SCHOOL OF ELECTRICAL ENGINEERING
AND TELECOMMUNICATIONS

Arabic Dialect Segmentation of Conversations

by

Madeline Younes

Student ID: z1234567

Thesis submitted as a requirement for the degree
Bachelor of Engineering (Electrical Engineering)

Submitted: November 16, 2022
Supervisor: Dr Beena Ahmed

Abstract

This report preposes a thesis to develop a novel approach to Dialectal Identification (DID) through utilising pre-trained end-to-end speech recognition model for low resource dialects, particularly Arabic. It explores the motivation for a thesis that explores the use of using transfer learning and semi-self supervised machine learning methods in improving Dialectal Identification systems particularly for Arabic dialects. This report also, examines current literature on methods used in Language Identification (LID) and Dialect Identification (DID), as well as their shortcomings. There has been a significant amount of research into transfer learning for English based downstream tasks while the technology has not progressed to the same extent for low resource languages. In addition to this there has been limited research into using transfer based learning for LID and DID tasks. In the later half of this paper will present planning for the remainder of the thesis, as well as the preliminary work which has been conducted.

Acknowledgements

I wish to thank Dr Beena Ahmed for her guidance and encouragement she has provided so far. I would also like to thank Iain McCowan from Dubber for introducing Language and Dialectal Identification as a novel use case for transfer learning.

Abbreviations

BE Bachelor of Engineering

EE&T School of Electrical Engineering and Telecommunications

LaTeX A document preparation computer program

CNN Convolutional Neural Network

BiLTSM Bidirectional Long Short-Term Memory

LTSM Long Short-Term Memory

RNN Recurrent Neural Network

LID Language Identification

DID Dialectal Identification

ASR Automatic Speech Recognition

NLP Natural Language Processing

E2E End to End

MSA Modern Standard Arabic

NOR North African Arabic

EGY Egyptian Arabic

LEV Levantine Arabic

GLF Gulf Arabic

Contents

Chapter 1

Introduction

1.1 Problem Statement

In multicultural societies such as that in Sydney, NSW it is common to have speakers of different dialects interact. There are around 200 thousand Arabic speakers within NSW, majority speaking Levantine dialect of Arabic with then a spread among Egyptian and Gulf Arabic. It would be a common scenario where perhaps for a telehealth consultation the only Arabic speaker available doesn't speak the same dialect as a patient. The speaker may then code switch with the words, phrases etc. they know from one dialect to their main dialect to hold a conversation. Having a system that is able to accurately identify the dialects spoken in certain segments to then produce an accurate transcription would certainly be helpful in that scenario. Dialectal Arabic has limited available datasets, no commercially available datasets and is considered low reasource. Current methods use phoneme recognition or traditional machine learning but both these methods have flaws that limit their ability to reliably recognise Arabic dialects. As phoneme identifiers rely on phonemic differences between the dialects, where there are shared phonemes distinguishing the dialects become a difficult task. While traditional machine learning requires large amounts of labelled training data which is currently unavailable for Arabic dialects. This thesis will investigate an accurate and reliable method to segment conversations of dialectal Arabic into homogenous dialectal segments and identify the dialect of each segment. The dialectal identifier should accurately distinguish the four major Arabic dialects and then further extended to be able to distinguish between seventeen regional dialects.

1.2 Thesis Aims

The goal of this thesis is to answer the overarching question: *How can the existing limited resource Arabic dialect speech Corpa be used to improve the accuracy Arabic Dialectal Identification?*

The key aims of this thesis are:

- To assess the viability of using transfer learning to improve the accuracy of Arabic Dialectal Identification.
- To investigate the performance of a transfer learning based DID on low resource dialects. Through assessing the minimum amount of data needed to accurately fine tune a DID

system.

- To critically analyse which pretrained model and downstream model architecture is able to produce the most accurate DID system.
- To explore whether a transfer learning based DID system can be adapted to be accurately applied to a finer set of dialectal groups.

1.3 Chapter Outline

This report is organised as described. Chapter 2 details some background information surrounding the impacts of this thesis and describes the unique features of Arabic Dialects. Chapter 3 provides a detailed analysis of current LID and DID methodologies. As well as details the literature which supports the use of transfer learning for the application of LID and DID. Chapter 4 proposes a methodology for this thesis and details the expected results. Chapter 5 explores the preliminary work conducted for this thesis. Chapter 6 details the work to be conducted during this thesis and outlines the possible challenges Chapter 7 draws up conclusions and summaries the key takeaways of the report.

Chapter 2

Background

To define Language Identification (LID), it is the process of differentiating spoken audio and classifying the audio segment based on its corresponding language. LID systems are the critical first step in selecting the most accurate model to use for Automatic Speech Recognition (ASR), multilingual transcription or other automated speech processing systems. A dialect is a sub-variant of a language which is usually mutually intelligible by other speakers of that language despite the speaker using a different dialect. Dialects evolve within a certain region, area or within a class. Dialect Identification (DID) generally posses some interesting challenges compared to language identification that semi-self supervised systems could provide solutions to. These challenges include that dialects unlike languages are not standardised, generally have very limited labelled datasets and so, are often considered low resource and the differences between different dialects are often not as clear as the differences between languages.

2.1 Arabic Dialects

This thesis will focus on Arabic dialects as despite being a widely spoken language and dialects being the primary spoken form of Arabic, there are still significant improvements that can be made to Arabic DIDs with current systems achieving the highest accuracy of 86.29%.

Arabic is the official language of 25 countries and has 330 million native speakers.
Academically the regional dialects are usually grouped into 5 main groups North African (NOR), Egyptian (EGY), Levantine (LEV), Gulf (GLF) and Modern Standard Arabic (MSA). MSA is taught academically in most Arabic speaking countries and originates from the Gulf region but is not used for general conversation or outside academic setting.

Comparatively to MSA, the lack of standardisation in dialectal Arabic has resulted in more linguist complexities. Dialectal Arabic has a richer morphology and cliticisation system, accounting for circumfix negation, and for attached pronouns to act as indirect objects. As well as this some words are shared but are used for differing functions eg. For example,'Tyb' is used as an adjective in MSA but dialectaly as an interjection. North African has the largest amount of dialectical variation within the dialect and is the most different from the other Arabic dialects. Taking influences from French and Berber languages. Egyptian globally is the widely understood dialect due to the Egyptian movie and television industry. Levantine dialects differ slightly in pronunciation and intonation but are equivalent when transcribed. Closely related to Aramaic. Gulf is the form which is most closely related to MSA and preserves many of MSA verb conjugations. Understanding of different dialects depends on an indervidual's exposure outside of their own country. eg. due to the prevalence of Egyptian television and movies, many Arab people can understand the Egyptian Dialect but a Leventine speaker would not be able to understand the Moroccan dialect.

The differences between Arabic dialects are comparable to the differences present in North Germanic languages such as Norwegian, Swedish, Danish or the West Salvic languages eg. Czech, Slovak, Polish. Some linguistic variation between dialectal Arabic include incongruous morphemes, prepositions verb conjugations, word meanings, phonemes and pronunciation. Some examples of this is shown in the table 2.1. [15,16]

In addition to this majority of available pretrained models that are used in self supervised or semi supervised systems are trained on English datasets. Arabic has 6 vowels/diphthongs in MSA and 8-10 vowels in most dialects, 28 constants while English has 24 consonants and 22 vowels. [59] As well as this compared to English, Arabic and particularly dialectal Arabic has large amount of morphemes, rendering it unfeasible for the training data to contain all the possible morphemes. So, a system which has been built to operate well for a DID that is linguistically different to English should thereby be robust enough to be applied to other languages with similar complexity.

Hence, the two key challenges in creating an Arabic DID which will be explored in this thesis are:

- Dialectal Arabic is considered low resource, as there are no large commercially available datasets.
- There is a significant amount of complex linguist differences and similarities between Arabic dialects.

More details about the ADI17 dataset to be used are provided in Chapter 5, the dataset contains audio from 17 countries and be divided into the 4 widely spoken major dialect groups (NOR, EGY, LEV, GLF), as MSA is not spoken in general conversation it is not included in the dataset and will not be identified by the DID designed in this thesis.

English/Feature	MSA	LEV	GLF	EGY
Money	nqwd	mSAry	flws	flws
I want	Aryd	bdy	Abγý	ςAyz
Now	AlĀn	hlq	AlHyn	dlwqt
When?	mtý?	Aymtý?	mtý?	Amtý
alveolar affricate sound	dj	j	y	g
Handsome	djami:l	jami:l	yami:l	gami:l
consonant sound	Θ	t	Θ	t
Three	Θala:Θa	tla:te	Θala:Θa	tala:ta

Table 2.1: Examples of linguistic differences between Arabic Dialects

Chapter 3

Literature Review

3.1 An Introduction to LID and DID systems

Dialect identification (DID) is a specialised task of Language Identification (LID) which identifies the dialects within a language. It poses more challenges compared to LID, as dialects share many acoustic, linguistic features and speaker characteristics. An accurate LID or DID allows for more specialised models to be used for other speech related tasks including ASR, speech transcription and Natural Learning Processing (NLP). Over time the methodology of LID and DID systems has evolved. Traditionally Phonematic Modelling was used to construct Arabic DIDs which is discussed further in 3.2.1. Then traditional machine learning networks were used, which is explored in Section 3.2.2. Current research is exploring the viability of utilising transfer based learning methods for LIDs and DIDs, which is detailed in Section 3.3. Table 3.1 compares the accuracies that were achieved both with traditional machine learning methods and transfer learning.

Application	**Features**	**Pretrained Model**	**Downstream Model**	**Accuracy**	**Year, Paper**
Arabic DID (17 dialects)	80 dimensional Fbank	Transformer Based Network (trained on ADI17)	CNN	86.29%	(2020), [31]
Arabic DID (5 dialects)	i-vector + FBANK + word + char + phoneme	N/A	E2E CNN+RNN+FC, DNN+SNN (feature extraction)	81.36%	(2018), [49]
LID (English, Spanish, French, German, Russian, and Italian)	Spectrograms	Resnet50	CNN + RESnets	89%	(2019), [45]
LID (Arabic, English, Malay, French, Spanish, German, Persian, and Urdu)	Acoustic features (MFCC + GMM + i-vector)	N/A	ESA-ELM (Enhanced Self- Adjusting Extreme Learning Machine)	96.25%	(2018), [53]
LID (26 languages)	N/A	wav2vec 2.0	pooling layer + linear layer	95.5%	(2021), [38]

Table 3.1: Recent Machine Learning Implementations of LIDs & DIDs.

3.2 Traditional Methodologies

3.2.1 Phonematic Modeling

A phoneme in linguistics is the smallest unit of sound which can convey meaning (for instance, the sound /c/ in cat). Phonematic modeling utilises recognisers to identify the phonemes present within an audio segment. Different dialects usually have different phoneme combinations and so, the identified phonemes are mapped to identify a dialect. In the paper [15] this technique was used to construct an Arabic DID for 4 dialects (Gulf, Iraqi, Levantine, and Egyptian) plus MSA which took advantage of English, Arabic, Japanese phone recognisers to identify the pheoneme differences between the dialects as seen in Figure 3.2.1. This method was able to achieve high accuracy levels for identifying MSA with F-Measures above 98% and the highest of the dialects was Egyptian Arabic with an F-Measure of 90.2% with 30s test-utterances. As seen in Figure 3.2.1, phonematic modelling for Arabic struggled when given shorter utterances and had particularly low accuracies for the Gulf dialect. The key challenges with using phoneme modelling is that it relies on the distinguishing phonemes to be present in the test data and for finer regional dialects it requires there to be more shared phonemes between the dialects.

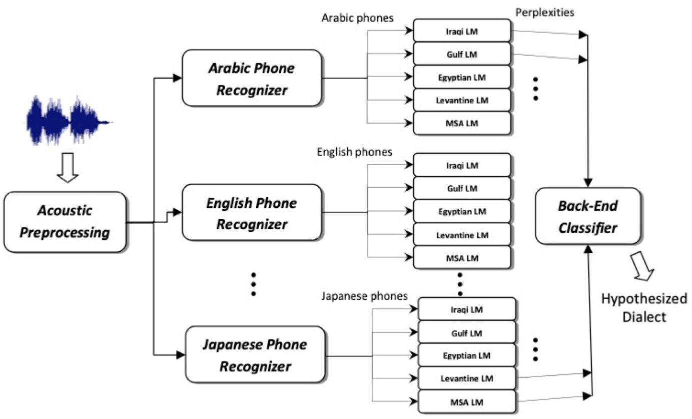


Figure 3.1: Parallel Phone Recognition Followed by Language Modeling (PRLM) for Arabic DID [15].

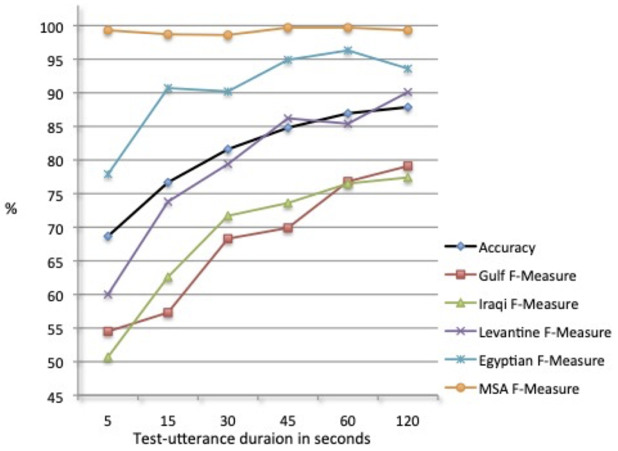
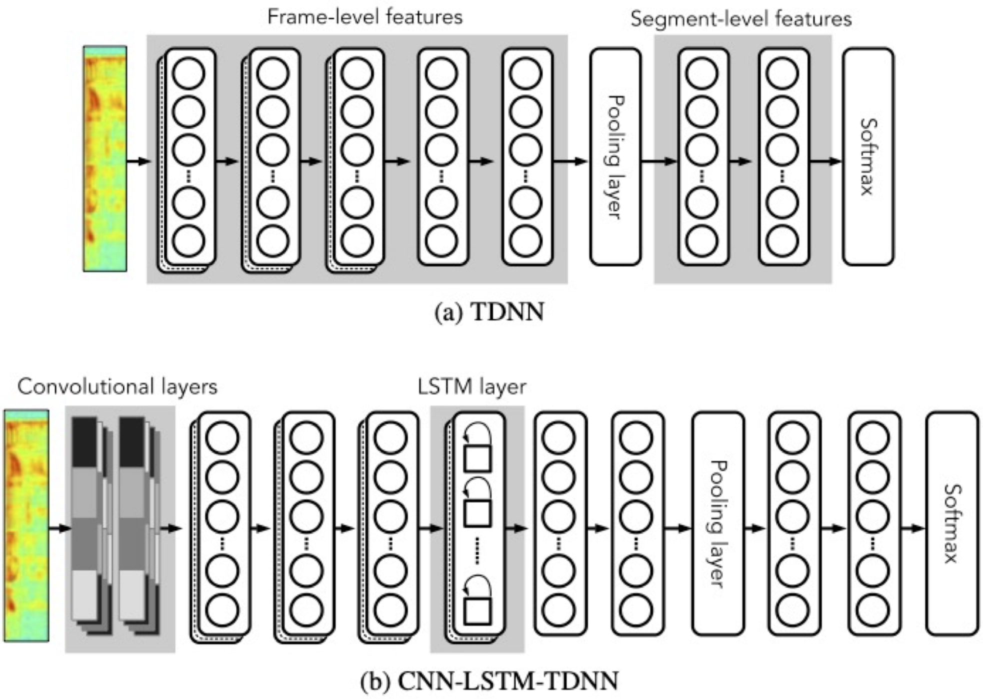


Figure 3.2: The accuracies and F-Measures of the five-way classification task with different test-utterance durations [15].

3.2.2 Traditional Machine Learning

Traditional Machine learning has been used for both LID and DID systems as explored in papers [3,31,35,36,44,49,53]. They operate by extracting key features from the training audio, which could be acoustic and/or linguistic then using some form of traditional machine learning structure to learn the differences between dialects based on the extracted features. Papers [31,35,36,49] use transformer based networks which are constructed with a similar structure to that shown in Figure 3.3. Simple transformer networks were compared to networks which used a combination of CNN, LTSM networks along with the transformer network. Convolutional Neural Networks (CNN) are composed of three types of layers, a convolutional layer, pooling layer and a fully-connected (FC) layer, with a greater amount of layers the complexity of the network increases. CNNs learn through using filters to detect certain features in the training data and adjusting its weights accordingly. The Arabic DID explored in paper [32] used the ADI17 dataset which will be used in this thesis was able to achieve the highest accuracy of 86.29% when cascaded with a CNN network. In contrast, Bidirectional Long Short Term Memory (BiLSTM) is created from two Recurrent Neural Networks (RNN). It has the ability to combine information from both past and future inputs. The structure of BiLSTM is shown in Figure 3.4. Although, there are no papers showing the effectiveness of using BiLSTMs specifically for Arabic DID, LSTMs were used in [35,36], transformer based networks and were able to achieve an accuracy of well over 90% for all the ADI17 dialects in [36]. As well as this the papers [46,59] explored the use of BiLTSMs in text based Arabic DIDs and paper [54] explored its use in Mandarin/English LID with XLS-R producing a 92.7% accuracy. Since, both CNN and BiLSTM networks have been used in traditional machine learning LIDs and DIDs they will be explored as possible downstream models to be used in this thesis.



(a) TDNN

(b) CNN-LSTM-TDNN

Figure 3.3: Transformer Network based LID [36].

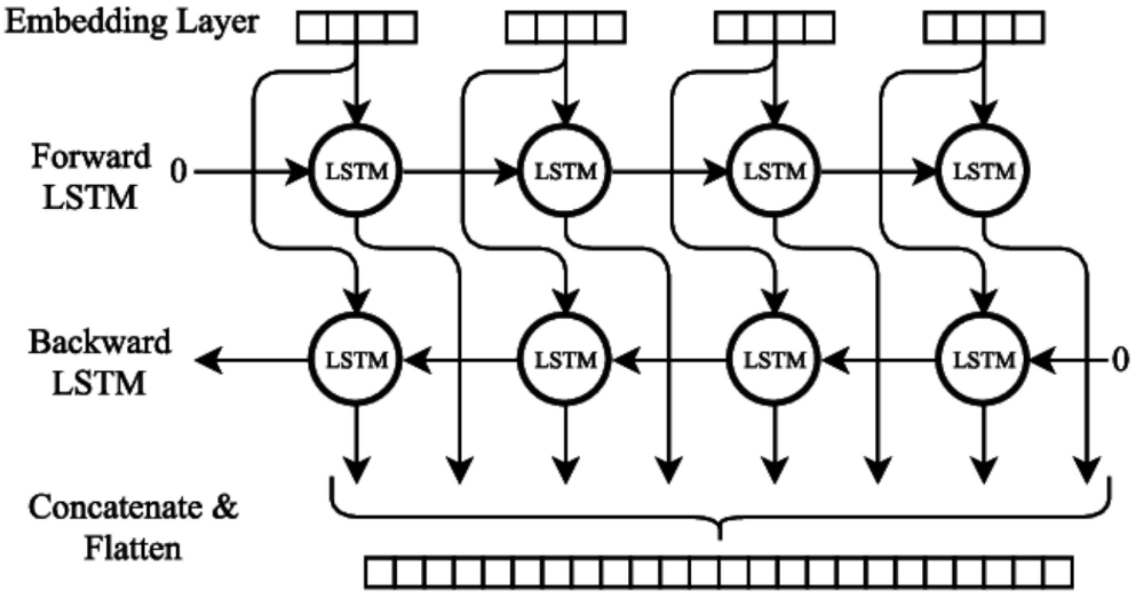


Figure 3.4: BiLTSM Structure [20].

3.3 Transfer Learning

Transfer learning is a form of deep machine learning which focuses on adapting a pretrained model to execute a similar task. The pretraining enables rapid learning for modeling the task and often requires less data to achieve a high level of accuracy in the secondary class. For the task of LID and DID, this often means applying a pretrained model designed for speech tasks like ASR, speech synthesis etc. Transfer learning is a relatively new method for optimising machine learning to complete a task and so, there are not a vast amount of papers exploring its use for some more niche tasks like LID and DID. Papers [8, 23, 38, 58] have shown significant increases in accuracy using pretrained models for tasks such as emotion recognition, ASR and NLP.

In addition to a pretrained model, a transfer learning system has a downstream model connected to it that allows the model to adapt to more specific tasks. The system can then be trained end to end (E2E) or fine-tuned in portions, tuning the pretrained model, then the downstream. It was found in the paper [58], which explored E2E training for wav2vec and HuBERT for the tasks Speaker Verification (SV), Intent Classification (IC) and Slot Filling (SF), that on average using E2E training provides more accurate systems. For example, looking at their results for wav2vec, E2E outperformed segmented training by 12.47% in SER, decreased EER by 3.26% in SV, improved accuracy in IC by 39.98% and SF by 36.66%. Thereby, this thesis will employ E2E training for the system.

3.3.1 Pretrained Models

The pretrained models are semi-self supervised machine learning models often designed by large tech companies, then trained on large amounts of unlabeled and small amounts of labelled data. There are several pretrained models available for use for speech processing some of which are shown in Table 3.2, although the main ones that will be explored are HuBERT, wav2vec 2.0 and XLS-R developed by Facebook. Wav2vec 2.0 is designed for speech data, consisting of a feature encoder, context network, quantisation module and a contrastive loss layer. The feature encoder is a 7-layer, 512 channel CNN that translates a waveform into feature vectors, reducing the dimension of the audio to 1D. It does this every 20ms and has a receptive field of 400 samples which is equivalent to 25ms of audio sampled at 16kHz. The quantisation layer addresses the continuous nature of speech data, automatically learning discrete speech units such as phonemes and words. While the transformer encoder composed of 12 transformer blocks and learns from the vectors from the CNN. Wav2vec is pretrained using a contrastive task, masking a unit in the feature vector then predicting what should be in that unit. In the case where the prediction is wrong a negative score is given and when right a positive, and the network then adjusts its weights accordingly. HuBert is a hidden unit bidirectional and shares a structure with wav2vec 2.0 using a transformer based networks and contrastive based learning, although it uses BERT. BERT is able to process a segment of speech simultaneously learning the surrounding context of a word. It aimed to improve wav2vec through the use of BERT prediction loss and was able to produce up to 19% and 13% relative WER reduction for a 1B parameter model. XLS-R is a fine-tuned variant of wav2vec 2.0, that is trained using data from 128 different languages collected from BABEL, MLS, CommonVoice and VoxPopuli speech corpa. Tuning the model on languages other than English reduced error rates 14-34% relative on average [10]. It has also shown to operate with a higher degree of accuracy on low resource languages compared to other models as shown in Figure 3.3.1. The paper [10] compared the accuracy when using no pretrained model, wav2vec 2.0 and XSL-R on a 26 language LID. The highest accuracy was consistently achieved by XLS-R as seen in Figure 3.3.1, the highest being 95.7% with 100hrs of labelled training data. Hence, this thesis will be using XLS-R and benchmarking it against wav2vec 2.0, HuBERT will not be tested as for the scope of this thesis it is too ambitious to explore more than two pretrained models.

There hasn't been any research into the application of pretrained models for Arabic DIDs but there has been limited reaserch into using wav2vec for LID systems. The papers [10,38,53] demonstrate it as a possible methodology for LID. The paper [38] was able to achieve an accuracy of 95.5% for their 26 language LID utilising only a simple pooling layer and linear layer as their downstream model as shown in Figure 3.5 and so this thesis will use this as the benchmark downstream model.

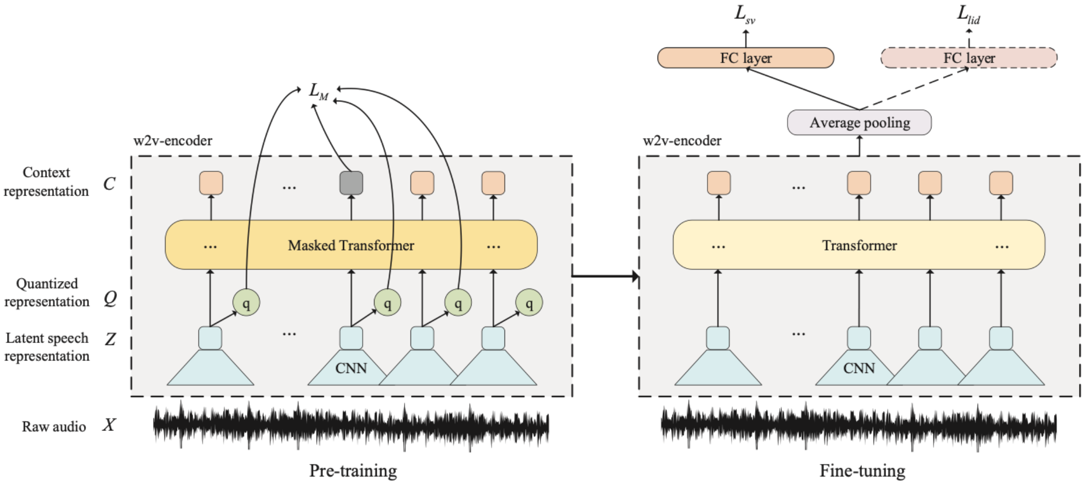
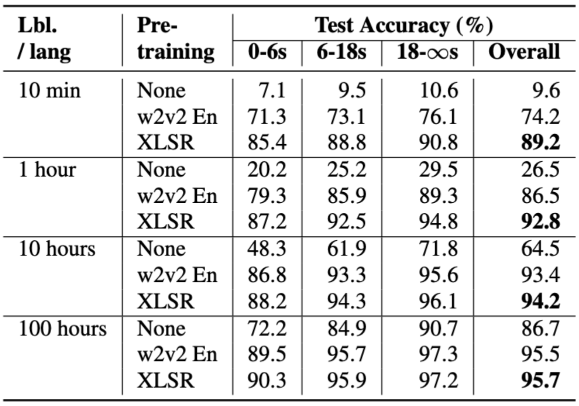


Figure 3.5: wav2vec 2.0 LID [38].



Lbl. / lang	Pre-training	Test Accuracy (%)			
		0-6s	6-18s	18-∞s	Overall
10 min	None	7.1	9.5	10.6	9.6
	w2v2 En	71.3	73.1	76.1	74.2
	XLSR	85.4	88.8	90.8	**89.2**
1 hour	None	20.2	25.2	29.5	26.5
	w2v2 En	79.3	85.9	89.3	86.5
	XLSR	87.2	92.5	94.8	**92.8**
10 hours	None	48.3	61.9	71.8	64.5
	w2v2 En	86.8	93.3	95.6	93.4
	XLSR	88.2	94.3	96.1	**94.2**
100 hours	None	72.2	84.9	90.7	86.7
	w2v2 En	89.5	95.7	97.3	95.5
	XLSR	90.3	95.9	97.2	**95.7**

Figure 3.6: 26 language LID test accuracy [10].

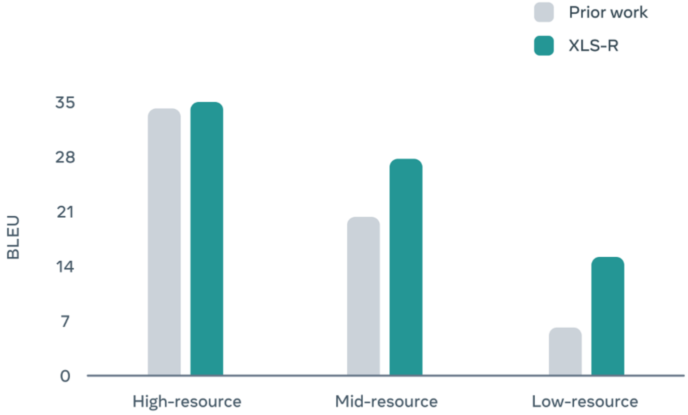


Figure 3.7: XLS-R BLEU Accuracy when Translating to English [1]

Model	Key Features	Year, Paper
Hubert	**Training:** 60k hrs of English unlabeled speech and 1hr of labelled speech. **Structure:** Masked hidden units.	(2021), [24]
Albert	Bert variant designed to be more lightweight.	(2020), [29]
w2v-Bert	wav2vec and HuBert hybrid. Slight accuracy improvement on some downstream tasks 5% to 10% improvement in WER reduction, but has a more complex structure.	(2021), [19]
wav2vec 2.0	**Training:** 53k hrs of English unlabeled speech and 1hr of labelled speech. **Structure:** Masking and contrastive based learning.	(2020), [12]
XLS-R	Fined tuned version of wav2vec 2.0. **Training:** 436k hrs unlabeled speech data in 128 languages.	(2021), [4]

Table 3.2: Possible Pretrained Model Choices.

Chapter 4

Thesis Outline

4.1 Proposed Approach

As discussed in the Chapter 3, using transfer learning for designing DID is a feasible method which may be able to address some challenges with Arabic DID, as discussed in Sections 1.1 and 2. This thesis will do this through constructing different transfer learning systems and comparing their accuracy. A block diagram giving an overview of the system is shown in Figure 4.1. The portions which will remain consistent in the system throughout experimentation will be the initial data preprocessing which will use the python packages noisereduce to filter the data and perform channel normalisation. (The importance of preprocessing is covered in the section 5.1.) As well as the outputting softmax layer which will classify the dialects based on the given categorisations. In Thesis B, the classification of 4 generalised dialects will be explored and Thesis C will look to see if the system constructed can then be applied to 17 regional dialects. The portions which will be investigated include the pretrained model and the downstream model. The choices for each have been explained in Sections 3.3.1 and 3.2.2. The benchmark model for each section respectively will be wav2vec 2.0 and a simple pooling + linear layer as the downstream model. It is expected that using XLS-R and BiLSTM will produce the system with the highest accuarcy DID. This thesis will also investigate the minimum amount of data required to create an effective system, to validate the claim that transfer learning is a useful method for low resource languages and will do this through varying both the amount of training data and the length of utterances for the test data. A breakdown of the tasks and testing which will be performed are in Section 6.2.2.

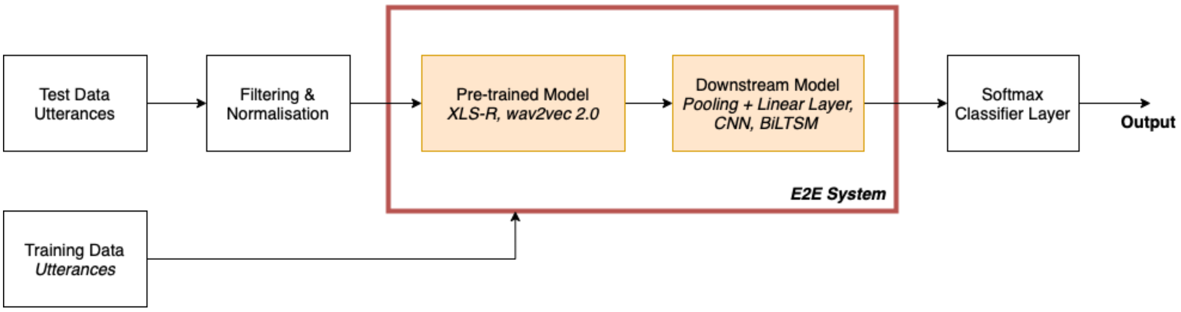


Figure 4.1: Block Diagram of Proposed Approach.

4.2 Expected Outcomes

The expected outcomes of this thesis is to have designed an Arabic DID with at least an accuracy of 85% for both the 4 generalised dialects and the 17 finer dialects. This will demonstrate that transfer learning is a viable methodology for creating DIDs, particularly for low resource languages and dialects. It is also expected that using the more generalised language pretrained model XLS-R will outperform a model utlising wav2vec 2.0.

Chapter 5

Preliminary Work

During the course of Thesis A, some preliminary work was conducted to prepare for the work which will be done in Thesis B and Thesis C. This included gaining access to NCIS Supercomputer and setting up the SSH on my personal computer. Setting up the development environment with jupyter notebook, python. Downloading the training ADI17 dataset and performing some analysis on it which is detailed in the Section 5.1.

5.1 Dataset Analysis

5.1.1 Dataset Selection

There are limited dialectal Arabic datasets available, ADI17 [50], QASR [39] and the ArPod [34] dataset were assessed as viable options to be used as training data for the DID to be designed in this thesis. An exploration of their features and negative characteristics is in the Table 5.1.

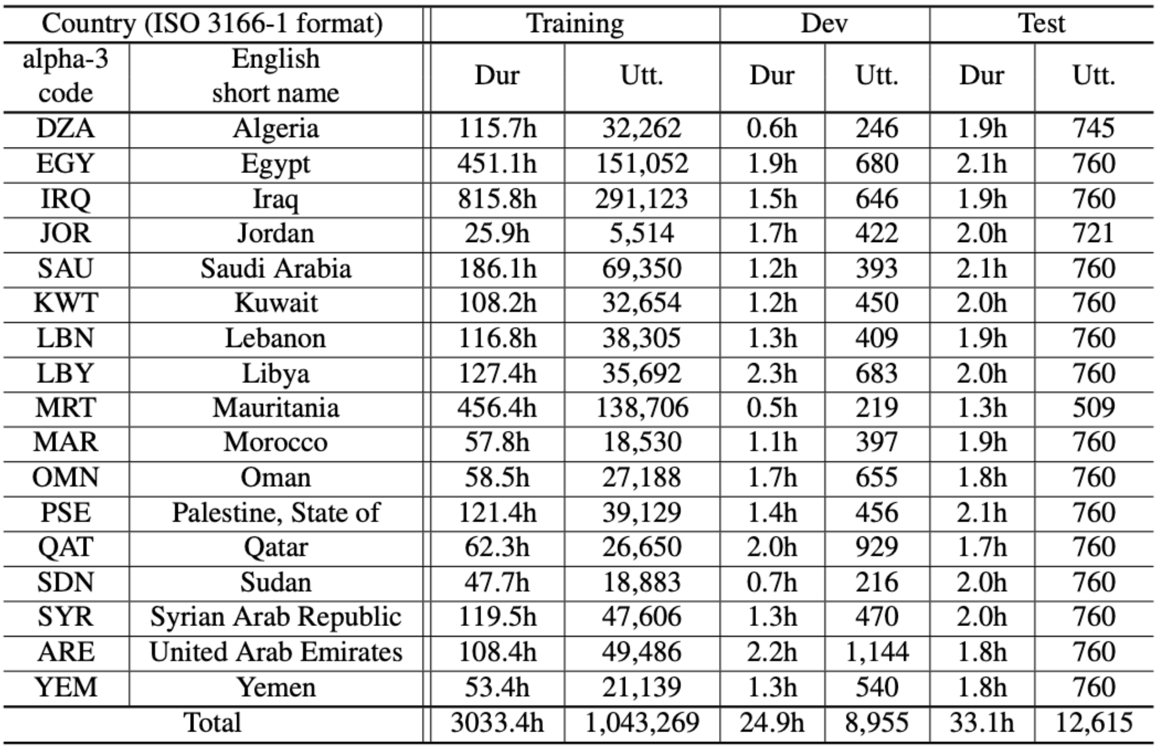
Dataset	Features	Associated Challenges	Access Status
ADI17	**Amount:** 3000hrs of conversational audio on various topics. **Source:** YouTube videos. **Languages/Dialects:** 17 regional dialects. Contains codeswitching.	The dataset is noisy with significant amount of acoustic variation.	Access granted.
QASR: QCRI Aljazeera Speech Resource	**Amount:** 2000hrs of broadcast audio on various topics. **Source:** Aljazeera broadcasting network. **Languages/Dialects:** 3 Languages (English, French, Arabic), 5 dialects (MSA, GLF, LEV, NOR, EGY). Contains codeswitching.	Even though this dataset has many favourable features, no contact information is provided to access the dataset.	Access unavailable (no contact information provided)
ArPod	**Amount:** 8hrs of high quality conversational audio. **Source:** Podcasts. **Languages/Dialects:** 2 Languages (English, Arabic), 5 dialects (MSA, SAU, EGY, LEB, SYR).	The dataset is very small compared to alternate datasets and only contains data from a limited set of regional dialects.	Access not granted but can be obtained through contacting Dr. Mourad Abbas

Table 5.1: Possible Dataset Choices.

The **ADI17** dataset has been chosen to be used for this thesis as it was the most suitable due to a couple of reasons. Particularly it was designed with the intention to be used for DID systems and has the largest amount of associated resources.

5.1.2 ADI17 Dataset

The (ADI17) comprised of audio segments from known Youtube videos with dialects from 17 different Middle Eastern and North African countries. The dataset is divided into training, development and test data groups. The training set contains 3000hrs of audio total while the development and test combined is 57hrs of audio. The specifics of the dataset can be seen in Figure 5.1. The data was collected from around 30 different Youtube channels per country and the primary dialect each Youtube channel used was verified by a human annotator. Using the Youtube channel's dialect audio segment's dialectal label was allocated. The training data relies on this for its labelling, whilst the test and development data was annotated by a human annotator. The audio segments are split into utterances, which are small portions of audio generated by segmenting the original audio at silence points. These silence points are usually natural pauses in conversation and a threshold is used to determine how long the silence must be before the audio is split. The creators of the ADI17 dataset have not specified the threshold that they used. The dataset is labelled using 17 regions, Thesis B explores creating a DID of 4 generalised dialects that encompass this finer set of regional dialects as shown in Figure 5.1.2. So, for training in Thesis B a portion of data from each region is taken to construct the training set for the generalised dialects as shown in Figure 5.1.2. The core challenges with the ADI17 dataset are that the acoustics are unbalanced across each of the dialect regions and the amount of data provided is unbalanced. The amount of noise in each of the regions datasets is shown in Figure 5.6, although its effect on the DID will be mitigated through using channel normalisation and filtering, as the papers [16, 42] have shown is effective at increasing accuracy of Arabic DIDs. The dataset is also unbalanced in terms of amount of training data for each region as shown in Figure 5.1.2, with Jordan having the least amount of data. This will not affect Thesis B, as the generalised dialect groups will collate portions of the data. While for Thesis C, the training data provided will be restricted to 10hrs to ensure that the training sets for each region are even and balanced.



Country (ISO 3166-1 format)		Training		Dev		Test	
alpha-3 code	English short name	Dur	Utt.	Dur	Utt.	Dur	Utt.
DZA	Algeria	115.7h	32,262	0.6h	246	1.9h	745
EGY	Egypt	451.1h	151,052	1.9h	680	2.1h	760
IRQ	Iraq	815.8h	291,123	1.5h	646	1.9h	760
JOR	Jordan	25.9h	5,514	1.7h	422	2.0h	721
SAU	Saudi Arabia	186.1h	69,350	1.2h	393	2.1h	760
KWT	Kuwait	108.2h	32,654	1.2h	450	2.0h	760
LBN	Lebanon	116.8h	38,305	1.3h	409	1.9h	760
LBY	Libya	127.4h	35,692	2.3h	683	2.0h	760
MRT	Mauritania	456.4h	138,706	0.5h	219	1.3h	509
MAR	Morocco	57.8h	18,530	1.1h	397	1.9h	760
OMN	Oman	58.5h	27,188	1.7h	655	1.8h	760
PSE	Palestine, State of	121.4h	39,129	1.4h	456	2.1h	760
QAT	Qatar	62.3h	26,650	2.0h	929	1.7h	760
SDN	Sudan	47.7h	18,883	0.7h	216	2.0h	760
SYR	Syrian Arab Republic	119.5h	47,606	1.3h	470	2.0h	760
ARE	United Arab Emirates	108.4h	49,486	2.2h	1,144	1.8h	760
YEM	Yemen	53.4h	21,139	1.3h	540	1.8h	760
Total		3033.4h	1,043,269	24.9h	8,955	33.1h	12,615

Figure 5.1: ADI17 Dataset Details.

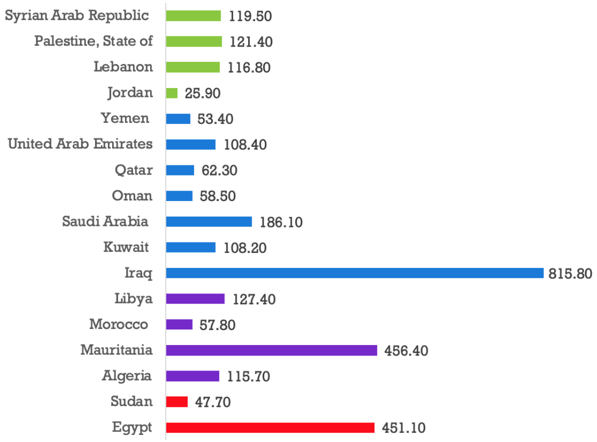


Figure 5.2: Plot of ADI17 Training Data.

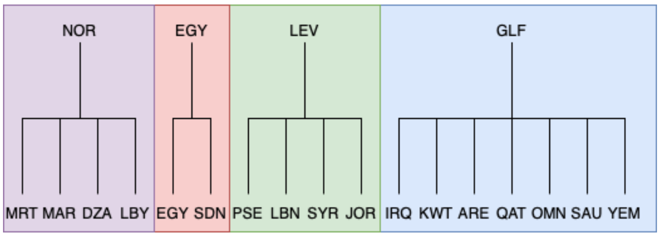
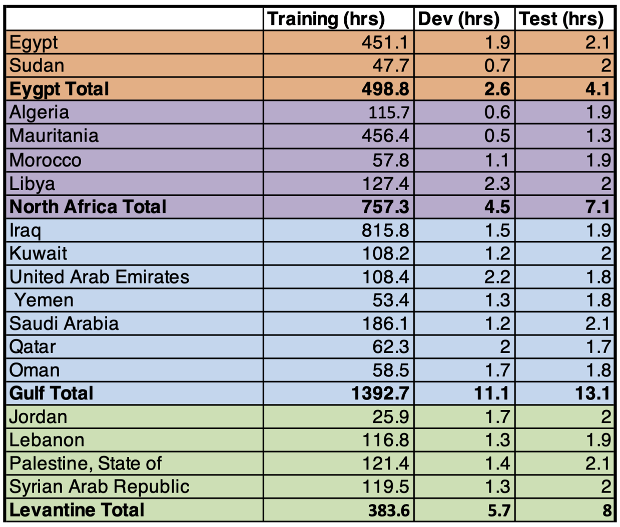
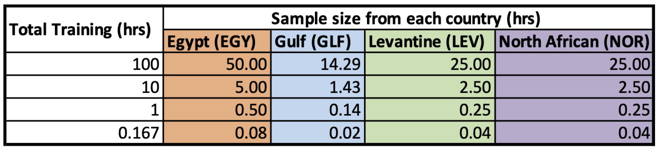


Figure 5.3: ADI17 Grouped into 4 major dialects.



	Training (hrs)	Dev (hrs)	Test (hrs)
Egypt	451.1	1.9	2.1
Sudan	47.7	0.7	2
Eygpt Total	**498.8**	**2.6**	**4.1**
Algeria	115.7	0.6	1.9
Mauritania	456.4	0.5	1.3
Morocco	57.8	1.1	1.9
Libya	127.4	2.3	2
North Africa Total	**757.3**	**4.5**	**7.1**
Iraq	815.8	1.5	1.9
Kuwait	108.2	1.2	2
United Arab Emirates	108.4	2.2	1.8
Yemen	53.4	1.3	1.8
Saudi Arabia	186.1	1.2	2.1
Qatar	62.3	2	1.7
Oman	58.5	1.7	1.8
Gulf Total	**1392.7**	**11.1**	**13.1**
Jordan	25.9	1.7	2
Lebanon	116.8	1.3	1.9
Palestine, State of	121.4	1.4	2.1
Syrian Arab Republic	119.5	1.3	2
Levantine Total	**383.6**	**5.7**	**8**

Figure 5.4: Total data for each of the 4 dialects.



Total Training (hrs)	Sample size from each country (hrs)			
	Egypt (EGY)	Gulf (GLF)	Levantine (LEV)	North African (NOR)
100	50.00	14.29	25.00	25.00
10	5.00	1.43	2.50	2.50
1	0.50	0.14	0.25	0.25
0.167	0.08	0.02	0.04	0.04

Figure 5.5: ADI17 Grouped into 4 dialect groups.

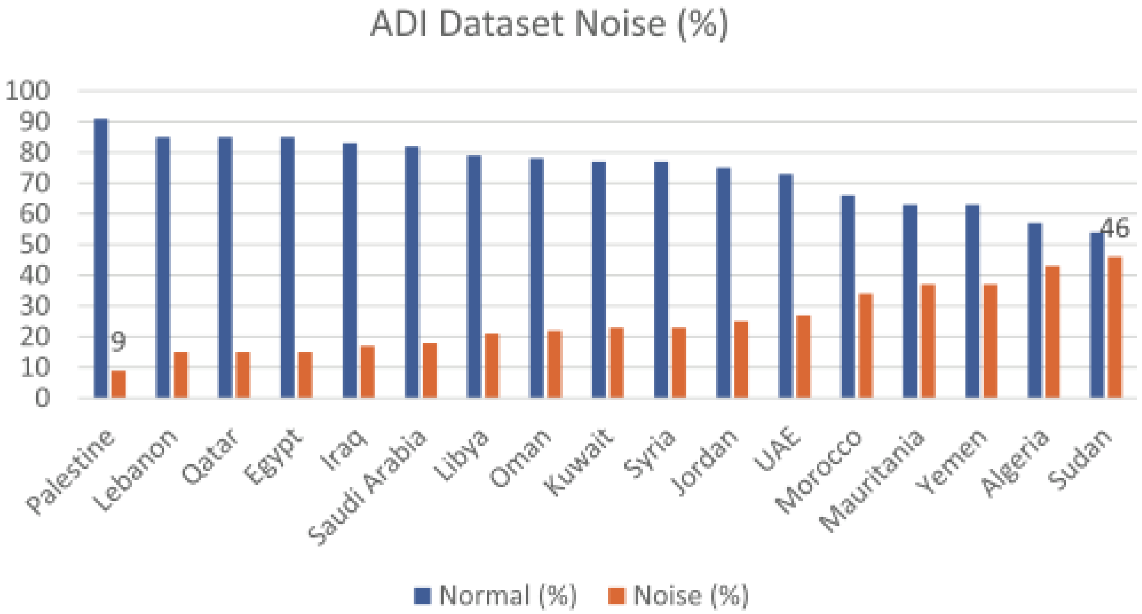


Figure 5.6: ADI17 Dataset Noise levels. [6]

Chapter 6

Thesis Plan

6.1 Timeline Overview

An overview of summarising the planned timeline for this thesis is presented in Figure 6.1. Tasks have been broken into manageable 2-3 week intervals and testing, evaluating and assignment preparation have also been accounted for in the timeline. Thesis A has comprised mainly of finding a focus topic, developing a sound understanding of the background information on dialectal Arabic, LIDs, DIDs and transfer learning. Evaluated possible datasets, prebuilt models, python packages and downstream models to be implemented in Thesis B. As well as gaining access to the Gadi supercomputer, dataset and performing some data analysis on the set.

The majority of the implementation and experimentation for this thesis will occur in Thesis B. The dataset preprocessed using python audio processing packages such as NoiseReduce and PyAudio. The main machine learning model with the pretrained model and downstream model will be implemented using the SpeechBrain Python toolkit.

Thesis C focus will focus on evaluating the performance of the model developed in Thesis B and further developing it. This evaluation will be done through changing the amount of classifier groups from 4 to 17. In addition to testing the model with various utterance lengths. There has been significant time allocated in Thesis C for reflecting, testing and iterating upon the model designed.

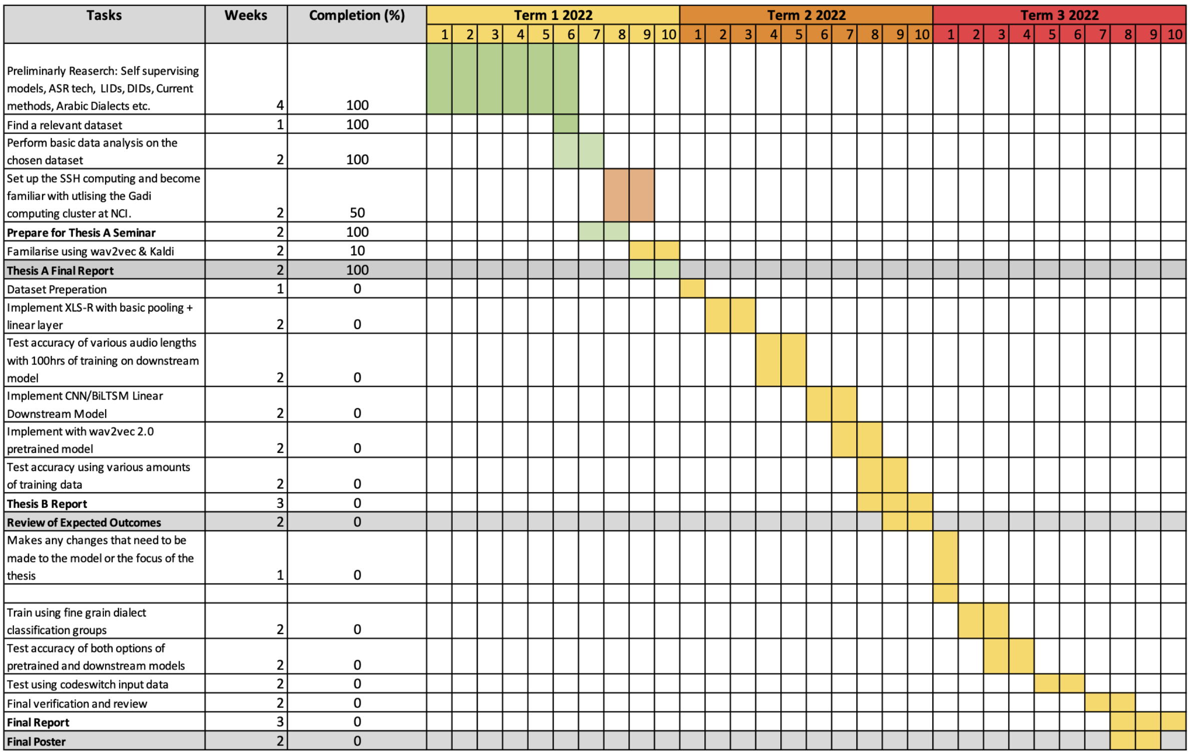
6.2 Thesis B

6.2.1 Aims

Thesis B will investigate various methods for designing a DID for the 4 major Arabic dialects (EGY, GLF, LEV, NOR). All testing done in Thesis B will be done with 120sec utterance test data.

The goals of Thesis B are as follows:

- Investigate the system's performance and determine if it is robust for low resource dialects.
- Benchmark a generalised language pretrained model against an English speech pretrained model.
- Determine the most accurate downstream model architecture.



Tasks	Weeks	Completion (%)	Term 1 2022										Term 2 2022										Term 3 2022									
			1	2	3	4	5	6	7	8	9	10	1	2	3	4	5	6	7	8	9	10	1	2	3	4	5	6	7	8	9	10
Preliminarly Reaserch: Self supervising models, ASR tech, LIDs, DIDs, Current methods, Arabic Dialects etc.	4	100	🟩	🟩	🟩	🟩	🟩	🟩																								
Find a relevant dataset	1	100						🟩																								
Perform basic data analysis on the chosen dataset	2	100						🟩	🟩																							
Set up the SSH computing and become familiar with utlising the Gadi computing cluster at NCI.	2	50								🟧	🟧																					
Prepare for Thesis A Seminar	2	100							🟩	🟩																						
Familarise using wav2vec & Kaldi	2	10									🟨	🟨																				
Thesis A Final Report	2	100									🟩	🟩																				
Dataset Preperation	1	0											🟨																			
Implement XLS-R with basic pooling + linear layer	2	0												🟨	🟨																	
Test accuracy of various audio lengths with 100hrs of training on downstream model	2	0														🟨	🟨															
Implement CNN/BiLTSM Linear Downstream Model	2	0																🟨	🟨													
Implement with wav2vec 2.0 pretrained model	2	0																	🟨	🟨												
Test accuracy using various amounts of training data	2	0																		🟨	🟨											
Thesis B Report	3	0																		🟨	🟨	🟨										
Review of Expected Outcomes	2	0																			🟨	🟨										
Makes any changes that need to be made to the model or the focus of the thesis	1	0																					🟨									
																							🟨									
Train using fine grain dialect classification groups	2	0																						🟨	🟨							
Test accuracy of both options of pretrained and downstream models	2	0																							🟨	🟨						
Test using codeswitch input data	2	0																									🟨	🟨				
Final verification and review	2	0																											🟨	🟨		
Final Report	3	0																												🟨	🟨	🟨
Final Poster	2	0																												🟨	🟨	

Figure 6.1: Gantt Chart.

6.2.2 Task Breakdown

1. Dataset preprocessing.
2. Build basic model (XLS-R, Linear Layer + Pooling).
3. Determine the minimum amount of labelled data required through training with:
    - 100hrs
    - 10hrs
    - 1hr
    - 10mins
4. Benchmark against wav2vec 2.0.
5. Build model with CNN downstream architecture.
    - XLS-R with labelled data:
        - 100hrs
        - 10hrs
        - 1hr
        - 10mins
    - wav2vec 2.0 with labelled data:
        - 100hrs
        - 10hrs

        - 1hr
        - 10mins

6. Build model with BiLSTM downstream architecture.
    - XLS-R trained with labelled data:
        - 100hrs
        - 10hrs
        - 1hr
        - 10mins
    - wav2vec 2.0 with labelled data:
        - 100hrs
        - 10hrs
        - 1hr
        - 10mins
    - Experiment fine tuning with unlabeled data

6.3 Thesis C

6.3.1 Aims

Thesis C will extend upon the conclusions found in Thesis B, assessing the robustness and performance of the DID designed through using a finer set of 17 dialect classifications (DZA, EGY, IRQ, JOR, SAU, KWT, LBN, LBY, MRT, MAR, OMN, PSE, QAT, SDN, SYR, ARE, YEM).
The goals of Thesis C are as follows:

- Prove a model can be used to identify a finer set of dialectal groups
- Evaluate the performance of the DID with utterances of various lengths.

6.3.2 Task Breakdown

1. Assess the most accurate model from Thesis B and adapt it to include the 17 classifier groups.
2. Observe its accuracy with labelled training data:
    - 10hrs
    - 1hr
    - 10mins
3. Evaluate the DID's accuracy with utterances of:
    - 10hrs
    - 1hr
    - 10mins

6.4 Possible Challenges

The major challenges likely to be encountered in this thesis are:
Quality of Dataset: There is acoustic variation across the dataset such as environmental noise and variation in the volume. The acoustic variation could lead to the network overfitting and classifying based upon acoustic rather than linguistic dialectal differences. As a non-Arabic speaker, I will be unable to audibly verify the accuracy of the dataset or identify any mislabeling errors. Filtering will be used to mitigate the acoustic variation within the dataset and since, multiple papers have utlised the ADI17 dataset there is confidence that the dataset is reliable.

Time: Machine learning systems especially when training with large amounts of data (100hrs) take time to process. So, tuning metaparameters and making design changes may take several hours to present results. This time delay will make debugging challenging, although by anticipating for this time delay and staying organised the effect this will have can be reduced.

Tuning Metaparamaters: There are various different metaparameters, pruning methods and optimisers which can be used in designing a machine learning model. Tuning these will take a significant amount of time, writing a script to cycle through different metaparameters can increase the efficiency of this task.

Negative Transfer: In transfer learning, there is the risk that previous knowledge of a pretrained model may damage the performance of specific task. This may cause significant issues so, by testing multiple amounts of training data and models, the occurance of negative transference can be assessed. [62]

Chapter 7

Conclusion

This report has explored the challenges in designing a reliable Arabic DID and the need for more accurate systems. It proposes using semi-self supervised transfer learning as an approach to improve Arabic DIDs. It then has provided a literature review of current methodologies for LID and DID systems. As well as the justification behind the pretrained models and downstream models which will be researched during this thesis. It has outlined the proposed method and then went into some preliminary work performed, particularly some data analysis on the ADI17 dataset. Finally, this report has provided a timeline, task break down and the challenges that may be encountered during this thesis were examined.

Bibliography

[1] XLS-R: Self-supervised speech processing for 128 languages, November.

[2] Muhammad Abdul-Mageed, Chiyu Zhang, Houda Bouamor, and Nizar Habash. NADI 2020: The First Nuanced Arabic Dialect Identification Shared Task. *arXiv:2010.11334 [cs]*, November 2020. arXiv: 2010.11334.

[3] Musatafa Abbas Abbood Albadr, Sabrina Tiun, Fahad Taha AL-Dhief, and Mahmoud A. M. Sammour. Spoken language identification based on the enhanced self-adjusting extreme learning machine approach. *PLOS ONE*, 13(4):e0194770, April 2018.

[4] Alexei Baevski, Wei-Ning Hsu, Qiantong Xu, Arun Babu, Jiatao Gu, and Michael Auli. The first high-performance self-supervised algorithm that works for speech, vision, and text, January 2022.

[5] David Alfter. Language Segmentation. *arXiv:1510.01717 [cs]*, October 2015. arXiv: 1510.01717.

[6] Zainab Alhakeem and Hong-Goo Kang. Confidence Learning from Noisy Labels for Arabic Dialect Identification. In *2021 36th International Technical Conference on Circuits/Systems, Computers and Communications (ITC-CSCC)*, pages 1–4, 2021.

[7] Ahmed Ali, Shammur Chowdhury, Amir Hussein, and Yasser Hifny. Arabic Code-Switching Speech Recognition using Monolingual Data. *arXiv:2107.01573 [cs, eess]*, July 2021. arXiv: 2107.01573.

[8] Ahmed Ali, Suwon Shon, Younes Samih, Hamdy Mubarak, Ahmed Abdelali, James Glass, Steve Renals, and Khalid Choukri. The MGB-5 Challenge: Recognition and Dialect Identification of Dialectal Arabic Speech. In *2019 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, pages 1026–1033, 2019.

[9] Arun Babu, Changhan Wang, Andros Tjandra, Kushal Lakhotia, Qiantong Xu, Naman Goyal, Kritika Singh, Patrick von Platen, Yatharth Saraf, Juan Pino, Alexei Baevski, Alexis Conneau, and Michael Auli. XLS-R: Self-supervised Cross-lingual Speech Representation Learning at Scale. *arXiv:2111.09296 [cs, eess]*, December 2021. arXiv: 2111.09296.

[10] Arun Babu, Changhan Wang, Andros Tjandra, Kushal Lakhotia, Qiantong Xu, Naman Goyal, Kritika Singh, Patrick von Platen, Yatharth Saraf, Juan Pino, Alexei Baevski, Alexis Conneau, and Michael Auli. XLS-R: Self-supervised Cross-lingual Speech Representation Learning at Scale. *arXiv:2111.09296 [cs, eess]*, December 2021. arXiv: 2111.09296.

[11] Alexei Baevski, Michael Auli, and Abdelrahman Mohamed. Effectiveness of self-supervised pre-training for speech recognition. *arXiv:1911.03912 [cs]*, May 2020. arXiv: 1911.03912.

[12] Alexei Baevski, Henry Zhou, Abdelrahman Mohamed, and Michael Auli. wav2vec 2.0: A Framework for Self-Supervised Learning of Speech Representations. *arXiv:2006.11477 [cs, eess]*, October 2020. arXiv: 2006.11477.

[13] Trapit Bansal, Rishikesh Jha, Tsendsuren Munkhdalai, and Andrew McCallum. Self-Supervised Meta-Learning for Few-Shot Natural Language Classification Tasks. *arXiv:2009.08445 [cs]*, November 2020. arXiv: 2009.08445.

[14] Fadi Biadsy, Julia Hirschberg, and Daniel Ellis. Dialect and Accent Recognition Using Phonetic-Segmentation Supervectors. pages 745–748, January 2011.

[15] Fadi Biadsy, Julia Hirschberg, and Nizar Habash. Spoken Arabic dialect identification using phonotactic modeling. In *Proceedings of the EACL 2009 Workshop on Computational Approaches to Semitic Languages - Semitic '09*, page 53, Athens, Greece, 2009. Association for Computational Linguistics.

[16] Hynek Bořil, Abhijeet Sangwan, and John H. L. Hansen. Arabic dialect identification - "is the secret in the silence?" and other observations. In *Interspeech 2012*, pages 30–33. ISCA, September 2012.

[17] Mathilde Caron, Piotr Bojanowski, Armand Joulin, and Matthijs Douze. Deep Clustering for Unsupervised Learning of Visual Features. *arXiv:1807.05520 [cs]*, March 2019. arXiv: 1807.05520.

[18] Shammur Absar Chowdhury, Amir Hussein, Ahmed Abdelali, and Ahmed Ali. Towards One Model to Rule All: Multilingual Strategy for Dialectal Code-Switching Arabic ASR. *arXiv:2105.14779 [cs, eess]*, July 2021. arXiv: 2105.14779.

[19] Yu-An Chung, Yu Zhang, Wei Han, Chung-Cheng Chiu, James Qin, Ruoming Pang, and Yonghui Wu. W2v-BERT: Combining Contrastive Learning and Masked Language Modeling for Self-Supervised Speech Pre-Training. *arXiv:2108.06209 [cs, eess]*, September 2021. arXiv: 2108.06209.

[20] Savelie Cornegruta, Robert Bakewell, Samuel Withey, and Giovanni Montana. Modelling Radiological Language with Bidirectional Long Short-Term Memory Networks. *arXiv:1609.08409 [cs, stat]*, September 2016. arXiv: 1609.08409.

[21] Bilal Dendani, Halima Bahi, and Toufik Sari. Self-Supervised Speech Enhancement for Arabic Speech Recognition in Real-World Environments. *Traitement du Signal*, 38(2):349–358, April 2021.

[22] Zhiyun Fan, Meng Li, Shiyu Zhou, and Bo Xu. Exploring wav2vec 2.0 on speaker verification and language identification. 2020. Publisher: arXiv Version Number: 2.

[23] Nizar Habash, Owen Rambow, Mona Diab, and Reem Kanjawi-Faraj. Guidelines for annotation of Arabic dialectness. In *Proceedings of the LREC Workshop on HLT & NLP within the Arabic world*, pages 49–53, 2008.

[24] Wei-Ning Hsu, Benjamin Bolte, Yao-Hung Hubert Tsai, Kushal Lakhotia, Ruslan Salakhutdinov, and Abdelrahman Mohamed. HuBERT: Self-Supervised Speech Representation Learning by Masked Prediction of Hidden Units. *arXiv:2106.07447 [cs, eess]*, June 2021. arXiv: 2106.07447.

[25] Amir Hussein, Shinji Watanabe, and Ahmed Ali. Arabic Speech Recognition by End-to-End, Modular Systems and Human. *arXiv:2101.08454 [cs, eess]*, June 2021. arXiv: 2101.08454.

[26] Tommi Jauhiainen, Marco Lui, Marcos Zampieri, Timothy Baldwin, and Krister Lindén. Automatic Language Identification in Texts: A Survey. *arXiv:1804.08186 [cs]*, November 2018. arXiv: 1804.08186.

[27] Jonathan Bgn. HuBERT: How to Apply BERT to Speech, Visually Explained, October 2021.

[28] Muhammad Khalifa, Hesham Hassan, and Aly Fahmy. Zero-Resource Multi-Dialectal Arabic Natural Language Understanding. *International Journal of Advanced Computer Science and Applications*, 12(3), 2021. arXiv: 2104.06591.

[29] Zhenzhong Lan, Mingda Chen, Sebastian Goodman, Kevin Gimpel, Piyush Sharma, and Radu Soricut. ALBERT: A Lite BERT for Self-supervised Learning of Language Representations. *arXiv:1909.11942 [cs]*, February 2020. arXiv: 1909.11942.

[30] Hunter Lang and Hoifung Poon. Self-supervised self-supervision by combining deep learning and probabilistic logic. *arXiv:2012.12474 [cs, stat]*, December 2020. arXiv: 2012.12474.

[31] Wanqiu Lin, Maulik Madhavi, Rohan Kumar Das, and Haizhou Li. Transformer-based Arabic Dialect Identification. *arXiv:2011.00699 [eess]*, November 2020. arXiv: 2011.00699.

[32] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. RoBERTa: A Robustly Optimized BERT Pretraining Approach. *arXiv:1907.11692 [cs]*, July 2019. arXiv: 1907.11692.

[33] Ignacio Lopez-Moreno, Javier Gonzalez-Dominguez, David Martinez, Oldřich Plchot, Joaquin Gonzalez-Rodriguez, and Pedro J. Moreno. On the use of deep feedforward neural networks for automatic language identification. *Computer Speech & Language*, 40:46–59, November 2016.

[34] Khaled Lounnas, Mourad Abbas, and Mohamed Lichouri. Building a Speech Corpus based on Arabic Podcasts for Language and Dialect Identification. September 2019.

[35] Xiaoxiao Miao and Ian McLoughlin. LSTM-TDNN with convolutional front-end for Dialect Identification in the 2019 Multi-Genre Broadcast Challenge. *arXiv:1912.09003 [cs, eess]*, December 2019. arXiv: 1912.09003.

[36] Xiaoxiao Miao, Ian McLoughlin, and Yonghong Yan. A New Time-Frequency Attention Mechanism for TDNN and CNN-LSTM-TDNN, with Application to Language Identification. In *Interspeech 2019*, pages 4080–4084. ISCA, September 2019.

[37] Michael Ellis. *Accent Identification for English Speakers*. PhD thesis, University of New South Wales.

[38] Omar Mohamed and Salah A. Aly. Arabic Speech Emotion Recognition Employing Wav2vec2.0 and HuBERT Based on BAVED Dataset. *arXiv:2110.04425 [cs]*, October 2021. arXiv: 2110.04425.

[39] Hamdy Mubarak, Amir Hussein, Shammur Absar Chowdhury, and Ahmed Ali. QASR: QCRI Aljazeera Speech Resource – A Large Scale Annotated Arabic Speech Corpus. *arXiv:2106.13000 [cs, eess]*, June 2021. arXiv: 2106.13000.

[40] Scott Novotney, Rich Schwartz, and Sanjeev Khudanpur. Unsupervised Arabic dialect adaptation with self-training. In *Interspeech 2011*, pages 541–544. ISCA, August 2011.

[41] Charles Perreault and Sarah Mathew. Dating the Origin of Language Using Phonemic Diversity. *PLoS ONE*, 7(4):e35289, April 2012.

[42] Jouni Pohjalainen, Fabien Fabien Ringeval, Zixing Zhang, and Björn Schuller. Spectral and Cepstral Audio Noise Reduction Techniques in Speech Emotion Recognition. In *Proceedings of the 24th ACM international conference on Multimedia*, pages 670–674, Amsterdam The Netherlands, October 2016. ACM.

[43] G. Ramesh, C. Shiva Kumar, and K. Sri Rama Murty. Self-Supervised Phonotactic Representations for Language Identification. In *Interspeech 2021*, pages 1514–1518. ISCA, August 2021.

[44] Shauna Revay and Matthew Teschke. Multiclass Language Identification using Deep Learning on Spectral Images of Audio Signals. *arXiv:1905.04348 [cs, eess]*, May 2019. arXiv: 1905.04348.

[45] Mohammad Salameh, Houda Bouamor, and Nizar Habash. Fine-Grained Arabic Dialect Identification. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 1332–1344, Santa Fe, New Mexico, USA, August 2018. Association for Computational Linguistics.

[46] Younes Samih, Mohammed Attia, Mohamed Eldesouki, Ahmed Abdelali, Hamdy Mubarak, Laura Kallmeyer, and Kareem Darwish. A Neural Architecture for Dialectal Arabic Segmentation. In *Proceedings of the Third Arabic Natural Language Processing Workshop*, pages 46–54, Valencia, Spain, 2017. Association for Computational Linguistics.

[47] Younes Samih, Suraj Maharjan, Mohammed Attia, Laura Kallmeyer, and Thamar Solorio. Multilingual Code-switching Identification via LSTM Recurrent Neural Networks. In *Proceedings of the Second Workshop on Computational Approaches to Code Switching*, pages 50–59, Austin, Texas, 2016. Association for Computational Linguistics.

[48] Sanket Shah, Sunayana Sitaram, and Rupeshkumar Mehta. FirstWorkshop on Speech Processing for Code-switching in Multilingual Communities: Shared Task on Code-switched Spoken Language Identification. pages 24–28. Microsoft Research India, Microsoft Corporation, October 2020.

[49] Suwon Shon, Ahmed Ali, and James Glass. Convolutional Neural Networks and Language Embeddings for End-to-End Dialect Recognition. *arXiv:1803.04567 [cs, eess]*, April 2018. arXiv: 1803.04567.

[50] Suwon Shon, Ahmed Ali, Younes Samih, Hamdy Mubarak, and James Glass. ADI17: A Fine-Grained Arabic Dialect Identification Dataset. In *ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 8244–8248, Barcelona, Spain, May 2020. IEEE.

[51] Suwon Shon, Wei-Ning Hsu, and James Glass. Unsupervised Representation Learning of Speech for Dialect Identification. *arXiv:1809.04458 [cs, eess]*, September 2018. arXiv: 1809.04458.

[52] Guan-Lip Soon, Nur-Hana Samsudin, and Dennis Lim. Evaluating the Effect of Multiple Filters in Automatic Language Identification without Lexical Knowledge. *International Journal of Advanced Computer Science and Applications*, 11(10), 2020.

[53] Andros Tjandra, Diptanu Gon Choudhury, Frank Zhang, Kritika Singh, Alexis Conneau, Alexei Baevski, Assaf Sela, Yatharth Saraf, and Michael Auli. Improved Language Identification Through Cross-Lingual Self-Supervised Learning. *arXiv:2107.04082 [cs, eess]*, October 2021. arXiv: 2107.04082.

[54] Liang-Hsuan Tseng, Yu-Kuan Fu, Heng-Jui Chang, and Hung-yi Lee. Mandarin-English Code-switching Speech Recognition with Self-supervised Speech Representation Models. *arXiv:2110.03504 [cs, eess]*, October 2021. arXiv: 2110.03504.

[55] Michael D. Tyler and Anne Cutler. Cross-language differences in cue use for speech segmentation. *The Journal of the Acoustical Society of America*, 126(1):367–376, July 2009.

[56] Jörgen Valk and Tanel Alumäe. VoxLingua107: a Dataset for Spoken Language Recognition. *arXiv:2011.12998 [eess]*, November 2020. arXiv: 2011.12998.

[57] Anshul Wadhawan. Dialect Identification in Nuanced Arabic Tweets Using Farasa Segmentation and AraBERT. *arXiv:2102.09749 [cs]*, February 2021. arXiv: 2102.09749.

[58] Yingzhi Wang, Abdelmoumene Boumadane, and Abdelwahab Heba. A Fine-tuned Wav2vec 2.0/HuBERT Benchmark For Speech Emotion Recognition, Speaker Verification and Spoken Language Understanding. *arXiv:2111.02735 [cs, eess]*, November 2021. arXiv: 2111.02735.

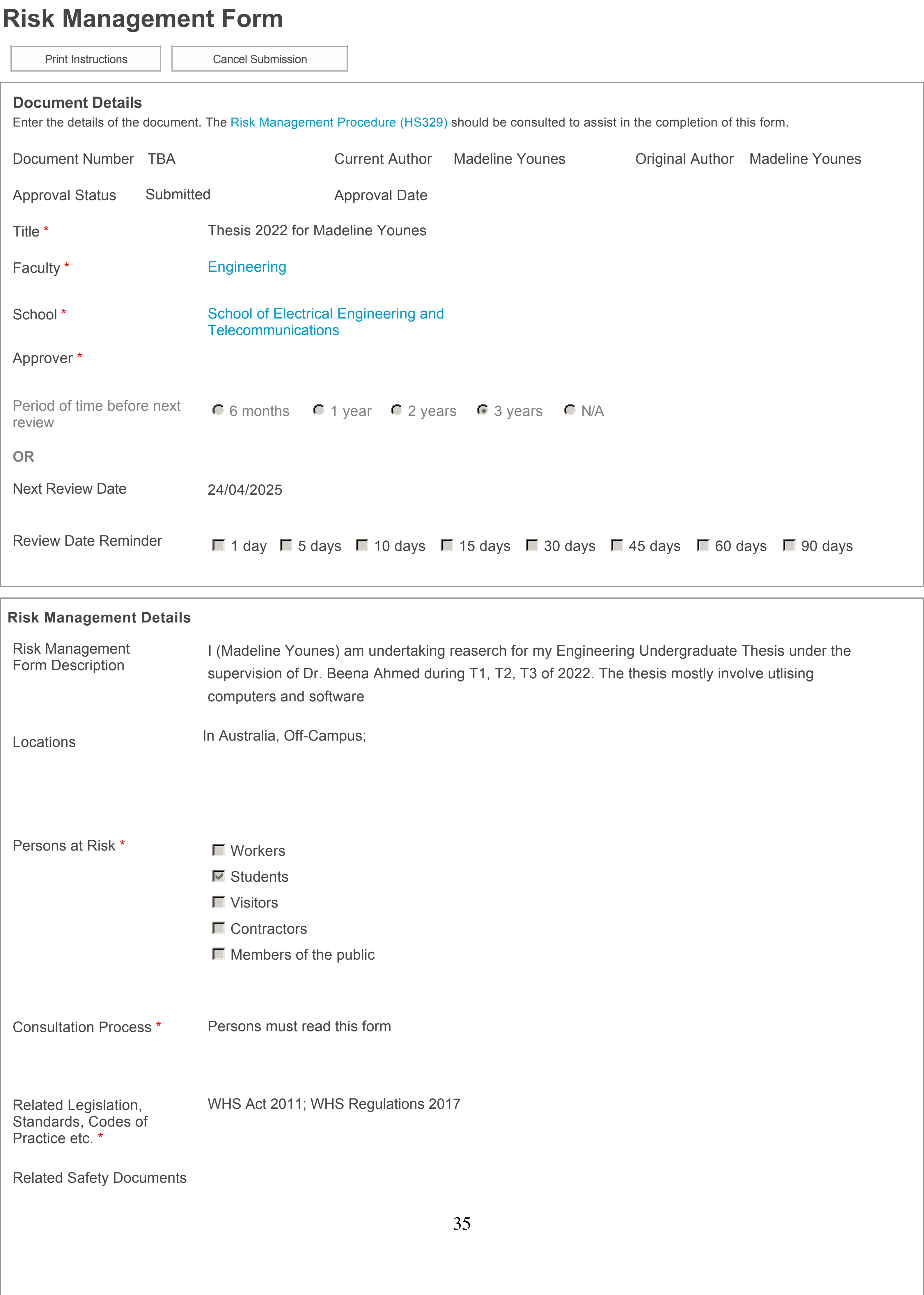
[59] Omar F. Zaidan and Chris Callison-Burch. Arabic Dialect Identification. *Computational Linguistics*, 40(1):171–202, March 2014.

[60] Chiyu Zhang and Muhammad Abdul-Mageed. No Army, No Navy: BERT Semi-Supervised Learning of Arabic Dialects. In *Proceedings of the Fourth Arabic Natural Language Processing Workshop*, pages 279–284, Florence, Italy, 2019. Association for Computational Linguistics.

[61] Qian Zhang and John H. L. Hansen. Language/Dialect Recognition Based on Unsupervised Deep Learning. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 26(5):873–882, 2018.

[62] Wen Zhang, Lingfei Deng, Lei Zhang, and Dongrui Wu. A Survey on Negative Transfer. *arXiv:2009.00909 [cs, stat]*, August 2021. arXiv: 2009.00909.

Appendix A - Risk Assessment Form



Risk Management Form

Document Details

Enter the details of the document. The Risk Management Procedure (HS329) should be consulted to assist in the completion of this form.

Document Number	TBA	Current Author	Madeline Younes	Original Author	Madeline Younes
Approval Status	Submitted	Approval Date			

Title * Thesis 2022 for Madeline Younes

Faculty * Engineering

School * School of Electrical Engineering and Telecommunications

Approver *

Period of time before next review ○ 6 months ○ 1 year ○ 2 years ● 3 years ○ N/A

OR

Next Review Date 24/04/2025

Review Date Reminder □ 1 day □ 5 days □ 10 days □ 15 days □ 30 days □ 45 days □ 60 days □ 90 days

Risk Management Details

Risk Management Form Description: I (Madeline Younes) am undertaking reaserch for my Engineering Undergraduate Thesis under the supervision of Dr. Beena Ahmed during T1, T2, T3 of 2022. The thesis mostly involve utlising computers and software

Locations: In Australia, Off-Campus;

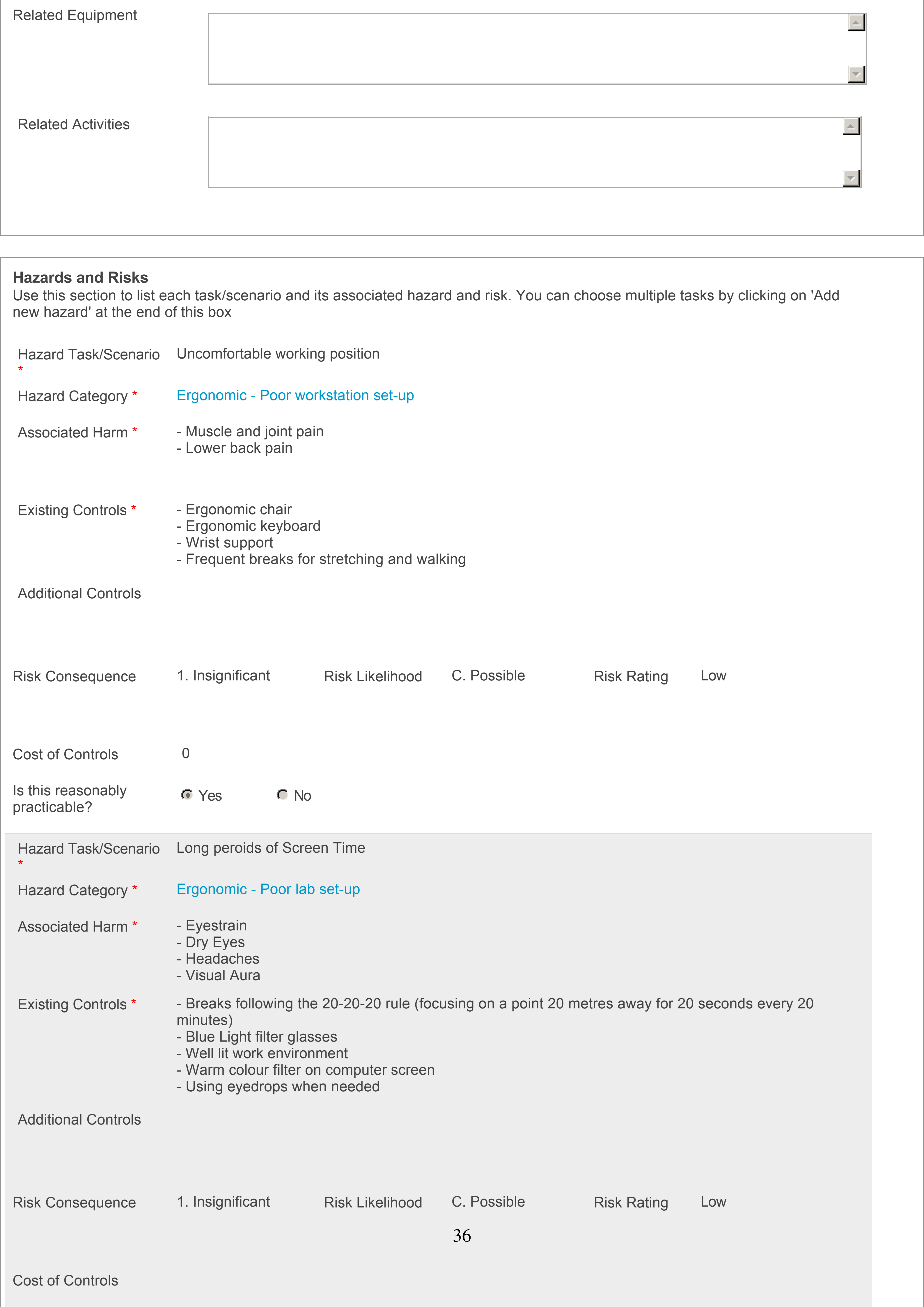
Persons at Risk *

- □ Workers
- ☑ Students
- □ Visitors
- □ Contractors
- □ Members of the public

Consultation Process * Persons must read this form

Related Legislation, Standards, Codes of Practice etc. * WHS Act 2011; WHS Regulations 2017

Related Safety Documents



Related Equipment	
Related Activities	

Hazards and Risks
Use this section to list each task/scenario and its associated hazard and risk. You can choose multiple tasks by clicking on 'Add new hazard' at the end of this box

Hazard Task/Scenario *	Uncomfortable working position
Hazard Category *	Ergonomic - Poor workstation set-up
Associated Harm *	- Muscle and joint pain - Lower back pain
Existing Controls *	- Ergonomic chair - Ergonomic keyboard - Wrist support - Frequent breaks for stretching and walking
Additional Controls	

Risk Consequence	1. Insignificant	Risk Likelihood	C. Possible	Risk Rating	Low

Cost of Controls	0
Is this reasonably practicable?	● Yes ○ No

Hazard Task/Scenario *	Long peroids of Screen Time
Hazard Category *	Ergonomic - Poor lab set-up
Associated Harm *	- Eyestrain - Dry Eyes - Headaches - Visual Aura
Existing Controls *	- Breaks following the 20-20-20 rule (focusing on a point 20 metres away for 20 seconds every 20 minutes) - Blue Light filter glasses - Well lit work environment - Warm colour filter on computer screen - Using eyedrops when needed
Additional Controls	

Risk Consequence	1. Insignificant	Risk Likelihood	C. Possible	Risk Rating	Low

Cost of Controls

Is this reasonably practicable?	● Yes ○ No

Hazard Task/Scenario *	Excessive Worload				
Hazard Category *	Psychological - Excessive workload				
Associated Harm *	- Stress - Headaches - Lack of sleep				
Existing Controls *	- Regular breaks - Regular meetings with supervisor - Regular self reflection - Setting reasonable goals and expectations				
Additional Controls					
Risk Consequence	1. Insignificant	Risk Likelihood	C. Possible	Risk Rating	Low
Cost of Controls					
Is this reasonably practicable?	● Yes ○ No				

Other Risk Management Details

Date All Controls Implemented	24/04/2022
Emergency Procedures *	Contact 000, or emergency contacts on mobilephone or those provided to the university. In the case where medical history is needed contact local GP (Dr Buddy Maroun).
Competency and Training Required	
Competency Levels *	1. Read Document

Only add descriptions below for competency levels chosen above

Training Description	
Knowledge Test Description	
License/Cert Description	
Other Competency Description	
Additional Documents	