine Learning Principles,” in *Data Intensive Industrial Asset Management.*, Cham, Springer, 2020.

[109] V. Rani, S. T. Nabi, M. Kumar, A. Mittal and K. Kumar, “Self-supervised Learning: A Succinct Review,” *Archives of Computational Methods in Engineering,* vol. 30, no. 4, pp. 2761-2775, 2023.

[110] I. H. Sarker, “Deep Learning: A Comprehensive Overview on Techniques, Taxonomy, Applications and Research Directions,” *SN Computer Science,* vol. 2, no. 6, 2021.

[111] L. Pauletto, “Neural network architecture search for extreme classification and in a partially labelled learning context,” Université Grenoble Alpes, 2022.

[112] E. R. Vankov and K. Gadhoumi, “Supervised Learning,” in *Statistical Methods in Epilepsy*, Chapman and Hall/CRC, 2024, pp. 272-301.

[113] N. L. Rane, M. Paramesha, S. P. Choudhary and J. Rane, “Machine Learning and Deep Learning for Big Data Analytics: A Review of Methods and

Applications," *Partners Universal International Innovation Journal,* vol. 2, no. 3, p. 172–197, 2024.

[114] D. Jurafsky and J. H. Martin, "Speech and Language Processing," *An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition,* 2023.

[115] A. Aljuaid and M. Anwar, "Survey of Supervised Learning for Medical Image Processing," *SN Computer Science,* vol. 3, no. 292, 2022.

[116] A. Ali, S. A. Razak, S. H. Othman, T. Eisa, A. Al-Dhaqm, M. Nasser, T. Elhassan, H. Elshafie and A. Saif, "Financial Fraud Detection Based on Machine Learning: A Systematic Literature Review," *Applied Sciences,* vol. 12, no. 19, p. 9637, 2022.

[117] S. Naeem, A. Ali, S. Anam and M. M. Ahmed, "An unsupervised machine learning algorithms: Comprehensive review," *International Journal of Computing and Digital Systems,* 2023.

[118] K. Tyagi, C. Rane, R. Sriram and M. Manry, "Chapter 3 - Unsupervised learning," *Artificial Intelligence and Machine Learning for EDGE Computing,* pp. 33-52, 2022.

[119] B. Shen, "E-commerce Customer Segmentation via Unsupervised Machine Learning," in *the 2nd International Conference on Computing and Data Science (CONF-CDS 2021)*, New York, 2021.

[120] A. Choppadandi, J. Kaur, P. K. Chenchala, A. Agarwal, V. Nakra and P. Kirupa, "Anomaly Detection in Cybersecurity: Leveraging Machine Learning Algorithms," *ESP Journal of Engineering & Technology Advancements,* vol. 1, no. 2, pp. 34-41, 2021.

[121] S. Gupta and G. Sharma, "Topic Modelling in Natural Language Processing," *International Journal of Engineering Research & Technology (IJERT),* vol. 10, no. 6, 2021.

[122] G. Li, Z. Yu, K. Yang, M. Lin and C. L. P. Chen, “Exploring Feature Selection with Limited Labels: A Comprehensive Survey of Semi-Supervised and Unsupervised Approaches,” *EEE Transactions on Knowledge and Data Engineering,* vol. 36, no. 11, pp. 6124-6144, 2024.

[123] J. M. Duarte and L. Berton, “A review of semi-supervised learning for text classification,” *An International Science and Engineering Journal,* vol. 56, p. 9401–9469, 2023.

[124] Y. Chen, M. Mancini, X. Zhu and Z. Akata, “Semi-Supervised and Unsupervised Deep Visual Learning: A Survey,” *IEEE Transactions on Pattern Analysis and Machine Intelligence,* vol. 46, no. 3, p. 46, 2024.

[125] D. M. Tadesse, S. R. Kebede, T. G. Debele and F. Gelana, “Application of deep reinforcement learning in various image processing tasks: a survey,” *Evolving Systems,* vol. 16, no. 13, 2025.

[126] M. Paliwal, “Deep Reinforcement Learning,” in *International Conference on Intelligent Emerging Methods of Artificial Intelligence & Cloud Computing (IEMAICLOUD 2021)*, Cham, 2022.

[127] K. Yu, K. Jin and X. Deng, “Review of Deep Reinforcement Learning,” in *2022 IEEE 5th Advanced Information Management, Communicates, Electronic and Automation Control Conference (IMCEC)*, Chongqing, China, 2022.

[128] Y. Li, “Deep reinforcement learning: An overview,” *arXiv preprint arXiv:1701.07274,* 2017.

[129] X. Wang, S. Wang, X. Liang, D. Zhao, J. Huang and X. Xu, “Deep Reinforcement Learning: A Survey,” *IEEE Transactions on Neural Networks and Learning Systems,* vol. 35, no. 4, pp. 5064 - 5078, 2024.

[130] M. Naeem, G. Paragliola and A. Coronato, “A reinforcement learning and deep learning based intelligent system for the support of impaired patients in home treatment,” *Expert Systems with Applications,* vol. 168, 2021.

[131] M. Khan, B. Jan, H. Farman, J. Ahmad, H. Farman and Z. Jan, “Deep learning methods and applications,” *Deep learning: convergence to big data analytics,* Vols. 31-42, 2019.

[132] Shipra and M. Chandra, “DSCC features for Hindi vowel classification,” in *International Conference on Wireless Communications, Signal Processing and Networking (WiSPNET)*, Chennai, India, 2017.

[133] E. C. Djamal, N. Nurhamidah and R. Ilyas, “Spoken word recognition using MFCC and learning vector quantization,” in *2017 4th International Conference on Electrical Engineering, Computer Science and Informatics (EECSI)*, Yogyakarta, Indonesia, 2017.

[134] V. Raveendran, M. R. Sanjeev, N. Paul and K. P. Jijina, “Speech only interface approach for personal computing environment,” *2016 IEEE International Conference on Engineering and Technology (ICETECH),* pp. 372-377, 2016.

[135] M. korba, H. Bourouba and R. Djemili, “Feature Extraction Algorithm Using New Cepstral Techniques for Robust Speech Recognition,” *Malaysian Journal of Computer Science,* vol. 33, no. 2, 2020.

[136] K. Gupta and D. Gupta, “An analysis on LPC, RASTA and MFCC techniques in Automatic Speech recognition system,” in *2016 6th International Conference - Cloud System and Big Data Engineering (Confluence)*, Noida, India, 2016.

[137] S. Davis and P. Mermelstein, “Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences,” in *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 1980.

[138] S. Sen, A. Dutta and N. Dey, Audio Processing and Speech Recognition Concepts, Techniques and Research Overviews, Singapore: Springer Singapore, 2019.

[139] M. A. A. Albadr, S. Tiun, M. Ayob, M. Mohammed and F. T. AL-Dhief, “Mel-Frequency Cepstral Coefficient Features Based on Standard Deviation and

Principal Component Analysis for Language Identification Systems," *Cognitive Computation,* vol. 13, p. 1136–1153, 2021.

[140] S. Mishra, A. Bhowmick and M. C. Shrotriya, "Hindi vowel classification using QCN-MFCC features," *Perspectives in Science,* vol. 8, pp. 28-31, 2016.

[141] L. Wang, Y. Xiong, Z. Wang, Y. Qiao, D. Lin, X. Tang and L. V. Gool, "Temporal Segment Networks: Towards Good Practices for Deep Action Recognition," in *14th European Conference on Computer Vision ECCV*, Amsterdam, The Netherlands, 2016.

[142] S.-S. Wang, P. Lin, Y. Tsao, J.-W. Hung and B. Su, "Suppression by Selecting Wavelets for Feature Compression in Distributed Speech Recognition," in *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 2018.

[143] P. Ramachandrula, S. Sruthi, K. Jaswanth, S. Palaniswamy and R. N. V. Sitaram, "Implementation of Discrete HMMs for Isolated Spoken Word Recognition," in *2023 World Conference on Communication & Computing (WCONF)*, RAIPUR, India, 2023.

[144] K. Ratkevicius, G. Paskauskaite and G. Bartisiute, "Advanced Recognition of Lithuanian Digit Names Using Hybrid Approach," *Elektronika Ir Elektrotechnika,* vol. 24, no. 2, pp. 70-73, 2018.

[145] C. Singla and S. Singh, "Emotion Recognition Using Text and Speech Through Machine Learning," in *International Conference on Advances in Data Science and Computing Technologies*, Singapore, 2022.

[146] S. Tejaswini, N. Sriraam and G. C. Pradeep, "Identification of risk neonates in NICUs: A pilot study using perceptual linear prediction cepstral coefficients with mobility features and machine learning classification," *Computer Science Engineering,* pp. 254-262, 2024.

[147] M. A. Mazumder and R. A. Salam, "Feature Extraction Techniques for Speech Processing: A Review," *International Journal of Advanced Trends in Computer Science and Engineering,* vol. 8, no. 1.3, pp. 285 - 292, 2019.

[148] M. Han, “Artificial intelligence-driven tone recognition of Guzheng: A linear prediction approach,” *Demonstration Mathematica,* vol. 57, no. 1, 2024.

[149] Y. Zouhir, I. B. Fredj, K. Ouni and M. Zarka, “Robust speaker recognition based on biologically inspired features,” *International Journal of Signal and Imaging Systems Engineering,* vol. 12, 2021.

[150] S. A. Alim and N. K. Rashid, Some commonly used speech feature extraction algorithms, London, UK: IntechOpen, 2018.

[151] K. A. Abdalmalak and A. Gallardo-Antolín, “Enhancement of a text-independent speaker verification system by using feature combination and parallel structure classifiers,” *Neural Computing and Applications,* vol. 29, pp. 637-651, 2018.

[152] K. Mehta, A. Dubey, R. Kalsariya, V. Prajapati and S. Pansambal, “Speech Recognition: General Idea and Overview,” *International Research Journal of Engineering and Technology (IRJET),* vol. 7, no. 10, 2020.

[153] P. Prithvi and K. Kumar, “Comparative Analysis of MFCC, LFCC, RASTA - PLP,” *International Journal of Scientific Engineering and Research (IJSER),* vol. 4, no. 5, 2016.

[154] M. A. A. Zulkifly and N. Yahya, “Relative spectral-perceptual linear prediction (RASTA-PLP) speech signals analysis using singular value decomposition (SVD),” *2017 IEEE 3rd International Symposium in Robotics and Manufacturing Automation (ROMA),* 2017.

[155] W. Jia, M. Sun, J. Lian and S. Hou, “Feature dimensionality reduction: a review,” *Complex & Intelligent Systems,* vol. 8, p. 2663–2693, 2022.

[156] P. Chhikara, N. Jain, R. Tekchandani and N. Kumar, “Data dimensionality reduction techniques for Industry 4.0: Research results, challenges, and future research directions,” *Software: Practice and Experience,* vol. 52, no. 3, pp. 658-688, 2022.

[157] M. Greenacre, P. Groenen, T. Hastie, A. I. D'Enza, A. Markos and E. Tuzhilina, "Principal component analysis," *Nature Reviews Methods Primers,* 2022.

[158] C. Labrín and F. Urdinez, "Principal Component Analysis," in *R for Political Data Science: A Practical Guide*, New York, Chapman and Hall/CRC, 2020.

[159] I. T. Jolliffe and J. Cadima, "Principal component analysis: a review and recent developments," *Philosophical Transactions,* 2016.

[160] T. M. Tang and G. I. Allen, "Integrated principal components analysis," *The Journal of Machine Learning Research,* vol. 22, no. 1, p. 8953–9023, 2021.

[161] N. Jakovljevic, D. Miskovic, N. Janev, S. M and V. Delic, "Comparison of Linear Discriminant Analysis Approaches in Automatic Speech Recognition," *Elektronika Ir Elektrotechnika,* vol. 19, no. 7, pp. 76-79, 2013.

[162] E. O. Abiodun, A. Alabdulatif, O. I. Abiodun, M. Alawida, A. Alabdulatif and R. S. Alkhawaldeh, "A systematic review of emerging feature selection optimization methods for optimal text classification: the present state and prospective opportunities," *Neural Computing and Applications,* vol. 33, p. 15091–15118, 2021.

[163] J. Yao and B. Yuan, "Optimization Strategies for Deep Learning Models in Natural Language Processing," *Journal of Theory and Practice of Engineering Science,* vol. 4, no. 5, p. 80–87, 2024.

[164] C. Almeida, C. Castro, V. Leiva, A. C. Braga and A. Freitas, "Optimizing Sentiment Analysis Models for Customer Support: Methodology and Case Study in the Portuguese Retail Sector," *Journal of Theoretical and Applied Electronic Commerce Research,* vol. 19, no. 2, pp. 1493-1516, 2024.

[165] M. Islam, A. Anjum, T. Ahsan and L. Wang, "Dimensionality reduction for sentiment classification using machine learning classifiers," in *IEEE Symposium Series on Computational Intelligence (SSCI)*, Xiamen, China, 2019.

[166] H.-M. Wong, X. Chen, H.-H. Tam, J. Lin, S. Zhang, S. Yan, X. Li and K.-C. Wong, “Feature Selection and Feature Extraction: Highlights,” in *ISMSI '21: Proceedings of the 2021 5th International Conference on Intelligent Systems, Metaheuristics & Swarm Intelligence*, Victoria, Seychelles, 2021.

[167] S. H. Haji and A. M. Abdulazeez, “Comparison of optimization techniques based on gradient descent algorithm,” *PalArch's Journal of Archaeology of Egypt/Egyptology,* vol. 18, no. 4, pp. 2715-2743, 2021.

[168] IBM, “Multinomial Logistic Regression,” 3 March 2023. [Online]. Available: https://www.ibm.com/docs/en/spss-statistics/29.0.0?topic=regression-multinomial-logistic. [Accessed 10 January 2024].

[169] S. Jalalvand, M. Negri, M. Turchi, J. de Souza, D. Falavigna and M. Qwaider, “Transcrater: a tool for automatic speech Recognition Quality Estimation,” in *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics—System Demonstrations*, Berlin, Germany, 2016.

[170] A. S. Bisht and U. Soni, “Application of Linear Regression Modeling on Continuous and Categorical Data using R Programming Scripts,” *International Journal of Applied Engineering Research,* vol. 14, no. 4, pp. 965-976, 2019.

[171] C. Starbuck, “Logistic Regression,” in *The Fundamentals of People Analytics*, Springer, Cham, 2023.

[172] R. Yacouby and D. Axman, “Probabilistic Extension of Precision, Recall, and F1 Score for More Thorough Evaluation of Classification Models,” in *Proceedings of the First Workshop on Evaluation and Comparison of NLP Systems (Eval4NLP)*, 2020.

[173] H. Dalianis , “Evaluation Metrics and Evaluation,” in *Clinical Text Mining*, Springer Cham, 2018.

[174] G. Naidu, T. Zuva and E. M. Sibanda, “A Review of Evaluation Metrics in Machine Learning Algorithms,” in *Proceedings of 12th Computer Science On-line Conference,* 2023.

[175] O. Rainio, J. Teuho and R. Klén, "Evaluation metrics and statistical tests for machine learning," *Scientific Reports,* vol. 14, no. 1, p. 6086, 2024.

[176] O. Oyedele, "Determining the optimal number of folds to use in a K-fold cross-validation: A neural network classification experiment," *Research in Mathematics,* vol. 10, no. 1, p. 2201015, 2023.

[177] L. N. Habibi, T. Matsui and T. Tanaka, "Critical evaluation of the effects of a cross-validation strategy and machine learning optimization on the prediction accuracy and transferability of a soybean yield prediction model using UAV-based remote sensing," *Journal of Agriculture and Food Research,* vol. 16, p. 101096, 2024.

[178] V. W. Lumumba, D. Kiprotich, M. L. Mpaine, G. M. Njoka and D. K. Musyimi, "Comparative Analysis of Cross-Validation Techniques: LOOCV, K-folds Cross-Validation, and Repeated K-folds Cross-Validation in Machine Learning Model," *American Journal of Theoretical and Applied Statistics,* vol. 13, no. 5, pp. 127-137, 2024.

[179] N. J. Saam, "Validation Benchmarks and Related Metrics," in *Computer Simulation Validation. Simulation Foundations, Methods and Applications*, Springer, Cham, 2019.

[180] T. H. Cormen, E. L. Charles, L. R. Ronald and S. Clifford, Introduction to algorithms, MIT press, 2022.

[181] M. Alawad and G. Tourassi, "Computationally Efficient Learning of Quality Controlled Word Embeddings for Natural Language Processing," in *2019 IEEE Computer Society Annual Symposium on VLSI (ISVLSI)*, Miami, FL, USA, 2019.

[182] S. Fei, D. Xu, Z. Chen, Y. Xiao and Y. Ma, "MLR-based feature splitting regression for estimating plant traits using high-dimensional hyperspectral reflectance data," *Field Crops Research,* vol. 293, no. 15, 2023.

[183] K. Fronckova, P. Prazak and A. Slaby, “Singular Value Decomposition and Principal Component Analysis in Face Images Recognition and FSVDR of Faces,” *Information Systems Architecture and Technology: Proceedings of 39th International Conference on Information Systems Architecture and Technology – ISAT 2018,* vol. 853, 2018.

[184] D. Khurana, A. Koli, K. Khatter and S. Singh, “Natural language processing: state of the art, current trends and challenges,” *Multimed Tools Appl,* vol. 82, p. 3713–3744, 2022.

[185] E. M. Ponti, H. O’Horan, Y. Berzak, I. Vulić, R. Reichart, T. Poibeau, E. Shutova and A. Korhonen, “Modelling Language Variation and Universals: A Survey on Typological Linguistics for Natural Language Processing,” *Computational Linguistics,* vol. 45, no. 3, p. 559–601, 2019.

[186] K. Kayabol, “Approximate Sparse Multinomial Logistic Regression for Classification,” *IEEE Transactions on pattern analysis and machine intelligence,* 2019.

[187] M. Malik, M. K. Malik, K. Mehmood and I. Makhdoom, “Automatic speech recognition: a survey,” *Multimedia Tools and Applications,* vol. 80, p. 9411–9457, 2020.

[188] M. Li, M. Liu and H. Masanori, “End-to-end Speech Recognition with Adaptive Computation Steps,” in *ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Brighton, UK, 2019.

[189] M. M. Kabir, M. F. Mridha, J. Shin, I. Jahan and A. Q. Ohi, “A Survey of Speaker Recognition: Fundamental Theories, Recognition Methods and Opportunities,” *IEEE Access,* vol. 9, pp. 79236-79263, 2021.

[190] S. Basak, H. Agrawal, S. Jena, S. Gite, M. Bachute, B. Pradhan and M. Assiri, “Challenges and Limitations in Speech Recognition Technology: A Critical Review of Speech Signal Processing Algorithms, Tools and Systems,”

Computer Modelling in Engineering & Sciences, vol. 135, no. 2, pp. 1053-1089, 2023.

[191] J. W. Creswell and J. D. Creswell, Research Design: Qualitative, Quantitative, and Mixed Methods, 5th ed., Los Angeles: SAGE Publications, Inc., 2018.

[192] E. Babbie and J. D. Edgerton, Fundamentals of Social Research, Canada: Cengage, 2023.

[193] Y. S. Park, K. Lars and A. Anthony, "The positivism paradigm of research," *Academic medicine,* vol. 95, no. 5, pp. 690-694, 2020.

[194] U. M. Mbanaso, L. Abrahams and K. C. Okafor, Research Techniques for Computer Science, Information Systems and Cybersecurity, Germany: Springer Nature Switzerland, 2023.

[195] J. Creswell and V. L. Plano Clark, Designing and Conducting Mixed Methods Research, India: SAGE Publications, 2018.

[196] B. J. Oates, M. Griffiths and R. McLean, Researching Information Systems and Computing, United Kingdom: SAGE Publications, 2022.

[197] J. O. De Sordi, Design Science Research Methodology, Springer International Publishing, 2021.

[198] A. R. Kivaisi, Q. Zhao and J. T. Mbelwa, "Swahili Speech Dataset Development and Improved Pre-Training Method for Spoken Digit Recognition," *CM Transactions on Asian and Low-Resource Language Information Processing,* 2023.

[199] D. Davis, "Swahili: News Classification Dataset," Zenodo, 2020.

[200] GitHub, "Masakhane: Let's put Africa on the NLP map," 2024. [Online]. Available: https://github.com/masakhane-io. [Accessed 27 02 2024].

[201] S. A. Agarwal and N. K. S. Chauhan, Basic Concepts of Research Methodology & Statistics, Bhopal, India: Academic Guru Publishing House, 2024.

[202] V. C. Sinha and A. Gupta, Statistical Analysis, SBPD Publications, 2023.

[203] Y. Gong, G. Liu, Y. Xue, R. Li and L. Meng, “A survey on dataset quality in machine learning,” *Information and Software Technology,* vol. 162, p. 107268, 2023.

[204] A. Paullada, I. D. Raji, E. M. Bender, E. Denton and A. Hanna, “Data and its (dis)contents: A survey of dataset development and use in machine learning research,” *Patterns,* vol. 2, no. 11, p. 100336, 2021.

[205] P. Pakray, A. Gelbukh and S. Bandyopadhyay, “Natural language processing applications for low-resource languages,” *Natural Language Processing,* vol. 31, no. 2, p. 183–197, 2025.

[206] L. Wang, M. Han, X. Li, N. Zhang and H. Cheng, “Review of Classification Methods on Unbalanced Data Sets,” *IEEE Access,* vol. 9, pp. 64606 - 64628, 2021.

[207] C. A. Ramezan, T. A. Warner, A. E. Maxwell and B. S. Price, “Larger samples yield more reliable estimates. In general statistics, “the magnitude of sampling error is inversely proportional to the sample size.,” *Remote Sensing,* vol. 13, no. 3, p. 368, 2021.

[208] S. Mittal, P. Rajput and S. Subramoney, “A Survey of Deep Learning on CPUs: Opportunities and Co-Optimizations,” *IEEE Transactions on Neural Networks and Learning Systems,* vol. 33, no. 10, pp. 5095 - 5115, 2022.

[209] S. Minaee, N. Kalchbrenner, E. Cambria, N. Nikzad, M. Chenaghlu and J. Gao, “Deep Learning--based Text Classification: A Comprehensive Review,” *ACM Computing Surveys (CSUR),* vol. 54, no. 3, pp. 1 - 40, 2021.

[210] E. Orlov and O. Serikov, “Less than Necessary or More than Sufficient: Validating Probing Dataset Size,” in *International Conference on Analysis of Images, Social Networks and Texts*, Cham, 2024.

[211] A. Althnian, D. AlSaeed, H. Al-Baity, A. Samha, A. B. Dris, N. Alzakari, A. A. Elwafa and H. Kurdi, “Impact of Dataset Size on Classification Performance: An Empirical Evaluation in the Medical Domain,” *Applied Sciences,* vol. 11, no. 2, p. 796, 2021.

[212] M. Bakaev, S. Heil, V. Khvorostov and M. Gaedke, “How Many Data Does Machine Learning in Human–Computer Interaction Need? Re-Estimating the Dataset Size for Convolutional Neural Network-Based Models of Visual Perception,” *IT Professional,* vol. 25, no. 2, pp. 23-29, 2023.

[213] A. Nanyonga and G. Wild, “Impact of Dataset Size & Data Source on Aviation Safety Incident Prediction Models with Natural Language Processing,” in *Global Conference on Information Technologies and Communications (GCITC)*, Bangalore, India, 2023.

[214] T. H. Nguyen, H. Nguyen, Z. Ahmadi, T.-A. Hoang and T.-N. Doan, “On the Impact of Dataset Size: A Twitter Classification Case Study,” in *WI-IAT '21: IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology*, New York, USA, 2021.

[215] G. Inoue, B. Alhafni, N. Baimukan, H. Bouamor and N. Habash, “The Interplay of Variant, Size, and Task Type in Arabic Pre-trained Language Models,” *arXiv preprint,* 2021.

[216] C. Aliferis and G. Simon, “Overfitting, Underfitting and General Model Overconfidence and Under-Performance Pitfalls and Best Practices in Machine Learning and AI,” in *Artificial Intelligence and Machine Learning in Health Care and Medical Sciences. Health Informatics*, Cham, Springer, 2024.

[217] H. S. G. Yaamini, S. K. J. S. R, V. K, M. N and T. Jipeng, “Deep Learning Approach for Karnataka Snacks Recognition,” in *2023 IEEE International Conference on Cloud Computing in Emerging Markets (CCEM)*, Mysuru, India, 2023.

[218] K. Lin, T. Alrashed and N. Noy, “Relationships Are Complicated! An Analysis of Relationships Between Datasets on the Web,” in *23rd International Semantic Web Conference*, Baltimore, MD, USA, 2024.

[219] H. Bichri, A. Chergui and M. Hain, “Investigating the Impact of Train / Test Split Ratio on the Performance of Pre-Trained Models with Custom Datasets,” *international Journal of Advanced Computer Science and Applications,* vol. 15, no. 2, 2024.

[220] Q. H. Nguyen, H.-B. Ly, L. Ho, N. Al-Ansari,. H. V. Le, V. Q. Tran, I. Prakash and. B. T. Pham, “Influence of Data Splitting on Performance of Machine Learning Models in Prediction of Shear Strength of Soil,” *Mathematical Problems in Engineering,* vol. 1, p. 4832864, 2021.

[221] M. K. Dahouda and I. Joe, “A Deep-Learned Embedding Technique for Categorical Features Encoding,” *IEEE Access,* vol. 9, pp. 114381-114391, 2021.

[222] J. Kaur and R. S. Sohal, “Noise Estimation and Removal in Natural Language Processing,” *Handbook of Vibroacoustic, Noise and Harshness,* 2024.

[223] F. Rasheeda, X. Zhong and P.-C. Huang, “Padding Module: Learning the Padding in Deep Neural Networks,” *IEEE Access,* vol. 11, pp. 7348 - 7357, 2023.

[224] J. Sadaiyandi, P. Arumugam and A. K. Sangaiah, “Stratified Sampling-Based Deep Learning Approach to Increase Prediction Accuracy of Unbalanced Dataset,” *Electronics,* vol. 12, no. 21, p. 4423, 2023.

[225] K. M. Kahloot and P. Ekler, “Algorithmic Splitting: A Method for Dataset Preparation,” *IEEE Access,* vol. 9, pp. 125229 - 125237, 2021.

[226] S. Rutunda, K. Kabanda and A. Stan, “Kinyarwanda Tts: Using A Multi-Speaker Dataset to Build a Kinyarwanda Tts Model,” in *4th Workshop on African Natural Language Processing*, Kigali, Rwanda, 2023.

[227] A. Katumba, S. Kagumire, J. Nakatumba-Nabende, J. Quinn and S. Murindanyi, “Building Text-to-Speech Models for Low-Resourced Languages from Crowdsourced Data,” *Applied AI Letters,* vol. 6, no. 2, 2024.

[228] N. M. Müller, P. Kawa, W. H. Choong, E. Casanova, E. Gölge and T. Müller, “MLAAD: The Multi-Language Audio Anti-Spoofing Dataset,” in *2024 International Joint Conference on Neural Networks (IJCNN)*, Yokohama, Japan, 2024.

[229] J. Yu and Y. Luo, “Efficient Monaural Speech Enhancement with Universal Sample Rate Band-Split RNN,” in *ICASSP 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Rhodes Island, Greece, 2023.

[230] C. Li, H. Zhi, K. Yang, J. Qian, Z. Yan and L. Zhu, “A 0.61-μW Fully Integrated Keyword-Spotting ASIC With Real-Point Serial FFT-Based MFCC and Temporal Depth wise Separable CNN,” *IEEE Journal of Solid-State Circuits,* vol. 59, no. 3, pp. 867 - 877, 2024.

[231] Y.-C. Wu, I. D. Gebru, D. Marković and A. Richard, “Audiodec: An Open-Source Streaming High-Fidelity Neural Audio Codec,” *ICASSP 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP),* 2023.

[232] Deepanshu, K. Srinivas and A. C. Kumari, “Convolutional Neural Network-Based Automated Acute Lymphoblastic Leukaemia Detection and Stage Classification from Peripheral Blood Smear Images,” *International Journal of Scientific Research in Computer Science and Engineering,* vol. 12, no. 3, pp. 21-28, 2024.

[233] H. S. G. Yaamini, S. J, S. H. R, V. K, M. N and T. Jipeng, “Deep Learning Approach for Karnataka Snacks Recognition,” in *2023 IEEE International Conference on Cloud Computing in Emerging Markets (CCEM)*, Mysuru, India, 2023.

[234] A. Suresh, A. Jain, K. Mathur and P. Gambhir, “Comparison of Modelling ASR System with Different Features Extraction Methods Using Sequential Model,” in *International Conference on Artificial Intelligence and Speech Technology*, Cham, 2022.

[235] A. F. Atanda, “Multinomial logistic regression probability ratio-based feature vectors for Malay vowel recognition,” Universiti Utara Malaysia, 2021.

[236] Z. Abdul, “Kurdish speaker identification based on one dimensional convolutional neural network,” *Comput. Methods Differ. Equ,* vol. 7, no. 4, pp. 566-572, 2019.

[237] X. Zhao and D. Wang, “Analysing noise robustness of MFCC and GFCC features in speaker identification,” in *IEEE Int. Conf. Acoust. Speech Signal Process*, 2013.

[238] H. Yin, V. Hohmann and C. Nadeu, “Acoustic features for speech recognition based on gammatone filter bank and instantaneous frequency,” *Speech Commun,* vol. 53, no. 5, pp. 707-715, 2022.

[239] M. Ismail, “Ideal dataset splitting ratios in machine learning algorithms: general concerns for data scientists and data analysts,” in *7th international Mardin Artuklu scientific research conference*, Mardin, Turkiye, 2022.

[240] Q. H. Nguyen, H.-B. Ly, L. S. Ho, N. Al-Ansari, H. V. Le, V. Q. Tran, I. Prakash and B. T. Pham, “Influence of Data Splitting on Performance of Machine Learning Models in Prediction of Shear Strength of Soil,” *Mathematical Problems in Engineering,* vol. 2021, no. 1, p. 4832864, 2021.

[241] A. Rácz, D. Bajusz and K. Héberger, “Effect of Dataset Size and Train/Test Split Ratios in QSAR/QSPR Multiclass Classification,” *Molecules,* vol. 26, no. 4, p. 1111, 2021.

[242] M. K. Singh, “Multimedia application for forensic automatic speaker recognition from disguised voices using MFCC feature extraction and

classification techniques," *Multimedia Tools and Applications,* vol. 83, p. 77327–77345, 2024.

[243] J. v. Brocke, A. Hevner and A. Maedche, "Introduction to Design Science Research," in *Design Science Research. Cases*, 2020.

[244] J. Waller, "Performance Benchmarking of Application Monitoring Frameworks," in *Kiel Computer Science Series*, Department of Computer Science, 2014.

[245] M. Patten, Empirical Validity, 10th ed., New York: Routledge, 2017.

[246] S. Sarma and N. Pathak, "Design and Implementation of an Assamese Language Chatbot Using," *International Journal of Scientific Research in Computer Science and Engineering,* vol. 11, no. 6, pp. 13-18, 2023.

[247] T. M. Maina, "The Swahili Digraph Corpus," *Mendeley Data,* vol. 2, 2024.

[248] C. S. Shivachi, R. Mokhosi, Z. Shijie and L. Qihe, "Learning Syllables Using Conv-LSTM Model for Swahili Word Representation and Part-of-speech Tagging," *Transactions on Asian and Low-Resource Language Information Processing,* vol. 20, no. 4, pp. 1 - 25, 2021.

[249] A. Zafar, M. Aamir, N. M. Nawi, A. Arshad, S. Riaz, A. Alruban, A. K. Dutta and S. Almotairi, "A Comparison of Pooling Methods for Convolutional Neural Networks," *Applied Sciences,* vol. 12, no. 17, 2022.

[250] A. Ziaee and E. ÇAno, "Batch Layer Normalization A new normalization layer for CNNs and RNNs," in *ICAAI '22: Proceedings of the 6th International Conference on Advances in Artificial Intelligence*, 2023.

[251] G. Fotopoulos, P. Popovich and N. H. Papadopoulos, "Review Non-Convex Optimization Method for," 2024.

[252] W. Olin-Ammentorp, K. Beckmann, C. D. Schuman, J. Plank and N. Cady, "Stochasticity and robustness in spiking neural networks," *Neurocomputing,* vol. 419, pp. 23-36, 2021.

[253] N. Kumar, M. Sharma, V. P. Singh, C. Madan and S. Mehandia, "An empirical study of handcrafted and dense feature extraction techniques for lung and colon cancer classification from histopathological images," *Biomedical Signal Processing and Control,* vol. 75, p. 103596, 2022.

[254] S. Selvakumari and M. Durairaj, "A Comparative Study of Optimization Techniques in Deep Learning Using the MNIST Dataset," *Indian Journal of Science and Technology,* vol. 18, no. 10, pp. 803-810, 2025.

[255] S. Mehra, V. Ranga, R. Agarwal and S. Susan, "Speaker independent recognition of low-resourced multilingual Arabic spoken words through hybrid fusion," *Multimedia Tools and Applications,* vol. 83, no. 35, pp. 82533-82561, 2024.

[256] B. Bala and S. Behal, "A Brief Survey of Data Preprocessing in Machine Learning and Deep Learning Techniques," in *8th International Conference on I-SMAC (IoT in Social, Mobile, Analytics and Cloud) (I-SMAC)*, Kirtipur, Nepal, 2024.

[257] D. Singh and B. Singh, "Feature wise normalization: An effective way of normalizing data," *Pattern Recognition,* vol. 122, p. 108307, 2022.

[258] A. N. Çayır and T. S. Navruz, "Effect of Dataset Size on Deep Learning in Voice Recognition," in *3rd International Congress on Human-Computer Interaction, Optimization and Robotic Applications (HORA)*, Ankara, Turkey, 2021.

[259] H. Gao, M. Hasegawa-Johnson and C. D. Yoo, "G2PU: Grapheme-To-Phoneme Transducer with Speech Units," in *ICASSP 2024 - 2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Seoul, Korea, 2024.

[260] C. Batista, A. L. Dias and N. Neto, “Free resources for forced phonetic alignment in Brazilian Portuguese based on Kaldi toolkit,” *EURASIP Journal on Advances in Signal Processing,* vol. 1, 2022.

[261] J. I. Ramírez-Hernández, A. Manzo-Martínez, F. Gaxiola, L. González-Gurrola, V. Álvarez-Oliva and R. López-Santillán, “A Comparison Between MFCC and MSE Features for Text-Independent Speaker Recognition Using Machine Learning Algorithms,” in *Fuzzy Logic and Neural Networks for Hybrid Intelligent System Design. Studies in Computational Intelligence,* vol. 1061, Springer, Cham, 2023.

[262] D. R. KS, R. MD and S. G, “Comparative performance analysis for speech digit recognition based on MFCC and vector quantization,” in *Global Transitions Proceedings*, 2021.

[263] Y. Hu and H. Jiang, “Predicting Change-Proneness with a Deep Learning Model: Incorporating Structural Dependencies,” in *IEEE 48th Annual Computers, Software, and Applications Conference (COMPSAC)*, Osaka, Japan, 2024.

[264] Z. K. Abdul and A. K. Al-Talabani, “Mel Frequency Cepstral Coefficient and its Applications: A Review,” *IEEE Access,* vol. 10, pp. 122136 - 122158, 2022.

[265] Z. J. M. Ameen and A. A. Kadhim, “Machine learning for Arabic phonemes recognition using electrolarynx speech,” *International Journal of Electrical and Computer Engineering (IJECE),* vol. 13, no. 1, pp. 400-412, 2023.

[266] T. R. McIntosh, T. Susnjak, N. Arachchilage, T. Liu, D. Xu and P. Watters, “Inadequacies of Large Language Model Benchmarks in the Era of Generative Artificial Intelligence,” *IEEE Transactions on Artificial Intelligence (Early Access),* 2025.

[267] Q. Jiang, Y. Gu, C. Li, R. Cong and F. Shao, “Underwater Image Enhancement Quality Evaluation: Benchmark Dataset and Objective Metric,” *IEEE*

Transactions on Circuits and Systems for Video Technology, vol. 32, no. 9, pp. 5959 - 5974, 2022.

[268] S. Riezler and M. Hagmann, Validity, reliability, and significance: Empirical methods for NLP and data science, 2nd ed., Switzerland: Springer Nature, 2024.

[269] M. Nasution, M. Elveny and R. Syah, "Research methodology for computer science," in *2nd International Conference on Advanced Information Scientific Development (ICAISD)*, Jakarta, Indonesia, 2023.

APPENDICES

Appendix 1: Python Extraction Codes for Digraph Sequences from Swahili Datasets

```python
import os
import pandas as pd
import re

# Paths to the directories containing the datasets
data_dirs = [
    r'D:\PhD\objecive_one\DataSets\New folder',
    r'D:\PhD\objecive_one\DataSets\Mendeley A Kiswahili dataset',
    r'D:\PhD\objecive_one\DataSets\Zenedo Swahili data (1)\Swahili data',
    r'D:\PhD\objecive_one\DataSets\Kaggle Swahili Audio Dataset'
]

# Define the specific vowel sequences to look for
vowel_sequences = ["cha", "che", "chi", "cho", "chu",
                   "dha", "dhe", "dhi", "dho", "dhu",
                   "gha", "ghe", "ghi", "gho", "ghu",
                   "kha", "khe", "khi", "kho", "khu",
                   "ng'a", "ng'e", "ng'i", "ng'o", "ng'u",
                   "nya", "nye", "nyi", "nyo", "nyu",
                   "sha", "she", "shi", "sho", "shu",
                   "tha", "the", "thi", "tho", "thu",
                   "nga", "nge", "ngi", "ngo", "ngu"]

def extract_words_with_vowel_sequences(text, vowel_sequences):
    """Function to extract words containing specific vowel sequences from a given text."""
    vowel_word_list = []
    words = text.split()  # Split text into words
    for word in words:
        # Use regex to check if the word contains only alphabetic characters (excluding symbols and numbers)
        if re.match(r'^[a-zA-Z]+$', word):
            for vowel_seq in vowel_sequences:
                if vowel_seq.lower() in word.lower():  # Check if vowel sequence is in word
                    vowel_word_list.append([vowel_seq, word])
                    break  # If a vowel sequence is found, move to the next word
    return vowel_word_list

# List to hold all vowel sequences and their respective words
all_vowel_words = []

# Read data from each text file in the specified directories
for data_dir in data_dirs:
    if not os.path.isdir(data_dir):
        print(f"Directory {data_dir} does not exist. Skipping.")
        continue

    for file_name in os.listdir(data_dir):
        file_path = os.path.join(data_dir, file_name)
        if os.path.isfile(file_path) and file_name.endswith('.txt'):
            try:
                with open(file_path, 'r', encoding='utf-8') as file:
                    text = file.read()
                    vowel_word_list = extract_words_with_vowel_sequences(text, vowel_sequences)
                    all_vowel_words.extend(vowel_word_list)
            except Exception as e:
                print(f"Error reading file {file_path}: {e}")

# Create a DataFrame to list each vowel sequence with its respective words
df = pd.DataFrame(all_vowel_words, columns=["Vowel Sequence", "Word"])

# Remove duplicate words while preserving the first occurrence
df = df.drop_duplicates(subset=["Word"])

# Define the maximum rows per sheet
max_rows_per_sheet = 1048576

# Save the DataFrame to multiple sheets in an Excel file
output_file = r'D:\PhD\objecive_one\Swahili_vowel_sequences_word_list.xlsx'

```

```
try:
    with pd.ExcelWriter(output_file, engine='openpyxl') as writer:
        # Check if DataFrame is empty
        if df.empty:
            # In case no data is written (empty DataFrame), create an empty sheet
            df_empty = pd.DataFrame({"Message": ["No data available"]})
            df_empty.to_excel(writer, sheet_name="Empty_Data", index=False)
        else:
            # Flag to check if at least one sheet is created
            sheet_created = False

            # Loop over chunks of DataFrame, creating sheets
            for i in range(0, len(df), max_rows_per_sheet):
                df_chunk = df.iloc[i:i + max_rows_per_sheet]
                sheet_name = f'Sheet_{i//max_rows_per_sheet + 1}'
                df_chunk.to_excel(writer, sheet_name=sheet_name, index=False)
                sheet_created = True  # Mark that at least one sheet has been created

            # Ensure at least one sheet is visible (active)
            if sheet_created:
                first_sheet_name = list(writer.sheets.keys())[0]
                writer.book.active = writer.sheets[first_sheet_name]

    print(f"Results have been saved to {output_file}")
except Exception as e:
    print(f"Error saving Excel file: {e}")

```

Appendix 2: Approval Of Research Proposal and Supervisors

MURANG'A UNIVERSITY OF TECHNOLOGY

Office of the Director, Postgraduate Studies

E-mail: hps@mut.ac.ke
Website: www.mut.ac.ke

P.O. Box 75-10200
Murang'a, Kenya

Ref: MUT/ARP/PGS/19/2020 /VOL.1 **Date:** 10th July, 2024

Mr. TIRUS MUYA MAINA - SC501/5213/2019

School of Computing and Information Technology,
Murang'a University of Technology,
P.O. Box 75-10,200
Murang'a.

Dear Mr. Maina,

RE: APPROVAL OF RESEARCH PROPOSAL AND SUPERVISORS.

I am pleased to inform you that the Directorate of Postgraduate Studies on 1st July 2024 considered and approved your Doctoral research proposal entitled ***"Optimizing Swahili digraph recognition using a feature vector reduction model employing multinomial logistic regression"*** and appointed the following as supervisors:

1. Dr. Aaron Oirere - Murang'a University of Technology
2. Dr. Stephen Kahara – Murang'a University of Technology

You may now proceed with your data collection subject to obtaining research permit from NACOSTI, if required. You should also begin consulting your supervisors and submit through them quarterly progress reports to the Director Postgraduate Studies through your CoD and School Dean. Progress Reports forms can be accessed in the Student Portal.

It is the policy and regulations of the University that you observe deadlines. The guidelines on Postgraduate supervision can be accessed in the post graduate Handbook.

Your responsibilities as a student will include, among others;

i. Maintain regular consultation with your supervisor(s), at least once a month

ii. Submit quarterly reports on time, through your supervisors, CoD, Dean and to the Director of Postgraduate Studies;
iii. Ensure quality work all through;
iv. Present your research findings at 2-3 seminars/conferences prior to thesis examination.
v. Publish two articles from your research findings in a refereed journal prior to thesis examination.

For any further clarification please contact the undersigned.

Yours Sincerely,

Prof. Geoffrey Muchiri, PhD
Director, Postgraduate Studies
GM/PIM

DIRECTORATE OF POSTGRADUATE STUDIES
MURANG'A UNIVERSITY OF TECHNOLOGY
10 JUL 2024
P. O. Box 75 - 10200, MURANG'A
Tel: 0771463515, Email: info@mut.ac.ke

Copy to:
1. Registrar – Academics and Students Affairs
2. Dean, School of Computing and Information Technology

Appendix 3: Approval Letter from NACOSTI



REPUBLIC OF KENYA

NATIONAL COMMISSION FOR SCIENCE,TECHNOLOGY & INNOVATION

Ref No: **565251**

Date of Issue: **03/February/2025**

RESEARCH LICENSE

This is to Certify that Mr.. Tirus Muya Maina of Murang'a University of Technology, has been licensed to conduct research as per the provision of the Science, Technology and Innovation Act, 2013 (Rev.2014) in Muranga on the topic: OPTIMIZING SWAHILI DIGRAPH RECOGNITION USING A FEATURE VECTOR REDUCTION MODEL EMPLOYING MULTINOMIAL LOGISTIC REGRESSION for the period ending : 03/February/2026.

License No: **NACOSTI/P/25/415430**

565251

Applicant Identification Number

Director General
NATIONAL COMMISSION FOR SCIENCE,TECHNOLOGY & INNOVATION

Verification QR Code

NOTE: This is a computer generated License. To verify the authenticity of this document, Scan the QR Code using QR scanner application.

See overleaf for conditions

Appendix 4: Publications

The following papers were published based on this thesis.

[1] T. M. Maina, A. M. Oirere and S. Kahara, "A CNN-Based Digraph Extraction Model for Enhanced Swahili Natural Language Processing," *International Journal of Scientific Research in Computer Science and Engineering,* vol. 12, no. 6, pp. 41-51, 2024.

[2] T. M. Maina, A. M. Oirere and S. Kahara, "Development of the Annotated Swahili Digraph Corpus Using a CNN-Based Digraph Extraction Model," *International Journal of Scientific Research in Computer Science and Engineering*, vol. 12, no. 6, pp. 58-65, 2024.