Speech-based Traumatic Brain Injury (TBI) Assessment using Deep Learning Methods with Limited Labeled Data

by
Apiwat Ditthapron

A Thesis
Submitted to the Faculty
of the
WORCESTER POLYTECHNIC INSTITUTE
In partial fulfillment of the requirements for the
Degree of Master of Science
in
Computer Science
by

May 2021

APPROVED:

Professor Emmanuel O. Agu, Thesis Advisor

Professor Adam C. Lammert, Thesis Co-Advisor

Professor Elke A. Rundensteiner, Thesis Reader

Professor Craig E. Wills, Head of Department

ABSTRACT

People afflicted with Traumatic Brain Injury (TBI) experience long-term impairments, which require follow-up assessments or rehospitalization in some cases. To promote recovery, continuous monitoring of TBI patients with long-term impairments is an area requiring urgent research in public health care. This master thesis proposes a Deep Neural Network (DNN) system for non-invasive, speech-based assessment of long-term impairments following TBI that runs passively on smartphones. Notably, we tackle the overfitting problem that arises from an insufficient amount of TBI speech, which is infrequently collected and expensive to acquire. Overfitting prevents the DNN from learning generalized features of TBI speech and hinders the TBI detection accuracy. In this master's thesis, we investigate three learning methods leverage knowledge from other datasets to mitigate the limited data problem in order to improve DNN-based TBI assessment accuracy. Specifically, we investigate transfer learning, multi-task learning, and meta-learning for improving our proposed cascading DNN for TBI speech assessment that combines sequential features with a backbone model. The results indicate that all three limited labeled data learning methods mitigate the overfitting problem and improve the TBI classification accuracy by 34% and TBI regression error by 31%. Moreover, we explore the related few-shot problem that determines the minimum amount of data required and extrapolate our results in order to estimate a full trajectory of expected performance for various input data sizes for each limited data method.

ACKNOWLEDGMENTS

First, I would like to express my sincere gratitude to my thesis advisors, Prof. Emmanuel O. Agu and Prof. Adam C. Lammert, for their guidance and support of my research. Their advice and feedback helped me shape my research and writing of this thesis in the right direction. I would like to thank Prof. Elke A. Rundensteiner for being the reader of my thesis reader and attending my thesis presentation.

I would like to thank the Defense Advanced Research Projects Agency (DARPA) and the Royal Thai government for financial support of my research at Worcester Polytechnic Institute. Lastly, I would like to thank my family and friend for their supports and encouragement throughout my years of study.

"This material is based on research sponsored by DARPA under agreement number FA8750-18-2-0077. The U.S. Government is authorized to reproduce and distribute reprints for Governmental purposes notwithstanding any copyright notation thereon." "The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies or endorsements, either expressed or implied, of DARPA or the U.S. Government."

NOMENCLATURE

TBI	Traumatic Brain Injury
mTBI	mild Traumatic Brain Injury
PTSD	ost-Traumatic Stress Disorder
GCS	Glasgow Coma Scale
DNN	Deep Neural Network
CNN	Convolution Neural Network
GRU	Gated recurrent units
ReLU	Rectified Linear Unit
RNN	Recurrent Neural Network
GAN	Generative Adversarial Network
MFCC	Mel-frequency cepstral coefficients
VTLN	Vocal Tract Length Normalization
BAC	Balanced ACuracy
AUC	Area under the ROC Curve
RMSE	Root Mean Square Error
MAE	Mean Absolute Error
GPU	Graphics Processing Unit
WOZ	Wizard-of-OZ dataset

TABLE OF CONTENTS

Page

ABSTRACT ii

ACKNOWLEDGMENTS iii

NOMENCLATURE iv

TABLE OF CONTENTS v

LIST OF FIGURES vii

LIST OF TABLES viii

1. INTRODUCTION 1

1.1 Background on TBI 1
1.2 TBI assessment from speech 1
1.3 Analyzing speech to assess TBI using Machine Learning 2
1.4 DNN-based TBI assessment from speech 2
1.5 The limited data problem in DNN-based TBI speech assessment 3
1.6 Limited data mitigation learning methods 7
1.7 Thesis overview 8
1.8 Thesis research questions 8
1.9 Novelty of this thesis in relation to prior work 9
1.10 Thesis contributions 10
1.11 Thesis outline 11

2. BACKGROUND AND RELATED WORK 12

2.1 Traumatic brain injury 12
2.2 TBI-impaired speech: background 12
2.3 Prior work on assessing TBI from speech 13
2.4 Machine learning analyses to assess TBI from speech 14
2.5 Learning techniques in DNN 14
2.5.1 Single task learning (Baseline) 14
2.5.2 Transfer learning 15
2.5.3 Multi-task learning 16
2.5.4 Meta-learning 18
2.6 DNNs and audio-based backbone network 18
2.6.1 Convolution Neural Network (CNN) 18

2.6.2 Gated Recurrent Unit (GRU) 19
2.6.3 VGGish 20
2.6.4 Wav2Vec 23
2.6.5 SincNet 25

3. METHODOLOGY 28

3.1 Cascading DNN for TBI classification and regression 29
3.2 Audio datasets utilized 30
3.2.1 Source dataset 31
3.2.2 Target dataset 32
3.2.3 Data preprocessing and feature extraction 33
3.3 Evaluation method 34
3.3.0.1 Transfer learning 35
3.3.0.2 Multi-task learning 35
3.3.0.3 Meta-learning 36
3.3.1 Evaluation metrics and baseline models 36
3.3.1.1 Evaluation Metrics 36
3.3.1.2 Baseline: single task learning and conventional machine learning . 37
3.3.1.3 Extrapolation of TBI classification accuracy on numbers of training sample 37

4. RESULT 39
4.0.1 Transfer learning 39
4.0.2 Multi-task learning 41
4.0.3 Meta-learning 43

5. DISCUSSION 50

6. CONCLUSION 53

6.1 Future work 53

REFERENCES 54

LIST OF FIGURES

FIGURE Page

1.1 Passive speech-based TBI assessment 2

1.2 Principle of parsimony 4

1.3 Trade-off between model complexity and error 5

1.4 Early stopping in deep learning 6

2.1 Architecture of Convolution Neural Network (CNN) 19

2.2 Architecture of Gated Recurrent Unit (GRU) 20

2.3 Architecture of VGGish model 21

2.4 Mel-spectrogram 24

2.5 Architecture of WAV2VEC model 25

2.6 SincNet architecture 26

3.1 Proposed learning methods for TBI detection with limited-data 28

3.2 Proposed cascading DNN architecture 30

3.3 Source and target datasets for limited-data learning method 33

4.1 Training and validation losses using transfer learning 41

4.2 Training and validation losses with single-task learning (baseline) 42

4.3 Training and validation losses using multi-task learning 44

4.4 Few-shot learning in TBI classification 47

4.5 Projection of TBI detection improvement from an increasing size of training sample using linear fit 48

4.6 Projection of TBI detection improvement from an increasing size of training sample using quadratic fit 49

LIST OF TABLES

TABLE Page

2.1 Previous studies in speech-based TBI detection 14

2.2 Equation notation 15

2.3 VGGish model architecture 22

2.4 Wav2Vec model architecture (encoder network) 24

2.5 SincNet model architecture 27

3.1 Input and output dimensions of the backbone models 29

3.2 Hyperparameters in training the cascading DNN 30

3.3 Summary of source and target datasets used to train the TBI detection model 34

4.1 TBI classification results using transfer learning technique 40

4.2 TBI regression results using transfer learning technique 40

4.3 TBI classification results using multi-task learning technique 43

4.4 TBI regression results using multi-task learning technique 45

4.5 TBI classification results using meta-learning technique 46

4.6 TBI regression results using meta-learning technique 46

5.1 TBI classification improvement in SincNet using limited-data learning methods 52

5.2 TBI regression improvement in Wav2Vec using limited-data learning methods 52

1. INTRODUCTION

1.1 Background on TBI

According to the United States (U.S.) Department of Health and Human Services (DHHS), at least 1.4 million people suffer from TBI in the U.S. annually [1]. Leading causes of TBI include falls, motor vehicle traffic, and assault. In terms of severity, Glasgow Coma Scale (GCS) has been used as the gold standard to classify TBI as mild, moderate, or severe, based on person's level of consciousness after the injury. Mild-TBI (mTBI), also known as a concussion, is the most common type of TBI in in both civilian and military populations [1]. Although a patient with mTBI has little to no physical injury, long-term TBI sequelae can occur in the form of neuro-cognitive deficits that disrupt a patient's life. Common TBI symptoms include reduced processing speed, poor concentration, memory difficulty, and increased fatigue. Some patients require rehospitalization with common causes of infectious, neurological, and neurosurgical disorders [2]. Hammond et al. indicated the needs of monitoring and screening for TBI symptoms after patients are discharged [2]. To this end, we attempt to develop a speech-based TBI assessment that passively monitors sequelae of TBI on smartphones.

1.2 TBI assessment from speech

Typically, TBI assessment is performed after the injury to measure neuro-cognitive deficits through a cognitive exam, neurological exam, neuroimaging, and ocular-motor screening. These kinds of TBI assessment are costly and invasive, making their use limited to the severe case [3]. To increase the accessibility of TBI assessment, a noninvasive TBI screening was previously proposed to detect TBI using sensing data, such as speech and gait [4, 5, 6, 7, 8, 9, 10]. Among the noninvasive features, speech has recently gained more attraction as individual with TBI frequently manifests symptoms of speech and language disorders associated with TBI in the form of poor speech production and speech comprehension [4, 5, 6, 7, 11]. Moreover, there is evidence indicating high correlations between language disorder and cognitive impairment, which is prevalence in

all TBI severities [11]. The TBI medical literature often lists speech and language disorders, such as dysarthria, apraxia and aphasia, as prevalent disorders associated with TBI [4]. This study aims to monitor individual's speech after the TBI in natural setting on the smartphone, as visualized in Figure 1.1.

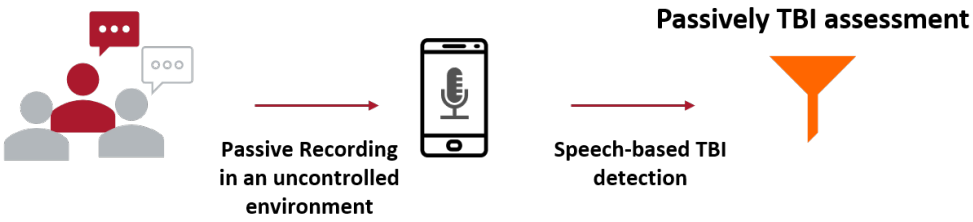

Figure 1.1: Proposed pipeline for passive speech-based TBI assessment

1.3 Analyzing speech to assess TBI using Machine Learning

Machine learning analysis of audio produced by TBI subjects has been proposed to facilitate continuous monitoring. In an area of automatic speech-based TBI detection, Falcone et al. exploited the acoustic characteristics of vowels to classify TBI and found that fundamental speech components such as formants, jitter, shimmer, and harmonic-to-noise ratio are prominent speech features of TBI patients [6]. Helfer et al. investigated the use of time-delay correlation of speech formants to predict patients' IMPACT score, one of the assessments for acute TBI [7]. Both sets of features mentioned earlier were classified using the Support Vector Machine (SVM) classifier [12]. These previous studies prioritized developing a feature set that improves TBI assessment performance from speech and did not consider Deep neural networks (DNNs) that have the potential to work as an end-to-end model without feature engineering.

1.4 DNN-based TBI assessment from speech

DNNs have demonstrated performance breakthroughs in various domains, especially in computer vision and natural language processing [13, 14, 15]. Unlike traditional machine learning

models, DNNs often perform better on raw data than on hand-crafted features [16], which obviates the need for feature engineering. Consequently, most machine learning research interest has shifted from machine learning that involves feature engineering to engineering and optimizing DNN. DNNs can learn robust feature representations through an end-to-end learning mechanism. However, training the DNN requires abundant training data to obtain a robust result and avoid overfitting problem.

According to bias-variance dilemma (Figure 1.2), having *too many* parameters leads to high variance (uncertainty) and low bias (accurate to provided data) while having *too few* parameters leads to low variance and high bias [17]. In practice, an optimum point in the trade-off is located by finding a saddle point on the validation error curve, plotted in Figure 1.3. A decreasing in validation loss often indicates an *underfitting* while an increasing indicates an *overfitting*. Generally, overfitting occurs in such a model with too many parameters, contradicting with underfitting that has too few parameters. The number of parameters is considered as a hyperparameters in DNN optimization and is determined on the conceptual trade-off between accuracy and certainty. Alternatively, *early stopping* can be executed to prevent the overfitting in a complex model, which stops the training when validation loss is increasing while training loss is still decreasing [18] as visualized in Figure 1.4.

1.5 The limited data problem in DNN-based TBI speech assessment

In practice, Zhang et al. demonstrated that a DNN with two layers of $2n + d$ parameters is adequate to represent a sample of size n in d dimensions [19]. Hence, training the model with less than n samples will lead to an *overfitting* problem, especially in the medical application that has data scarcity problem. Collecting medical data is costly due to the involvement of medical experts, the use of expensive instrumentation and the patients' potential discomfort [20]. Although the process of speech recording costs less than neuroimaging, the TBI discourse still requires a speech-language pathologist to perform a series of speech and language assessments, which has a charge starting at 39$ per hour [21]. Moreover, collecting medical data for a research needs a sufficient number of patients over broad demographics, such as gender, age, and education level, increasing

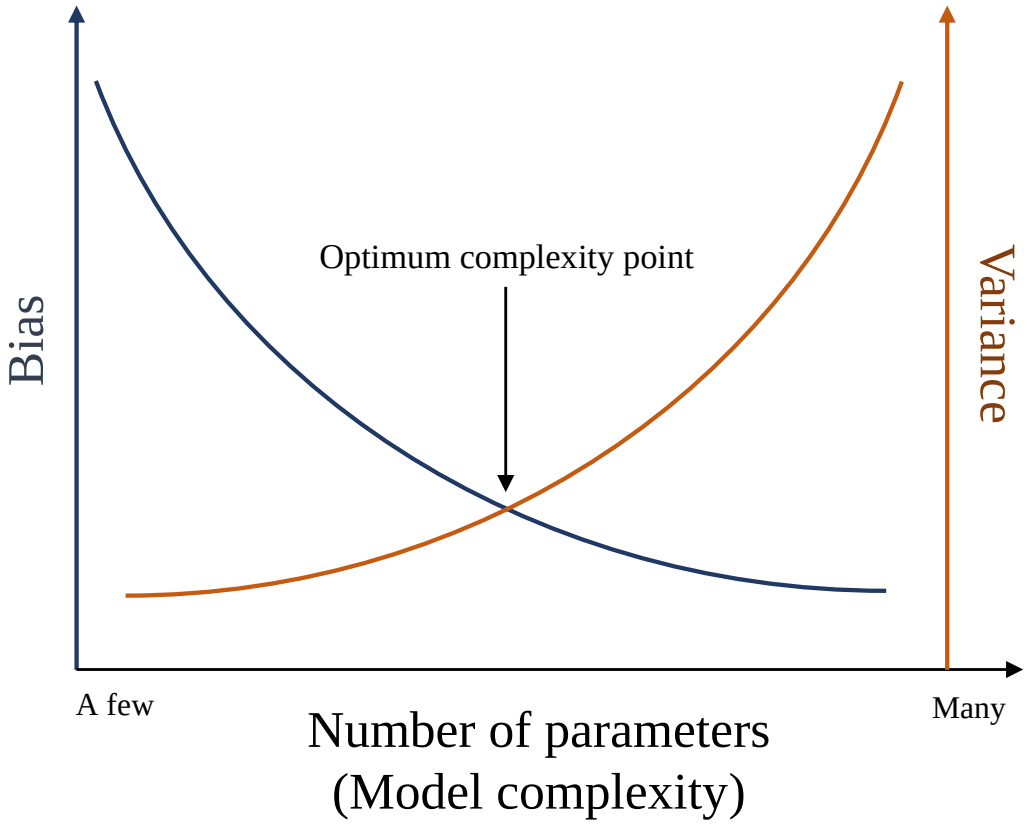

Figure 1.2: Principle of parsimony: A trade off between bias and variance is visualized based on model complexity.

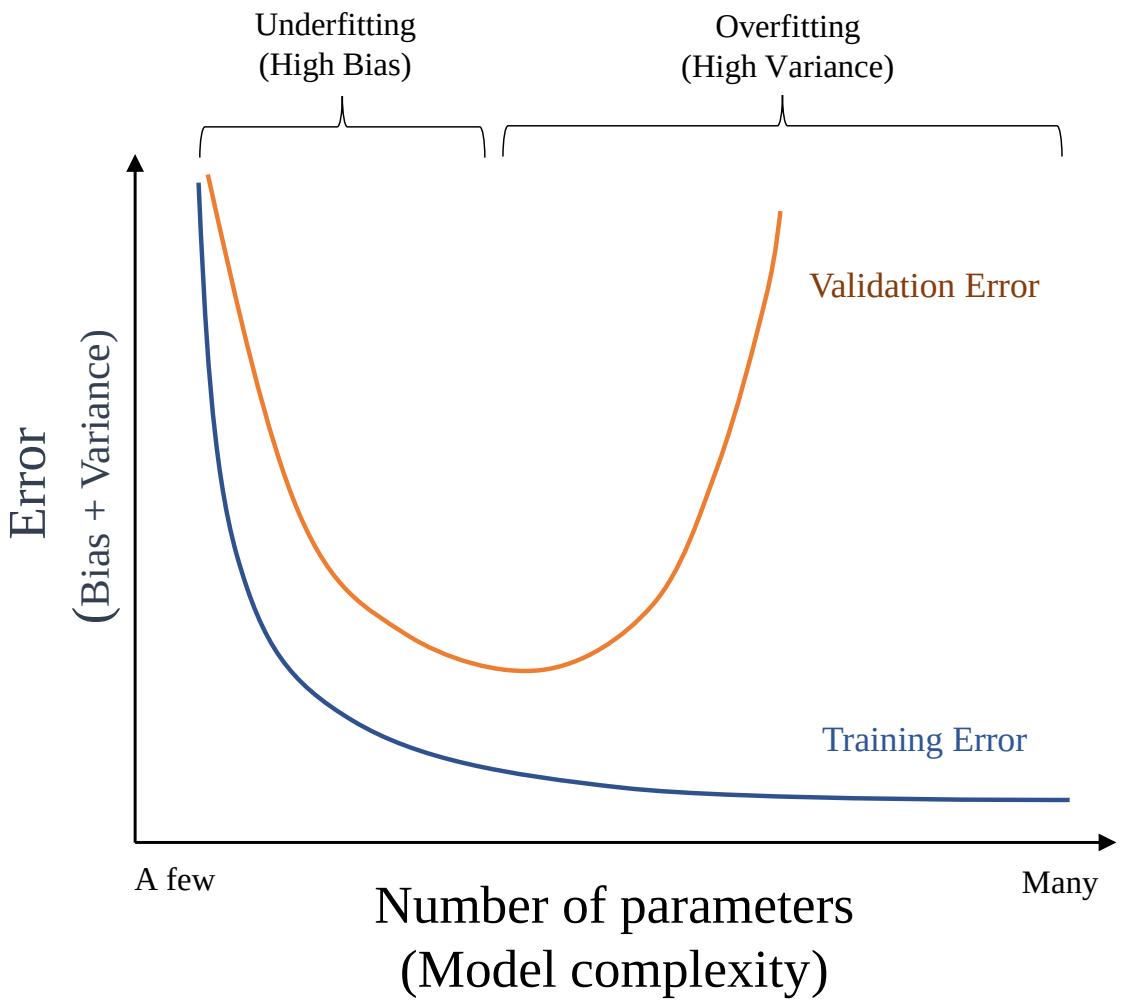

Figure 1.3: Impact of model complexity on training and validation errors

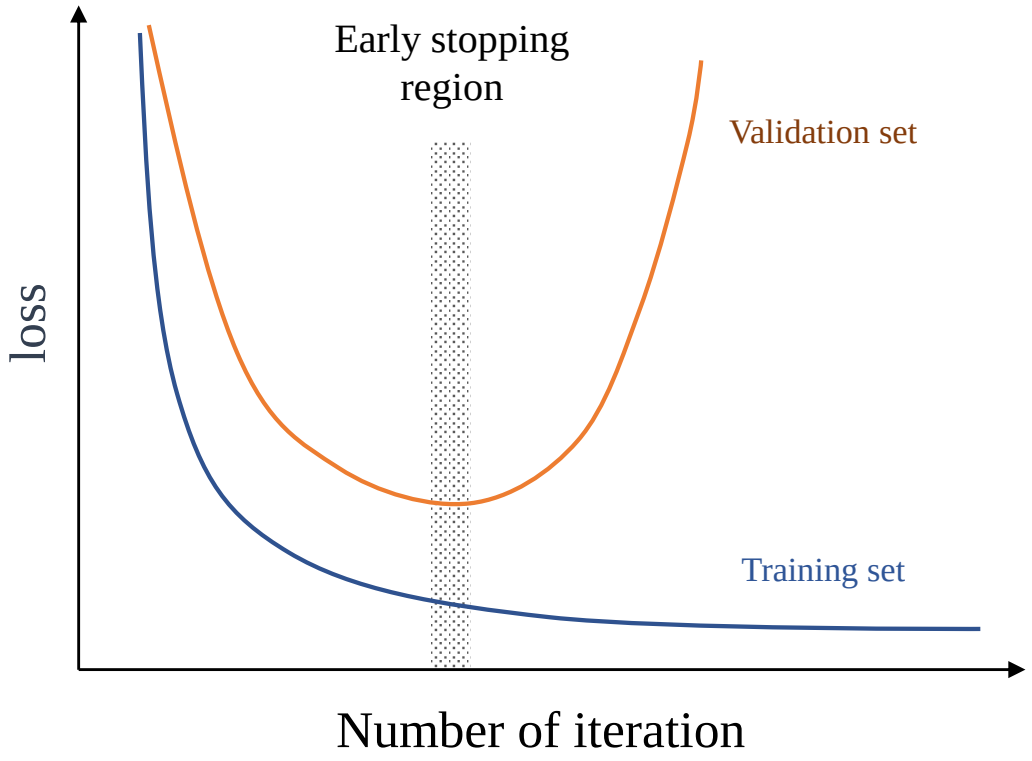

Figure 1.4: Early stopping in deep learning: training can be stopped after a sign of overfitting, an increasing in validation loss values.

the cost and duration of data collection significantly. Moreover, some medical conditions are rare, which needs even more effort to include in the cohort study. In this thesis, we consider Coelho and Togher copora, which are the two largest TBI corpus. In Coelho corpus, there are only 55 TBI subjects [22] whereas Togher corpus contains 58 TBI subjects [23]. However, only 28 subjects, out of 58 subjects, in the Togher corpus completed the speech assessment, which was conducted as a longitudinal study over two years – the absence of subjects in Togher's study due to medical condition and withdrawal that may be rooted from the lack of motivation to continue the study. It is evidenced that collecting medical data has a high cost and onerous process, which limits the availability of TBI data.

According to our preliminary TBI assessment results, the current available amount of TBI speech is insufficient to build such a robust TBI detection model using the Gated Recurrent Unit (GRU) dues to an overfitting problem. The TBI classification model, which is the first work to do so to the best of our knowledge, was developed on Coelho's corpus [22]. The corpus contains speech discourses of story retelling, story generation, and conversation from 55 subjects with TBI and 52 controlled subjects. Although our previously proposed method outperformed the machine learning baseline with a balanced classification accuracy of 62.74%, loss values demonstrate a severe overfitting problem resulted from limited TBI labeled data. According to [19], the GRU model needs at least 453 samples, as an upper bound, to train 789,909 parameters without overfitting; this estimation assumes that the data is well distributed, the length of speech in each recording is five minutes and no recurrent connection between GRUs.

1.6 Limited data mitigation learning methods

In this thesis, we focus on methods addressing the issue of insufficient labeled TBI speech data to improve our speech-based DNN model's performance and reduce overfitting. The overfitting problem can be resolved by either increasing the number of training examples or reducing the number of parameters in the DNN model. Techniques that increase the number of examples include data augmentation [24], and Generative Adversarial Networks (GANs) [25]. Alternatively, this thesis focuses on learning techniques that leverage additional external datasets with

related examples to improve the TBI task's learning performance The three learning techniques are *multitask-learning*, *transfer learning* and *meta-learning*. These limited-data mitigation methods include *source task* training that extracts speech representation from non-TBI task and applied the learned speech features either at the beginning or during the training of *target task*, the TBI assessment. Multi-task learning aims to learn the common low-level features of speech among all source and target (TBI assessment) tasks. Transfer learning and meta-learning extract feature from the source task as prior information to aid the learning in target task.

Challenges: The challenges in applying these learning techniques are incompatibility between source task and target task, determining appropriate source dataset, and locating layers that represent common features.

1.7 Thesis overview

In order to comprehensively compare all learning methods, we proposed an speech-based cascading DNN that comprises GRUs for TBI classification and TBI regression. The cascading DNN performs on sequential features that are extracted from any speech-based backbone model. We evaluated three learning methods and our proposed cascading model using three state-of-the-art DNNs, i.e., VGGish [26], WAV2VEC[27], SincNet [28], as backbone model.

We incorporated transfer learning, multitask-learning, and meta-learning methods into the proposed cascading DNN for TBI assessment. The three learning methods utilize external datasets, so-called *source dataset* in this study, which needs a careful selection to maximize the target task's performance – small dataset and dataset with high specificity might decreases the TBI detection performance [29].

1.8 Thesis research questions

In order to define the scope of our research, we formulate primary research questions as follows.

1. *Best limited-data mitigation method:* Which of these three learning methods is the most suitable method for improving TBI assessment from the speech with limited labeled data? Answering this question will guide us to find a suitable learning method for DNN-based

TBI assessment. We consider both TBI classification and TBI regression to determine the appropriate learning method.

2. *Best type of dataset for each limited data mitigation method:* How can the most suitable source task for TBI assessment be determined? One of the key factors in these learning methods is the data used in the source task training. Conducting experiments using datasets from different tasks as source dataset will help us establishing a rule to select a proper source dataset.

3. *Minimum number of examples required in each learning method:* For each learning method, what are the minimum target examples required to achieve a robust result? Finding the target minimum sample size will help future research to collect an adequate number of samples.

1.9 Novelty of this thesis in relation to prior work

This thesis investigates the learning methods exploiting knowledge from the external domain to cope with insufficient TBI samples. These techniques have been used in other speech-based assessments, but not yet for the TBI assessment. The recent works of depression assessment using speech include transfer learning that adapt the knowledge learned from the task with larger dataset to improve depression detection accuracy [30, 31, 32]. Zhao et al. transferred knowledge learned from speaker recognition model with attention mechanism to predict depression severity from raw speech [30]. Rejaibi et al. proposed a Recurrent Neural Network (RNN) that performs on Mel-Frequency Cepstral Coefficients (MFCC) with transfer learning using emotions recognition as source task to detect depression from speech [32]. Huang et al. investigated adaptation methods for depression detection that have source and target datasets recorded in different environments [31]. Their results demonstrate that knowledge transferred from dataset recorded in naturalistic environments are more difficult to adapt and adding more data from naturalistic environments into the source task training does not help the learning in target task, unlike transferring knowledge from clean speech to noisy speech.

Only a few works employed multi-task learning [33, 34] and meta-learning [35] in speech assessment. Qureshi et al. proposed an attention-based fusion network that fuses features from formant, head pose and eye gaze to simultaneously classify depression severity and predict depression score [33]. The multi-task learning improved their depression classification performance by 5.3% and and 4.9% in regression task. In another work, Chao et al. proposed a multimodel RNN combining audio features and facial features to discriminate depression and emotion simultaneously [34]. By jointly learning features from both tasks, it improved their depression detection performance by 13.7%. For meta-learning, the number of speech assessment that uses this technique is very limited. koluguri et al. recently introduced meta-learning into the speaker representation learning that accounts child-adult speaker classification problem for autism spectrum disorder by considering child and adult as two different tasks [35]. Their speaker representation method outperforms the baseline by 4.82%.

1.10 Thesis contributions

In this study we propose a comprehensive comparison between each learning method that has not yet been studied, especially for the TBI assessment. The contribution of this work is as follows.

1. An audio-based cascading DNN for TBI detection, which aggregates TBI prediction on temporal domain.

2. Comprehensive comparisons of transfer learning, multi-task learning, and meta-learning for TBI classification and regression that initially manifested overfitting as a consequence of insufficient training data.

3. Determining the limited-data mitigation method and source dataset, that provides the best balanced TBI detection accuracy, for the proposed TBI-detection cascading DNN.

4. Few-short learning using limited-data mitigation method, which uses only a few samples of TBI speech to learn TBI detection task.

5. An estimation of TBI training samples required for each limited-data mitigation method using interpolation and extrapolation of TBI detection accuracy.

1.11 Thesis outline

The thesis is presented in subsequent chapters as follows. Chapter 2 summarizes a background of TBI and its relation to speech disorder, followed by three limited-data learning methods and adopted backbone DNNs. In Chapter 3, we proposed a cascading network for TBI assessment that works on top of any backbone DNNs, followed by datasets and evaluation methods. The evaluation results are presented in Chapter 4 with a further discussion in Chapter 5. Finally, Chapter 6 concludes this study and discusses any possible future work.

2. BACKGROUND AND RELATED WORK

2.1 Traumatic brain injury

TBI was defined by CDC as "a disruption in the normal function of the brain that can be caused by a bump, blow, or jolt to the head, or penetrating head injury [36]." In each year, there are at least 1.4 million U.S. people suffer from Traumatic brain injury (TBI) with leading causes of falls, motor vehicle traffic, and assault [1]. Severity of TBI ranges from mild to severe, often classified using the Glasgow coma scale (GCS); a GCS of 3-8 is determined as severe TBI right after the concussion with at least 6 hours loss of consciousness; a GCS of 9-12 is classified as moderate TBI with loss of consciousness between 30 minutes and 6 hours; a GCS of 13-15 is classified as mild TBI (mTBI) with a loss of consciousness less than 30 minutes [37]. Most TBI patients have there physical injury recovered after a period of time, but some medicals condition can result in a long-term problem, such as cognitive and communication impairments, that should be monitored in a long-term to prevent fatality and re-hospitalization.

2.2 TBI-impaired speech: background

Although TBI symptom greatly varies among individual, the common symptoms are reduced processing speed, poor concentration, memory difficulty, and increased fatigue, which are considered as *communication impairments* [11]. After the incident, patients often found themselves lose their abilities to communicate with others precisely. Literature separates communication impairments into *speech impairment* and *language disorders*. In an acute stage of TBI, speech pathologist usually performs a series of speech-language assessment, specifically using assessments for dysarthria, apraxia, and aphasia [4, 5]. Dysarthria and apraxia are motor-speech disorder that results in a poor muscle controlling or motor planning. TBI patients with dysarthria suffer neurological damages to some cranial nerves and fail to control muscles used for pronunciation as they could before the injury [11]. Some TBI patients loss an ability to plan and coordinate articular movements in the brain as a result of apraxia [11]. TBI patients with aphasia, a language disor-

der, may suddenly have trouble understanding, speaking and reading a language as a subsequent following TBI and require language therapy in a long-term [11].

2.3 Prior work on assessing TBI from speech

During the discourse performed by speech-language pathologists, a patient undergoes several assessment and conversations to evaluate TBI's recovery. The conversational discourses include story retelling, story generation, picture descriptions, story narrative, procedural discourse, and casual conversation. Among the speech-language assessment, Western Aphasia Battery (WAB) was indicated to have a high correlation with TBI [4], and is assigned as a predicting target of our proposed regression model. The WAB score ranges from 0 to 100 measuring patient's ability to comprehend spontaneous speech, auditory verbal conprehension, repretation and word finding. Subjects are diagnosed with aphasia if their WAB scores are lower than 93.8.

Previous study indicates correlations between mTBI and speech components, including acoustic characteristics of a vowel, a contrast in phonics, fundamental speech components, and sentence structure [4, 5, 6, 7]. Machine learning, as a classifier, was built using the components in speech as acoustic features to detect mTBI [6, 7]; however, the speeches were produced as a dialogue speech which limits executive functioning in language impairment from the analysis.

As speech disorders are common in moderate to severe TBI, prior work has utilized changes in speech production as indicators of TBI status [4, 5, 6, 7, 8, 9, 10]. Falcone et al. exploited the acoustic characteristics of vowels to classify TBI and found that fundamental speech components such as formants, jitter, shimmer, and harmonic-to-noise ratio are prominent features in the speech of TBI patients [6]. Poellabauer et al.[8] and Daudet et al.[9] proposed a framework to record and process short sentence reading for TBI detection on a mobile tablet using temporal and spectral features, similarly to [6]. Helfer et al. investigated the use of time-delay correlation of speech formants to predict patients' IMPACT score, which is one of the assessments for acute TBI, and concludes that formant correlations is an effective high-level acoustic features in TBI classification [7]. The time-delay correlation matrices has also been utilized with facial expression to estimate cognitive status in mTBI [38] and with gate coordination to detect the mTBI [10].

2.4 Machine learning analyses to assess TBI from speech

Most of the TBI classification mentioned earlier, except [10], used Support Vector Machine (SVM) [12] as a classifier to detect TBI. These previous studies prioritized developing a feature set that improves TBI assessment performance from acoustic feature and did not consider DNNS that may due to an insufficient amount of training samples. Only [10] employed Convolution Neural Networks (CNNs) with the results outperforming Gaussian Mixture Models (GMMs). The acoustic features and TBI samples used in each study are summarized in Table 2.1.

Table 2.1: Previous studies in speech-based TBI detection

	Task	**Speech features**	**TBI subject**	**Result**
[6]	Digit reading	Fundamental components	7% (n=105)	F1: 87.51
[8, 9]	Sentence reading and pa-ta-ka	Temporal and spectral	16% (n=581)	AUC: 0.86
[7]	Passage reading	Formant correlation	29% (n=32)	AUC: 0.95
[10]	Passage reading	Formant correlation	55% (n=21)	AUC:0.90
[10]	Free Speech	Formant correlation	55% (n=21)	AUC:0.89
[10]	Pa-ta-ka	Formant correlation	55% (n=21)	AUC:0.89
[10]	All three combined	Formant correlation	55% (n=21)	AUC:0.96

2.5 Learning techniques in DNN

Although DNN has been developed and applied in many applications since 1943, it has just reached its prime time in 2010s [39]. The major challenges of DNN in an early stage are the lacking of computational power – until recent development of Graphics Processing Unit (GPU) has become available – and *data scarcity* that remains a challenging research topic and is addressed in this study. Specifically, we investigate deep learning learning methods that exploits knowledge from other domains for TBI detection, which has limited labeled data.

2.5.1 Single task learning (Baseline)

Single task learning is learning to perform task t without exploiting any external dataset or other tasks. We consider this method as one of the baselines in this study. The algorithm aims

to learn parameters θ_t that is initialized randomly by minimizing the loss function L_t, as in Equation 2.1 with a terminology explained in Table 2.2.

$$\min_{\theta} L_t(f_{\theta_t}) \tag{2.1}$$

Symbol	Definition
$T_{1,2,3,...,k}$	Learning task 1 to task k
θ_t	Model parameter of task T
L_t	Loss of task T.
f_θ	Model f with parameters θ

Table 2.2: Equation notation

This learning method is known as *training from scratch* and is susceptible to the overfitting problem. Without any data augmentation, it is possible to mitigate the overfitting problem by reducing the number of trainable parameters in the model and introducing a regularizer into a loss function. In the baseline, we include a dropout layer [40], a regularizer, with a tuned dropout rate to reduce overfitting. Dropout is included in many state-of-the-art DNN models, even in cases where there is adequate training data to prevent overfitting to training batch [41, 28]. Moreover, the number of parameters in each DNN is gradually reduced to obtain the best balance accuracy that equally reflects *sensitivity* (true positive rate) and *specificity* (true negative rate).

2.5.2 Transfer learning

To mitigate the overfitting problem in TBI detection, we introduce transfer learning into TBI detection model using knowledge learned from other audio tasks. Transfer learning, or *domain adaptation*, exploits knowledge learned from *source tasks* (T_s) with a goal of improving learning

on the*target task* (T_t). The method transfers low to mid-level feature representations by initializing the target task's parameters with weights from the pre-trained DNN, which reduces the total number of optimized parameters and mitigates the overfitting problem. Low to mid-level feature representations $\theta_{shallow}$ are obtain following Equation 2.2, and used to initialize weights in Equation 2.3.

$$\min_{\theta_s,\theta_{shallow}} \sum_{i=0}^{k} L_i(f_{\theta_i}(f_{\theta_{shallow}})) \tag{2.2}$$

$$\min_{\theta_t} L_t(f_{\theta_t}(f_{\theta_{shallow}})) \tag{2.3}$$

$$\min_{\theta_t,\theta_{shallow}} L_t(f_{\theta_t}(f_{\theta_{shallow}})) \tag{2.4}$$

Equation 2.3 is a domain adaptation process that only optimizes task-specific parameters (θ_t) and leaves $\theta_{shallow}$ unchanged to train task-specific model f_{θ_t}. After the learning of f_{θ_t} complete, all parameters ($\theta_{shallow}$ and θ_t) are trained fine-tuned as in Equation 2.4. Overfitting is rarely observed in the domain adaptation step but expected in the fine-tuning step as the number of trainable parameters is the same as in the single-task learning. However, single-task learning does not utilize pretraining weights, and overfitting occurs at a higher loss.

Fundamentally, in order to avoid the overfitting problem, the source task learning requires a large corpus in order to generate a decent pre-trained DNN. Transfer learning has previously been employed to improve the performance of various health assessment methods in order to mitigate the overfitting problem, including Post-Traumatic Stress Disorder (PTSD) and depression assessments from speech [42, 43, 44].

2.5.3 Multi-task learning

The majority of model's parameters reside in the early layers ($f_{shallow}$) of TBI detection model. Optimizing parameter $\theta_{shallow}$ across multiple tasks potentially prevents $f_{shallow}$ from learning bias

in TBI data which causes overfitting problem. The first few layers ($f_{shallow}$) of DNN extract generic information of the input, and the deeper into the model, the more complex patterns are extracted. $f_{shallow}$ usually contains more than half of the total parameters. The transformation between hidden layers reduces the number of features to learn more complex and important features. The concept of multi-task learning is to learn these first few layers together among task $T_{1,2,3,...,k}$, which prevents parameters $\theta_{shallow}$ in $f_{shallow}$ from overfitting. Parameters $\theta_{shallow}$ are shared among all tasks and are jointly optimized to simultaneously learn common data representations across all k tasks. The final prediction is made independently by a in-task network f_θ that follows $f_{shallow}$. The training can be explained as Equation 2.5.

$$\min_{\theta} \sum_{i=0}^{k} L_i(f_{\theta_i}(f_{\theta_{shallow}})) \tag{2.5}$$

For n tasks with an equal amount of data, the ratio of model parameters to the data sample is reduced significantly by 2^n. Thus, multi-task learning reduces the overfitting problem when n is large enough. However, previous work in healthcare application only considers tasks available within the dataset [45, 46], which may be insufficient to train the low-level feature; [45] considers gender classification as a secondary task to the the emotion recognition task while [46] includes relevant emotion classification as an additional task to aid the emotion classification. In this study, we utilized data from other datasets with considerable amounts of data to improve TBI detection performance.

Multi-task learning also has an additional advantage of increasing the network's ability to discriminate classes that it previously confuses for one another. Task t_1 is usually a target task with a small number of examples, while t_2 is a task with a large corpus in the closed domain. The training set is comprised of an equal number of examples from both tasks. However, the target task contributes more to the model parameter update. The multi-task learning method is also known as joint learning, where the first few layers of the network are shared to learn low-level feature representation, as in [45, 46].

2.5.4 Meta-learning

Meta-learning aims to learn well on the target task, TBI detection in this study, with minimal training data. It is usually defined as *learning to learn fast*, which is a broad term describing three learning types of meta-parameters (Φ). The three meta-learning methods are metric-based, model-based, and optimization-based learning. This study considers only optimization-based meta-learning since our objective is not to optimize the model architecture or to compare the distance function. Specifically, we adopted Model-Agnostic Meta-Learning (MAML) algorithm [47] in this study. MAML is a model-agnostic method that makes it compatible with any DNN with gradient descent.

The method contains two steps similar to transfer learning. The first step is a prior training step utilizing other related tasks. However, it differs from transfer learning in the sense that its objective is not to optimize for the task in pretraining but any other tasks, including the target task.

Meta-learning learns to generalize core knowledge from previous tasks while learning domain-specific bias on a new task. Unlike the transfer learning that freezes the weights transferred from the source network, meta-learning adapts parameters in the target model and bias according to the metadata. The metadata guides the learning to avoid the overfitting problem. Previous work successfully incorporated lifelong learning in target tasks that had very few audio samples [48, 49].

2.6 DNNs and audio-based backbone network

Although all three learning methods presented in the previous section are model-agnostic, the performance of each method might depend on the network architecture. This study adopted three audio-based DNNs as a backbone model to learn audio representation from TBI speech. The three DNNs are VGGish[26], Wav2Vec[27] and SincNet[26]. This section explains various types of DNNs used in this thesis followed by the three backbone networks.

2.6.1 Convolution Neural Network (CNN)

CNN is a neural network that applies convolution on topology input instead of matrix multiplication [50]. CNN's kernel dimension depends on the input type; one-dimensional CNN is

generally applied on time series data while two-dimensional CNN is applied on an image. For two-dimension CNN kernel (K), the feature map (S) with (i, j) coordination is defined as Equation 2.6 where I is an input and (m,n) is a filter size.

$$S(i,j) = \sum_m \sum_n I(m,n)K(i-m, j-n) \tag{2.6}$$

Because of CNN's property that maintains sparse interactions, parameter sharing, and equivariant representations, most state-of-the-art DNNs include CNN in the very first layers to extract invariant features with minimum parameters [39]. CNN is usually employed in a non-linear activation layer and a pooling layer, visualized in Figure 2.1. Non-linear activation, such as Rectified Linear Unit (ReLU) [51], acts as a feature detector while pooling layer, max-pooling [52] for example, helps CNN to be invariant to small translations of I over K.

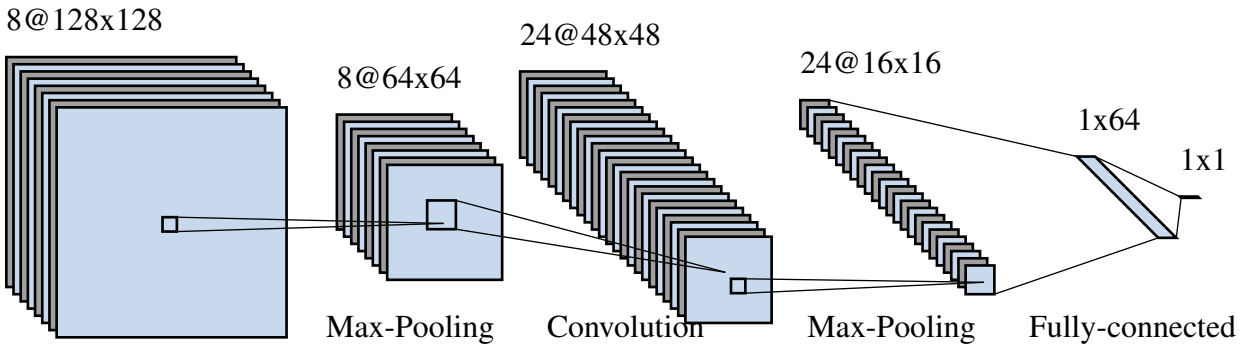

Figure 2.1: An example of CNN architecture comprises two convolution layers and two max-pooling layers

2.6.2 Gated Recurrent Unit (GRU)

GRU is a variation of Recurrent Neural Network (RNN) that specializes in sequence modeling [53]. GRU improves memory utility in the Long Short-Term Memory (LSTM) by combining forget gate and state unit into a single gating unit h_i^t (for time step t and cell i) as in Figure 2.2 and Equation 2.7 where the update gate u_i^t and reset gate r_i^t are defined in Equation 2.8 and Equation 2.9

respectively. The term W, U and b are recurrent weight, input weight and bias respectively. All gates contain a sigmoid function (σ) in the temporal loop, which weights decision to update, forget and ignore.

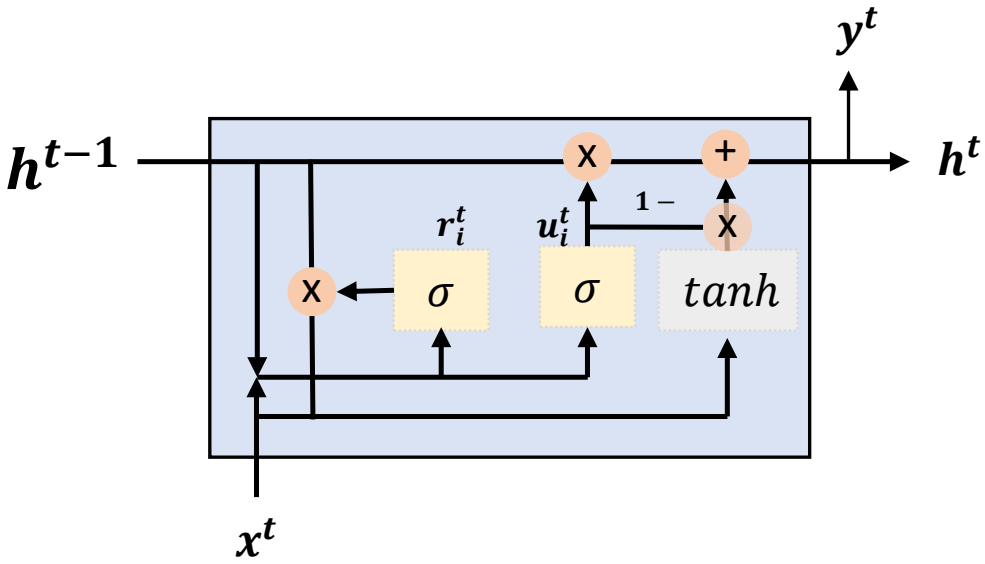

Figure 2.2: GRU cell(s): GRU contains connection between each input time step, called gating unit (h_i^t)

$$y_i^t = h_i^t = u_i^{t-1} h_i^{t-1} + (1 - u_i^{t-1}) tanh\big(b_i + \sum_j U_{i,j} x_j^t + W_{i,j} r_j^{t-1} h_j^{t-1}\big) \tag{2.7}$$

$$u_i^t = \sigma\big(b_i^u + \sum_j U_{i,j}^u x_j^t + W_{i,j}^u h_j^t\big) \tag{2.8}$$

$$r_i^t = \sigma\big(b_i^r + \sum_j U_{i,j}^r x_j^t + W_{i,j}^r h_j^t\big) \tag{2.9}$$

2.6.3 VGGish

Hershey et al. [26] demonstrated that the performance of Convolution Neural Network (CNN) in audio-based scene classification is as well as in image recognition. A VGGish model was

proposed by modifying the layers of original VGG [14] in the following way. The input size was changed to 96x64 for the log-Mel spectrogram, and only four groups of convolution/max-pooling layers were utilized instead of five. The full network architecture is shown in Figure 2.3 with a configuration listed in Table 2.3. The VGGish model was developed and fine-tuned using AudioSet corpus [54], an audio dataset with 632 classes collected from the soundtracks of YouTube videos.

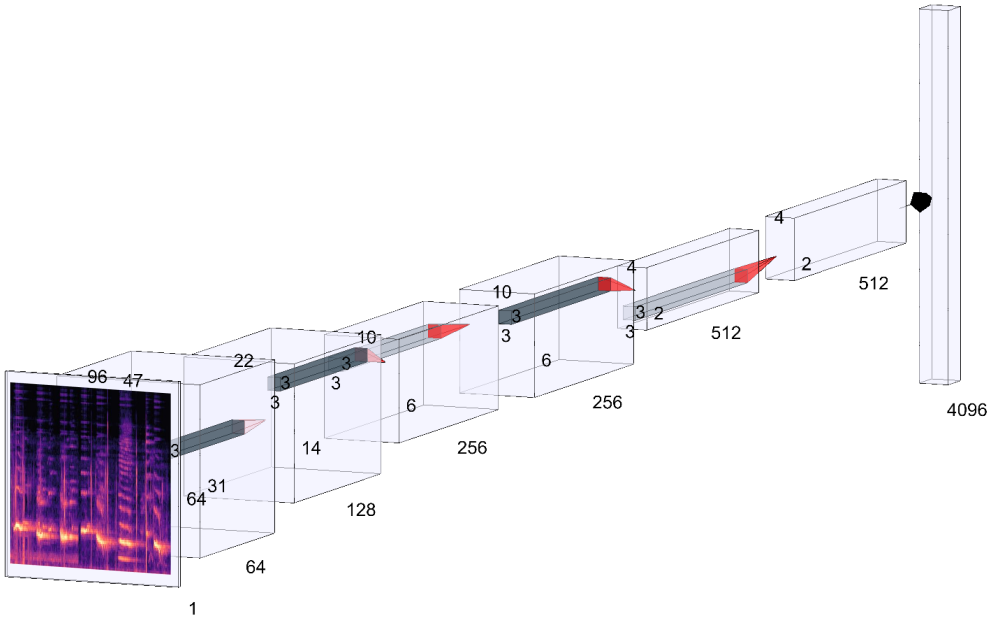

Figure 2.3: Architecture of VGGish: a backbone model for audio feature extraction

The Mel spectrogram represents an audio on frequency domain and can be generated as follows:

1. Signal pre-processing and framing: The signal processing method varies across the applications. However, it is common to apply a pre-emphasis filter to boost high frequency and apply a bandpass filter to remove some noise. The pre-emphasized signal (s') at time t is computed using the Equation 2.10, where s(t) is a raw audio signal at time t and α is a coefficient, usually, between 0.90 and 0.97. In speech processing, it is common to limit the maximum frequency to about the Nyquist frequency. In practice, [55] suggested to extract

Table 2.3: VGGish model architecture

Model	Number of filters	Filter size	Stride	Activation function
CNN 2D	64	3×3	1×1	ReLU
Max pooling 2D	-	2×2	2×2	-
CNN 2D	128	3×3	1×1	ReLU
Max pooling 2D	-	2×2	2×2	-
CNN 2D	256	3×3	1×1	ReLU
CNN 2D	256	3×3	1×1	ReLU
Max pooling 2D	-	2×2	2×2	-
CNN 2D	512	3×3	1×1	ReLU
CNN 2D	512	3×3	1×1	ReLU
Max pooling 2D	-	2×2	2×2	-

MFCC features between 40 Hz and 7800 Hz for audio signals with a sampling rate of 16000 Hz.

$$s'(t) = s(t) - \alpha s(t-1) \tag{2.10}$$

Then, the audio is split into a small frame of 20-40 ms with 10 ms. Subsequent extraction steps are performed on a short duration of frame audio signal.

2. Windowing and Discrete Fourier Transform (DFT): To smooth and prevent discontinuities in the signal, particularly, at the beginning and end of each frame, hamming window w_{ham}, as shown in Equation 2.12 is applied on the time domain signal ($s_i(n)$) during DFT, Equation 2.11, where N denotes frame size. The DFT transforms the signal at frame i (s_i) to a complex DFT (S_i) of DCT window size K. Then, power spectrum is obtained by Equation 2.13.

$$S_i(k) = \sum_{n=1}^{N} s_i(n) w_{\text{ham}}(n) e^{-j\frac{2\pi}{N}nk}, 1 \leq k \leq K \tag{2.11}$$

$$w_{\text{ham}}(n) = 0.54 - 0.46cos(\frac{2\pi n}{N-1}), 0 \leq n \leq N-1. \tag{2.12}$$

$$P_i(k) = \frac{|S_i(k)|^2}{N} \tag{2.13}$$

3. Mel Filterbank: Mel-filter banks mimic humans' sound perception that is non-linear and more discriminative at a lower frequency than higher frequency. Thus, magnitude coefficients are passed through triangular filter, accumulating weighted magnitude into filterbank spectral magnitude. M filters bank (H_m) can be computed as in the Equation 2.14-2.16. The final Mel-spectrogram is visualized in Figure 2.4.

$$m(f) = 2595\log_{10}(1 + \frac{f}{700}) \tag{2.14}$$

$$f(m) = 700(10^{m/2595} - 1) \tag{2.15}$$

$$H_m(k) = \begin{cases} 0 & k < f(m-1) \\ \frac{k - f(m-1)}{f(m) - f(m-1)} & f(m-1) \leq k \leq f(m) \\ \frac{f(m+1) - k}{f(m+1) - f(m)} & f(m) \leq k \leq f(m+1) \\ 0 & k > f(m+1) \end{cases} \tag{2.16}$$

2.6.4 Wav2Vec

Wav2Vec is an unsupervised model, originally built as part of the speech recognition model, to extract acoustic features from the waveform [27]. The pre-trained model consists of an encoder network ($\mathcal{Z}$) containing five CNN layers and a contextual network ($\mathcal{C}$) that combines the representations from the encoder network into a vector representing 210 ms of audio. The model was pretrained to minimize the contrastive binary loss ($\mathcal{L}_k$) of phoneme classification at time k, computed between current and next contextual representation as in Equation 2.17 where λ is a regularization rate of $\mathbb{E}$, computed over ten negatives examples (n) that are uniformly choosing from an audio sequence of length T with a probability of p_n.

$$\mathcal{L}_k = -\sum_{i=1}^{T-k} \Big(\log \sigma(\mathbf{z}_{i+k}^{\top} h_k(\mathbf{c}_i)) + \lambda \mathbb{E}_{\tilde{\mathbf{z}} \sim p_n} \left[\log \sigma(-\tilde{\mathbf{z}}^{\top} h_k(\mathbf{c}_i)) \right] \Big) \tag{2.17}$$

$$h_k(\mathbf{c_i}) = W_k \mathbf{c_i} + \mathbf{b_k} \tag{2.18}$$

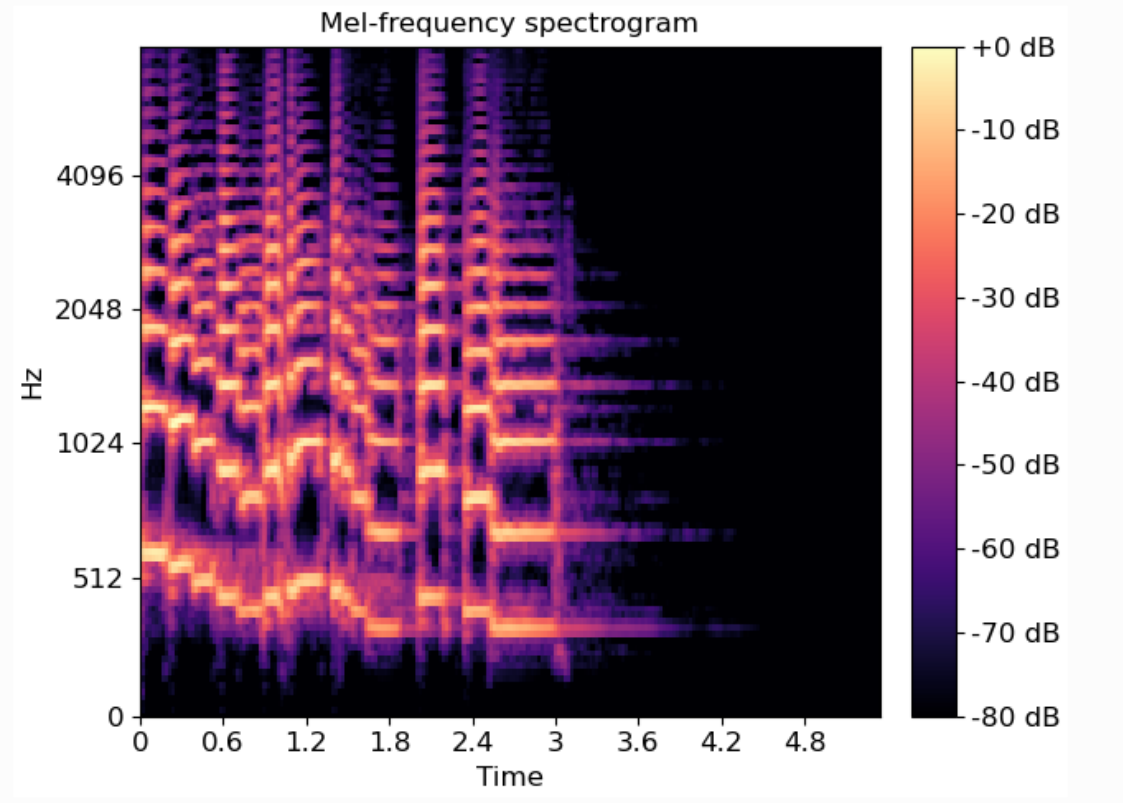

Figure 2.4: Mel-spectrogram: an input for VGGish backbone model

The trained representation was used to boost the speech recognition system's performance in place of log-Mel filterbank features [27]. Only the layers in encoder networks, listed in Table 2.4, are used as a backbone model of TBI detection model.

Table 2.4: Wav2Vec model architecture (encoder network)

Model	Number of filters	Filter size	Stride	Activation function
CNN 1D	512	10	5	ReLU
CNN 1D	512	8	4	ReLU
CNN 1D	512	4	2	ReLU
CNN 1D	512	4	2	ReLU
CNN 1D	512	4	2	ReLU

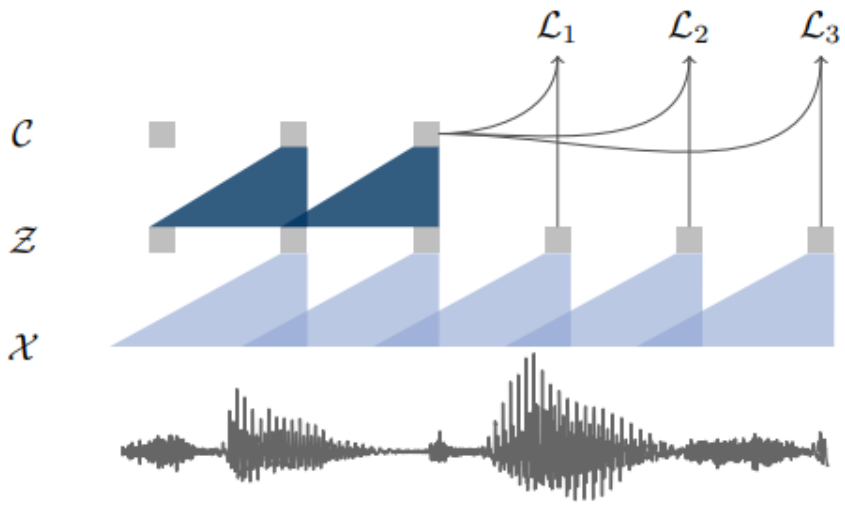

Figure 2.5: Architecture ofWAV2VEC model: a pre-trained model, originally proposed for an automatic speech recognition in [27]

2.6.5 SincNet

The last backbone network of our proposed TBI detection model is SincNet [28]. Although [27, 26] demonstrated that CNN provides excellent performance on audio data, the models do not process on the raw audios. CNN models require a data transformation step, such as log-Mel spectrograms, before the training, which is not a complete end-to-end DNN model. To fully utilize DNN as end-to-end learning, SincNet was proposed by incorporating trainable Sinc filters (Equation 2.19) in the first layer as vizualized in Figure 2.6 [28]. The Sinc filter acts as a bandpass filter allowing only signals within a specific frequency range to pass through to the subsequent CNN layers. [28] shows that SincNet outperforms CNN in the speaker recognition task. The network architecture is shown in Table 2.5 below.

$$\text{sinc}(x) = \frac{\sin(x)}{x} \tag{2.19}$$

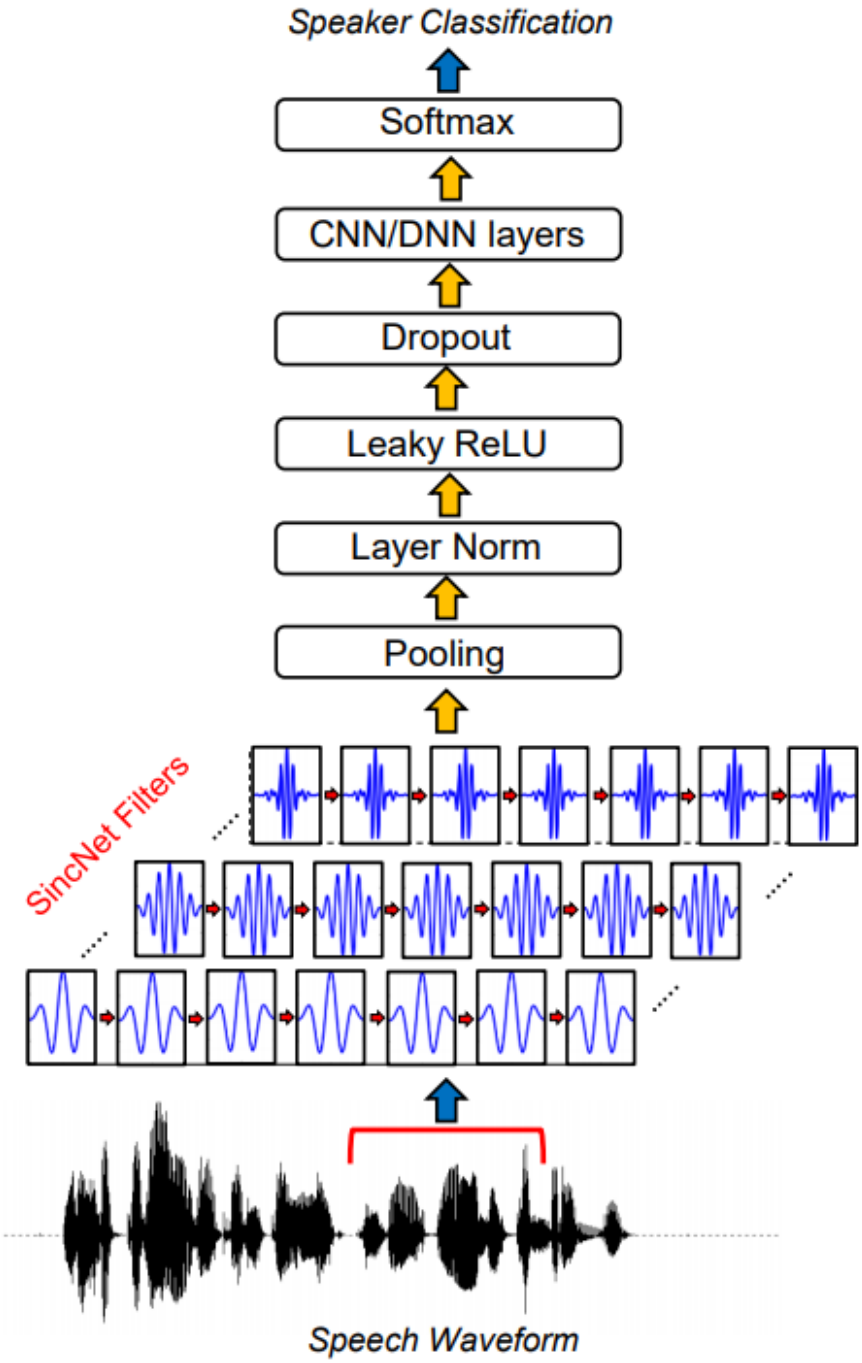

Figure 2.6: SincNet architecture [28]

Table 2.5: SincNet model architecture

Model	Number of filters	Filter size	Stride	Activation function
Sinc	80	251	-	ReLU
Max pooling 1D	-	-	3	-
CNN 1D	60	5	-	ReLU
Max pooling 1D	-	-	3	-
CNN 1D	60	5	-	ReLU
Max pooling 1D	-	-	3	-
CNN 1D	60	5	-	ReLU
Max pooling 1D	-	-	3	-

3. METHODOLOGY

In this chapter, we introduce a cascading DNN that utilized a backbone network to extract acoustic features followed by learning methods for limited data. Our methodology is conceptually visualized in Figure 3.1. The block in the left part of diagram shows limited-data mitigation method for DNN that exploits external data to train the TBI detection model. Multi-task learning is jointly trained with the proposed TBI assessment model, shown in the right block, while transfer learning and meta-learning learned pre-trained weights or meta-weights prior to the training of TBI assessment model. All three learning methods for limited data leverages the single task learning (baseline) by utilizing external data from other audio tasks (denoted as source task). The datasets and preprocessing steps are subsequently presented in this chapter.

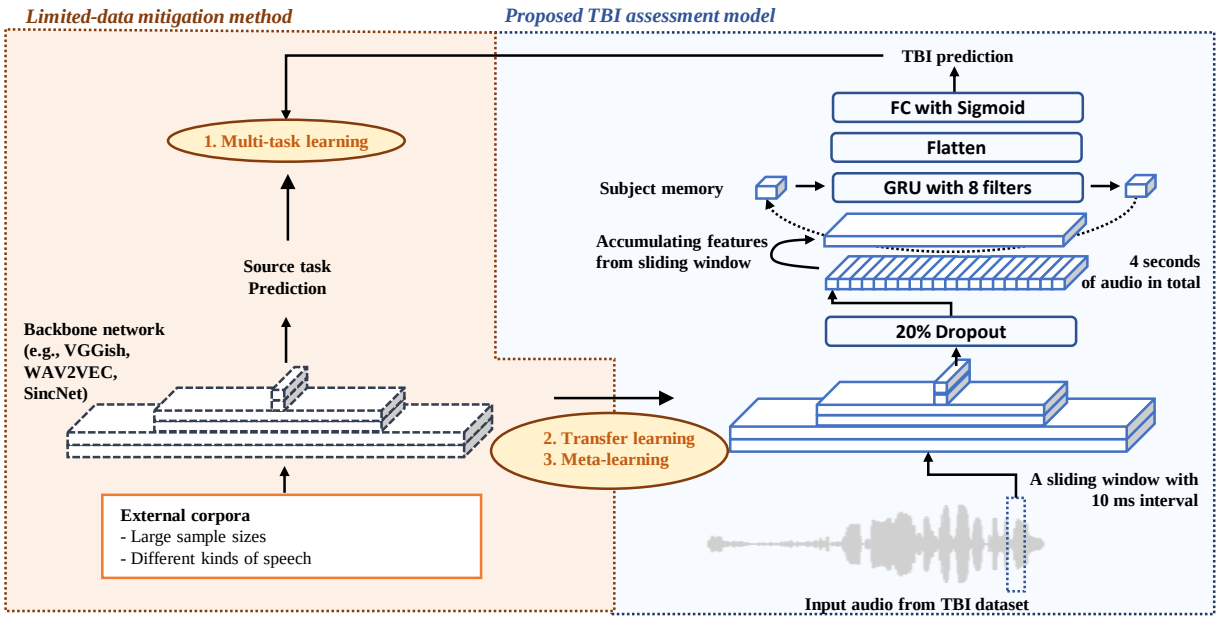

Figure 3.1: Proposed learning methods for TBI detection with limited-data: An overview of limited-data mitigation methods, i.e., multi-task learning, transfer learning, and meta-learning applied on the backbone DNNs for source task prediction and the proposed TBI assessment model, which includes layers from backbone DNNs.

3.1 Cascading DNN for TBI classification and regression

This study utilized three backbone models, previously introduced in section 2.6. Each backbone model processes on different input dimensions and yields different bottleneck-feature size, as shown in Table 3.1. The length of the input was initially designed depending on the task. In TBI detection, we previously found that the minimum speech duration for TBI assessment is 4 seconds. The audio length for TBI assessment and other hyperparameters were tuned using values reported in Table 3.2.Hence, we proposed a cascading network that performs on bottleneck-feature instead of modifying the backbone networks. The TBI-cascading aims to predict TBI over a sequence of bottleneck features provided by each backbone model. To formulate the input for each backbone model, a sliding window of the length specified in Table 3.1 with 10 ms interval was applied on the raw audio to format the input shape to match the backbone network. As the proposed method operates over a sequence of acoustic feature, the concatenation of outputs from the backbone model are considered as a combined representation to the TBI cascading model, which comprises one GRU layer. Since overfitting is the primary concern of this study, we employed GRU with a small filter size of 8. The final TBI prediction is made by a Fully Connected (FC) layer with Sigmoid activation for binary classification and linear activation for regression. Figure 3.2 visualizes a binary TBI prediction using the Sigmoid activation function and classifies TBI status based on four seconds of the audio signal.

Table 3.1: Input and output dimensions of the backbone models

Model	Original task	Input audio	Bottleneck feature size
VGG	Scene classification	25 ms	128
WAV2vec	Speech embedding	210 ms	512
SincNet	Speaker recognition	200 ms	2048

Hyperparameter	Values
Number of GRU filter	4,8,12,16
Number of GRU layers	1,2
Dropout rate	0,0.2,0.4,0.6,0.8
Learning rate	0.003,0.002,0.001,0.0005
Audio length (seconds)	1,2,4,6,8

Table 3.2: Hyperparameters in training the cascading DNN

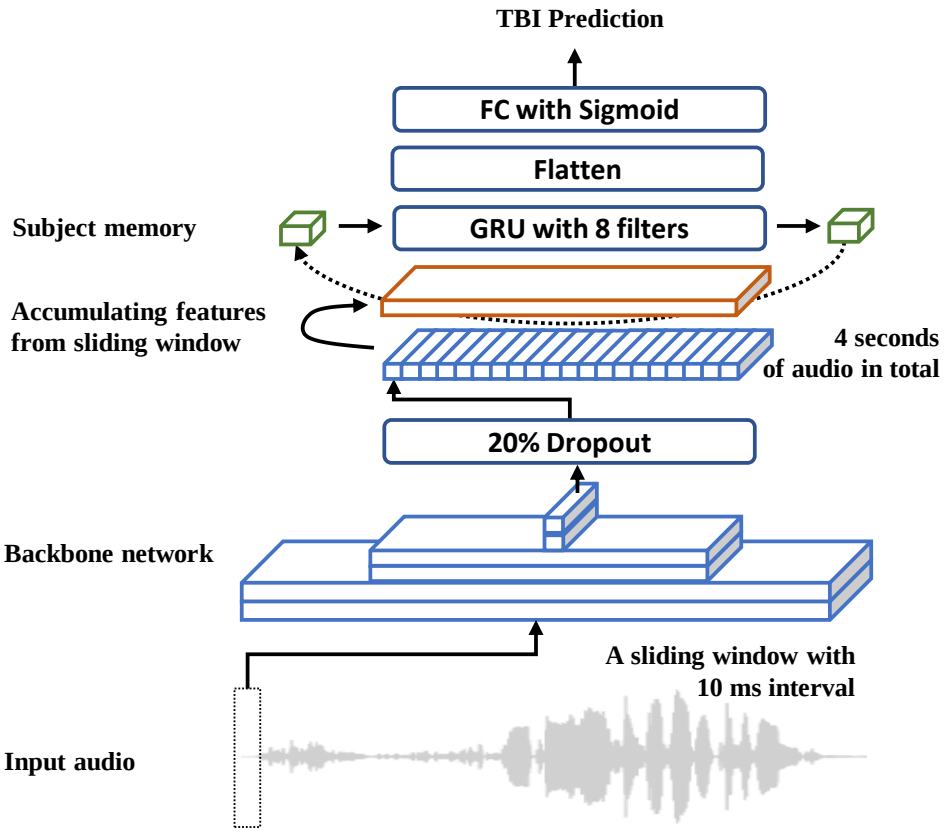

Figure 3.2: Architecture of our proposed cascading DNN for TBI classification and regression

3.2 Audio datasets utilized

This study investigates DNN learning methods that exploits *source dataset* to improve performance of the *target task*, TBI detection. Each learning method aims to transfer or share the knowledge from other tasks into a TBI detection, which consequently reduces the overfitting. The

major criteria we used to select a source dataset are based on the corpus size and domain distance. The source dataset should contains more samples than the task dataset and does not manifest an overfitting problem in the training of source task. In the transfer learning of image classification, [29] shows that the source task needs at least 100 samples per class to cope the overfitting and at least 1000-1300 samples per class until the overfitting is hardly observed. Another important factor in domain adaptation is the specificity of source tasks and distance between target and source tasks. [29] demonstrates that the adaptation of pre-trained network to the target task may fail due to representation specificity, especially, in the deeper layers. Source datasets that represent audio in generality are robust to most target tasks, but might be inferior to the high-specificity source task.

3.2.1 Source dataset

Three source datasets previously utilized for different learning tasks were explored and compared on our three limited data mitigation learning methods. We describe these datasets below, ordering them by the distance of their prior domains to our primary TBI corpus (from far to close).

1. **Google Audio Set** [54] is a large corpus containing 632 audio events for scene classification. It contains more than 2 million human-labeled of 10-second sound clips extracted from the videos on Youtube. While the corpus has a tremendous amount of audio examples, it contains a relatively small number of human speech examples.

2. **Librispeech corpus** [56] is a large corpus containing 1000 hours of audiobooks and has been frequently used to train speakers recognition and automatic speech recognition. Although all the audio in this dataset involved speech recorded from male and female speakers, the speech is scripted speech and does not capture all-natural speech variations.

3. **Wizard-of-Oz (WOZ)** interviews corpus [57] is part of the Distress Analysis Interview Corpus (DAIC). The corpus contains 50 hours of recording audio during 189 clinical interviews by a virtual interviewer, controlled by a human to assess depression. A Personal Health Questionnaire (PHQ-8) depression scale questionnaire was gathered and used as a target.

The speech collected is considered spontaneous, which is close to our TBI corpora and our study goals.

3.2.2 Target dataset

Two TBI datasets in the TBIbank discourse were selected based on sample size and speech quality. All TBI data from two datasets are spontaneous speech.

1. **Coelho corpus** [22] is a collection of discourses from individual with TBI. The corpus contains equal numbers of positive and negative cases, which prevents the imbalance class issue. The discourse includes story retelling, story generation, and conversation collected from 55 native English speakers with closed head injuries and 52 native English speakers with no brain injury. The subjects were labeled as TBI subjects if they had closed head injuries in the past. All 55 subjects with TBI meet the criteria of being recovered a high level of functional language – they had achieved fluent conversation. They did not demonstrate appreciable deficits on traditional clinical language tests. The cause of brain injury in this corpus are motor vehicle accident (44 subjects), fall (6 subjects), stuck by car (3 subjects) and others (2 subjects).

2. **Togher corpus** [23] collected speech during the TBI assessments of 55 subjects. All subjects in this study were diagnosed with TBI and were monitored for recovery at every three months, six months, nine months, twelve months, and two years post-injury. However, more than half of the participants did not attend twelve months and two years post-injury sessions. Therefore, we only considered data collected from three months, six months, and nine months. During the discourse, ten assessments were performed. However, we use only Western Aphasia Battery-revised (WAB) to train our regression models.

The size of each corpus is conceptually plotted against domain distance between the source dataset and TBI dataset in Figure 3.3. Google audio set and Librispeech corpus are the two largest corpora with public access. The Google audio set contains all kinds of audio, and Librispeech contains speech recorded in a set-up environment. The corpus size is beneficial to the learning

methods that exploit external data regardless, but the difference between the application domain is one of the questions in this study. WOZ-DAIC is a small dataset, but it contains spontaneous speech recorded for health assessment, which is similar to TBI datasets' nature.

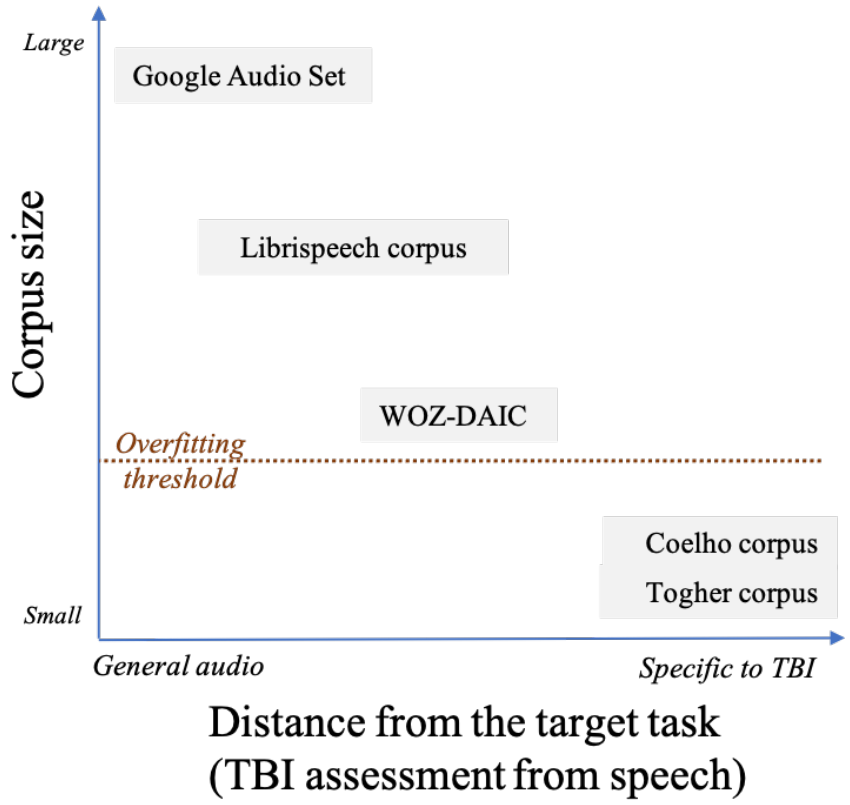

Figure 3.3: Characterization of our source datasets based on their size and distance from target task (TBI assessment). The overfitting threshold is estimated based on corpus size and model parameters.

3.2.3 Data preprocessing and feature extraction

The conversational speech produced by a target was extracted using onset time in the transcript provided by Coelho corpus and manually selected in the Togher corpus. Three TBI subjects and one control subject in the Coelho corpus were excluded from the experiments due to incomplete transcripts. Five subjects from the Togher corpus were excluded due to the low signal-to-noise

Name	Task	Recording type	Sample size (hours)
Source datasets			
Google Audio Set [54]	Scene classification	Youtube video	5790
Librispeech corpus [56]	Speaker recognition	Audio books	1000
WOZ interviews corpus [57]	Depression assessment	Interview	34
Target datasets			
Coelho corpus[22]	TBI assessment	Discourse	19
Togher corpus[23]	TBI assessment	Discourse	16

Table 3.3: Summary of source and target datasets used to train the TBI detection model

ratio. All corpus were normalized using Vocal Tract Length Normalization (VTLN) followed by a min-max normalization.

VTLN: VTLN was used to suppress the inter-speaker variation from vocal tract length between different genders and ages. The piece-wise frequency warping was performed on the frequency domain to optimize factor α between a range of 0.1 and 1.8 with 0.1 incremental. α is optimized using Equation 3.1 to increase phoneme classification performance (W) using Gaussian Mixture Model (GMM) and Hidden Markov Model (HMM), denoted as λ, on cepstrum features (X) of each speaker i [58, 59]. The warping factor of each speaker was determined using all speaking utterances.

$$\alpha_i = \arg\max_{\alpha} Pr(X_i^{\alpha} \mid \lambda, W_i) \tag{3.1}$$

The normalized audio was used as an input in SincNet and Wav2Vec models, but we applied Mel-spectrogram transformation on the normalized audio as an input in VGGish model.

3.3 Evaluation method

The three limited data mitigation learning methods that we explored require different training processes but were implemented and trained with PyTorch library [60] using NVIDIA tesla K80, P100 and V100 GPUs. The evaluation was performed using 10-fold cross-validation that was done in a subject-wise fashion to avoid any bias with multi-label stratification method [61] to maintain

demographic distributions (gender, age and TBI severity) in each fold. In each cross-validation run, two subjects from each TBI class were randomly selected as a validation set.

3.3.0.1 Transfer learning

The network was pre-trained on their original task, and the learned weights were utilized as initial weights for training the TBI assessment model. As described in section 2.5, only parameters in the TBI classifier were optimized in the domain adaptation step, freezing the pre-trained weights. After no loss improvement was observed for ten epochs, the fine-tuning process bagan with a scheduled learning rate, which is a concave function, allowing the optimizer to adjust to the pre-trained network before learning at a normal learning rate. The learning rate was scheduled to increase from 1×10^{-5} to 2×10^{-3} in the first ten epochs and continued training with a decay of 1×10^{-6} until convergence.

3.3.0.2 Multi-task learning

TBI assessment was trained jointly with other source datasets to learn low and mid-level features. The decision of which layers to be shared was based on the hyperparameter tuning as follows. Layer 1-7 of VGGish, listed in Table 2.3, layer 1-3 of Wav2Vec, listed in Table 2.4, and layer 1-6 of SincNet, listed in Table 2.5, were jointly trained among all tasks. In the evaluation, we trained all combinations of source tasks together with the TBI task, which is either TBI classifier or TBI regressor. For subsequent layers that are specific to each learning task, we applied FC(s) with the activation function specifically to the source task as follows. Google audio dataset has two FCs with softmax activation function as in the original VGGish model. Librispeech corpus has the two FCs with softmax activation function as in the original paper of SincNet model. Lastly, WOZ dataset is connected to the same classifier with Sigmoid activation function as in the TBI classifier. The gradient update in each task was balanced using GradNorm [62], which normalizes the gradient and weights the loss (w_t) in each training batch using Equation 3.2.

$$L_{grad}(\theta_t) = \sum_t \left| \|\nabla_\theta w_t f_t(\theta_t)\| - \overline{\|\nabla_\theta w f(\theta)\|} \right| \tag{3.2}$$

3.3.0.3 Meta-learning

We selected Model-Agnostic Meta-Learning (MAML) [47] method as a meta-learning due to its fast adaptation and agnostic to the model architecture. However, we found that it is onerous to tune hyperparameters in MAML. We adopted MAML++ [63], which was proposed to solve training instability and second order derivertive cost of MAML, to train the model. The experiment also considered few-shots learning of MAML algorithm to indicate minimum number of samples required to train each model. We considered all combinations of the source datasets, as in the multi-task learning, to train in meta-learning and use the target task in the meta-testing step. In meta-testing, only training data of TBI speech were used to optimize model's parameters.

3.3.1 Evaluation metrics and baseline models

3.3.1.1 Evaluation Metrics

Our classification metrics include Balanced Accuracy (BAC) calculated as in Equation 3.3, F1 score (F1) calculated as in Equation 3.4, and Area under the ROC Curve (AUC) calculated by summing up all the trapezoid area under true positive rate and false positive rate curve. For regression, we report Root Mean Square Error (RMSE) calculated as in Equation 3.5, Mean Absolute Error (MAE) calculated as in Equation 3.6, and R-squared calculated as in Equation 3.7. The estimation error of each metric is reported in a form of standard error computed following Equation 3.8 where x_i denotes reported performance of validation fold i out of m folds. We configured the hyperparameter based on BAC in classification and RMSE in regression to select the most optimal model.

$$\text{BAC} = \frac{\text{Sensitivity} + \text{Specificity}}{2} \tag{3.3}$$

$$\text{F1} = \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \tag{3.4}$$

$$\text{RMSE} = \sqrt{\sum_{i=1}^{n} \frac{(\text{Predicted value}_i - \text{True value}_i)^2}{n}} \tag{3.5}$$

$$\text{MAE} = \sum_{i=1}^{n} \frac{|\text{Predicted value}_i - \text{True value}_i|}{n} \tag{3.6}$$

$$\text{R-squared} = 1 - \frac{\sum_{i=1}^{n}(\text{Predicted value}_i - \text{True value}_i)^2}{\sum_{i=1}^{n}(\text{Averaged true value} - \text{True value}_i)^2} \tag{3.7}$$

$$\text{Standard Error} = \frac{\sum_{i=1}^{\text{m}}(x_i - \overline{x})^2}{\sqrt{m}} \tag{3.8}$$

3.3.1.2 Baseline: single task learning and conventional machine learning

Single task learning is learning without any external dataset and is used as one of the baselines. We included a dropout layer [40] with a tuned dropout rate as a regularization method to reduce overfitting. Dropout was included in many state-of-the-art DNN models, even in cases where there was adequate data to train the models [41, 28]. Moreover, the number of parameters in each DNN was gradually reduced to obtain the best balance accuracy that equally reflects sensitivity (true positive rate) and specificity (true negative rate).

For conventional machine learning, we considered three feature extraction methods, which are COMPARE feature [64], Bag-Of-Audio-Word (BOAW) [65, 66] and formant coordination [7]. The TBI classification was performed by Support Vector Machine (SVM), Random Forest (RF) and Multi-Layer Perceptron (MLP) on each of the feature set with hyperparameter fine-tuning.

3.3.1.3 Extrapolation of TBI classification accuracy on numbers of training sample.

This study collected classification accuracy using 1, 5, 10, 25, 30, 40 and 50 (all) subjects from each class of Coelho corpus to estimate the numbers of samples require to reach a certain accuracy. Extrapolations of 75, 100 and 125 were projected using linear spline (Equation 3.9) and quadratic spline (Equation 3.10). $S_{linear,k}(x)$ and $S_{quad,k}(x)$ were used to represent accuracy y_k at point k with x_k samples under three conditions:

1. $S(x) = S_k(x); x_k \leq x \leq x_{k+1}; k = 0, 1, ..., n$

2. $S(x_k) = y_k; k = 0, 1, ..., n$

3. $S_k(x_{k+1}) = S_{k+1}(x_{k+1}); k = 0, 1, ..., n$

$$S_{linear,k}(x) = y_k + \frac{y_{k+1} - y_k}{x_{k+1} - x_k}(x - x_k) \tag{3.9}$$

$$S_{quad,k}(x) = y_k + z_k(x - x_k) + \frac{z_{k+1} - z_k}{2(x_{k+1} - x_k)}(x - x_k)^2 \tag{3.10}$$

$$z_{k+1} = -z_k + 2\frac{y_{k+1} - y_k}{x_{k+1} - x_k}, \text{where } z_0 = 0 \tag{3.11}$$

4. RESULT

This section compares three different techniques for limited data using TBI assessment as a target task. For each technique, there are three backbone networks and three external TBI-related datasets to be considered. Results are reported differently for classification and regression.

4.0.1 Transfer learning

The TBI classification results are based on the Coelho corpus, with other datasets being used to pre-train the model. From Table 4.1, SincNet outperforms other networks. Although VGGish has been frequently adopted in various deep learning-based speech assessments and performs well during the pre-training step, we could not modify the model to perform well in TBI detection.

In all backbone networks, pre-training with Librispeech leads to better performances than other datasets. While Librispeech corpus is neither the biggest corpus nor the most relative to TBI, it contains scripted speeches from professional speakers collected in a controlled environment. The training and validation loss are plotted in Figure 4.1. Compared to the baseline (Figure 4.2), the training and validation losses in transfer learning are lower than the baseline. The evaluation of SincNet with Librispeech was performed at epoch 16, where validation loss decreases to the lowest value before continuously increasing while training loss decreases, which indicates that the model starts overfitting to the training set, and early stopping should be executed. Although, the graph shows that the transfer learning starts overfitting around epoch 16, the gaps between validation loss and training loss are smaller than the single task learning.

In a regression task, the SincNet model is competitive with Wav2Vec as shown in Table 4.2. Similarly to the classification, the model with VGGish as a backbone network is underfitting, which we attempted but failed to modify the model for the TBI regression task. In each backbone model, pre-training with Librispeech corpus leads to better performance. While SincNet has a slightly lower RMSE and MAE than the Wav2Vec, Wav2Vec has a narrow standard error, which may infer that the Wav2Vec model is more stable than the SincNet model.

Table 4.1: TBI classification results using transfer learning technique

Model	Source dataset	BAC	F1 score	AUC
SincNet	Google audio	67.34 (1.12)	70.86 (1.34)	70.75 (1.44)
	Librispeech	**83.87 (1.64)**	**85.09 (2.58)**	**87.44 (1.20)**
	WOZ	72.15 (1.34)	75.86 (1.82)	81.41 (1.78)
Wav2Vec	Google audio	58.84 (1.30)	76.10 (1.98)	61.10 (0.96)
	Librispeech	**76.13 (1.19)**	**77.13 (2.19)**	**79.70 (1.33)**
	WOZ	75.22 (1.29)	75.90 (1.61)	78.37 (1.68)
VGGish	Google audio	51.81 (0.63)	68.42 (1.09)	62.70 (1.11)
	Librispeech	**56.25 (0.68)**	72.49 (1.12)	**67.38 (1.16)**
	WOZ	55.53 (0.72)	**74.65 (1.13)**	64.93 (1.12)
Baseline (Single-task learning)				
SincNet	-	62.74 (1.22)	69.26 (1.25)	65.62 (0.87)
Wav2Vec	-	61.99 (1.14)	65.60 (0.98)	68.35 (1.11)
VGGish	-	50.12 (0.48)	55.31 (0.60)	52.87 (0.63)

Table 4.2: TBI regression results using transfer learning technique

Model	Source dataset	RMSE	MAE	R-squared
SincNet	Google audio	1.79 (0.06)	1.01 (0.07)	0.85 (0.011)
	Librispeech	**1.14 (0.08)**	**0.82 (0.05)**	**0.96 (0.013)**
	WOZ	1.22 (0.09)	0.93 (0.05)	0.92 (0.017)
Wav2Vec	Google audio	1.64 (0.05)	1.18 (0.07)	0.86 (0.009)
	Librispeech	**1.16 (0.06)**	**0.93 (0.05)**	**0.96 (0.012)**
	WOZ	1.27 (0.07)	0.93 (0.04)	0.94 (0.015)
VGGish	Google audio	2.14 (0.07)	1.58 (0.06)	0.72 (0.020)
	Librispeech	2.15 (0.07)	**1.41 (0.03)**	**0.75 (0.017)**
	WOZ	**2.06 (0.06)**	1.48 (0.04)	0.74 (0.021)
Baseline (Single-task learning)				
SincNet	-	2.06 (0.11)	1.47 (0.07)	0.78 (0.013)
Wav2Vec	-	1.92 (0.09)	1.45 (0.06)	0.80 (0.011)
VGGish	-	2.62 (0.13)	1.52 (0.08)	0.67 (0.024)

Transfer learning exhibits improvements in classification and regression performances in all combinations of the datasets and backbone models. Domain adaptation allows the model to use and adapt knowledge from other domains into TBI detection, which is more efficient than starting from random parameters. It also aligns with how humans learn; we learn a new knowledge base

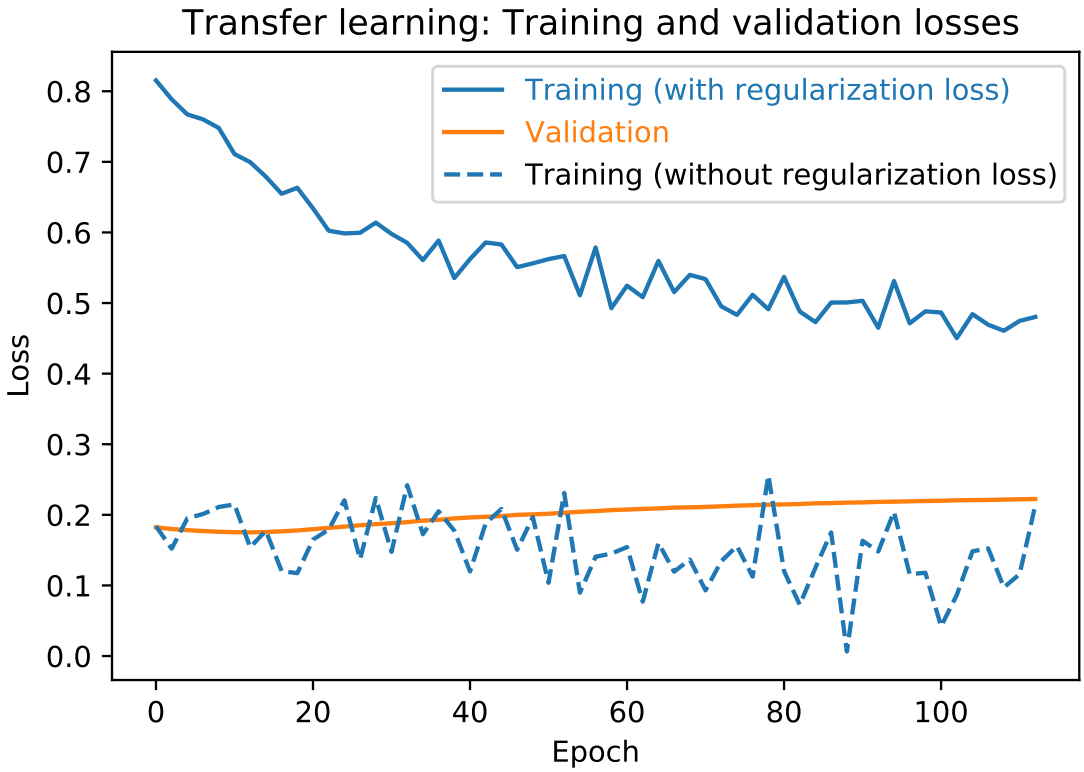

Figure 4.1: Training and validation losses of the SincNet model trained using transfer learning method with Librispeech as a source task

on the past knowledge. Compared to the other two techniques, transfer learning is straightforward and is easier to implement, which requires less effort for hyperparameter tuning and running time.

4.0.2 Multi-task learning

The multi-task learning results exclude VGGish model since it is not fitted with TBI detection tasks that depend on longer audio than in [26]. Results from 7 combinations of 3 datasets are shown in Table 4.3. TBI classification results obtained from training the Coelho dataset together with Librispeech and WOZ corpus outperform other dataset combinations in both SincNet and Wav2Vec models. Moreover, SincNet performs better than Wav2Vec in all dataset combinations. An increasing number of source datasets seems to improve the TBI detection performance, although using all datasets does not provide the highest balanced accuracy, which may occur because Google audio

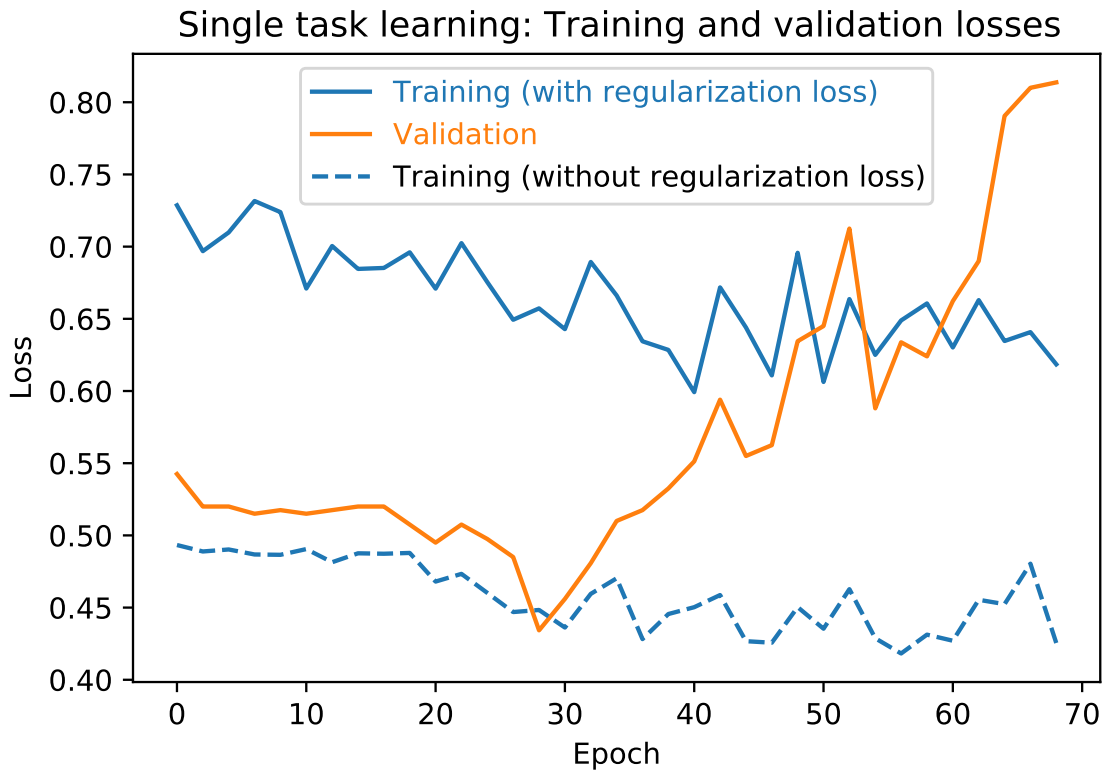

Figure 4.2: Training and validation losses of the SincNet model trained using single-task learning method (baseline)

does not specify to the speech domain but more generally to any types of audio recordings. A combination of Librispeech and WOZ that contains both scripted speech and spontaneous speech with training tasks to identify the speaker and assess depression aids the TBI classification model to learn common audio representations and, simultaneously, prevents the TBI classification model from overfitting. Overall, all SincNet and Wav2Vec models trained using multi-task provide better results than single-task learning.

For TBI regression task, a combination of Librispeech and WOZ outperforms other combinations, which is consistent with the classification results. However, the Wav2Vec model leads to a lower RMSE and MAE with high R-squares than the SincNet model. Comparatively, the RMSE gaps between the models using multi-task learning and the baseline are much larger than MAE, which indicates that the large error is penalized more than the small error in the training.

Table 4.3: TBI classification results using multi-task learning technique

Model	Source dataset	BAC	F1 score	AUC
SincNet	Google audio	66.13 (1.27)	74.75 (1.31)	71.14 (1.52)
	Librispeech	84.22 (1.00)	86.51 (1.59)	88.68 (1.61)
	WOZ	80.62 (1.49)	82.77 (1.67)	84.27 (1.69)
	Google audio + Librispeech	79.41 (1.32)	82.42 (1.44)	76.12 (1.65)
	Google audio + WOZ	78.92 (1.25)	80.02 (1.33)	76.25 (1.57)
	Librispeech + WOZ	**85.70 (1.18)**	**87.12 (1.51)**	**90.13 (1.43)**
	Google audio + Librispeech + WOZ	81.11 (1.62)	84.52 (1.28)	85.04 (1.77)
Wav2Vec	Google audio	67.42 (1.23)	74.26 (1.44)	69.03 (1.68)
	Librispeech	72.00 (1.35)	81.32 (1.20)	77.65 (1.52)
	WOZ	70.25 (1.34)	80.89 (1.32)	73.31 (1.43)
	Google audio + Librispeech	71.42 (1.18)	81.02 (1.30)	75.60 (1.44)
	Google audio + WOZ	70.27 (1.53)	79.88 (1.61)	72.68 (1.67)
	Librispeech + WOZ	**73.16 (1.35)**	**82.66 (1.31)**	**80.52 (1.60)**
	Google audio + Librispeech + WOZ	70.57 (1.37)	80.65 (1.51)	78.53 (1.52)
Baseline (Single-task learning)				
SincNet	-	62.74 (1.22)	69.26 (1.25)	65.62 (0.87)
Wav2Vec	-	61.99 (1.14)	65.60 (0.98)	68.35 (1.11)
VGGish	-	50.12 (0.48)	55.31 (0.60)	52.87 (0.63)

Overall, multi-task learning improves the baseline by jointly learning low-level feature representation of speech together with other tasks. When the model begins to overfit or capture any noise, the penalty from other tasks will be increased and prevent any learning that does not benefit all tasks. As visualized in Figure 4.3, training loss (without dropout) and validation loss attain the same error level throughout the training.

4.0.3 Meta-learning

According to BAC reported in Table 4.5, a combination of Google audio, Librispeech, and WOZ outperforms other source datasets. It also improves the baseline by 27% of BAC. The evaluation of meta-learning performs on the same source dataset combinations as in multi-task learning, but meta-learning seems to gain more advantage when there is more dataset. According to Table 4.5, a combination of Google audio, Librispeech and WOZ slightly outperforms the combination of Google audio and Librispeech, which is indicated as the best combination in multi-task

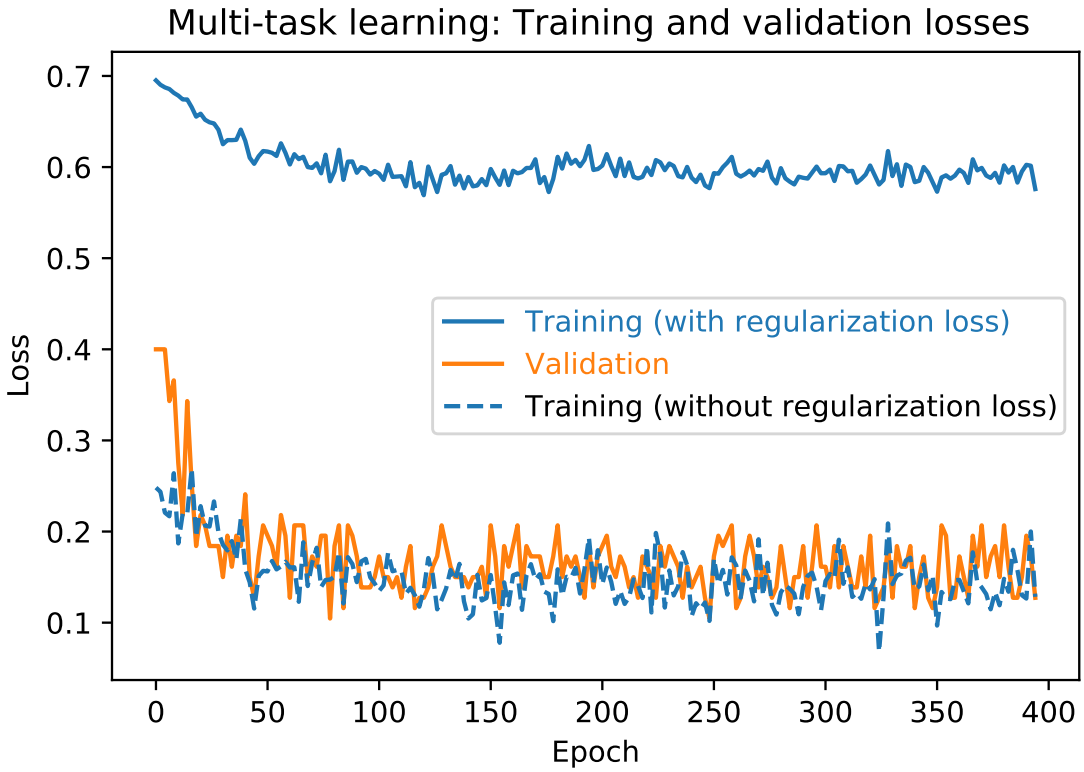

Figure 4.3: Training and validation losses of the SincNet model trained using multi-task learning method with Librispeech and WOZ as source tasks

learning. SincNet model performs better than the Wav2Vec model in all dataset combinations.

In the TBI regression task, a combination of Google audio, Librispeech, and WOZ outperforms other source dataset combinations, consistent with the TBI classification results. Wav2Vec outperforms the SincNet model, archiving an RMSE of 1.58, MAE of 1.16, and R-squared of 0.84.

MAML is also evidenced to perform well on few-shot learning, as shown in Figure 4.4. The TBI classification model reaches 79.32% BAC using only speeches from five TBI subjects and five healthy subjects. Meta-learning outperforms all other learning techniques when provided sample sizes are smaller than 40. We extrapolate the BAC plot in Figure 4.4 using linear fitting (Figure 4.5), and quadratic fitting (Figure 4.6). In the extrapolations of transfer learning and multi-task learning, the learning methods' BAC exceeds 100% around 52-55 samples. We are aware that there exists a plateau that prevents BAC from further increasing due to the bias in nature. From linear

Table 4.4: TBI regression results using multi-task learning technique

Model	Source dataset	RMSE	MAE	R-squared
SincNet	Google audio	1.82 (0.13)	1.32 (0.07)	0.81 (0.012)
	Librispeech	1.62 (0.12)	1.20 (0.06)	0.85 (0.012)
	WOZ	1.77 (0.11)	1.28 (0.06)	0.83 (0.011)
	Google audio + Librispeech	1.78 (0.13)	1.25 (0.08)	0.83 (0.013)
	Google audio + WOZ	1.80 (0.13)	1.26 (0.07)	0.83 (0.013)
	Librispeech + WOZ	**1.56 (0.13)**	**1.17 (0.08)**	**0.87 (0.012)**
	Google audio + Librispeech + WOZ	1.68 (0.13)	1.22 (0.06)	0.84 (0.011)
Wav2Vec	Google audio	1.71 (0.13)	1.28 (0.08)	0.83 (0.014)
	Librispeech	1.49 (0.13)	1.21 (0.07)	0.86 (0.013)
	WOZ	1.58 (0.15)	1.24 (0.07)	0.84 (0.015)
	Google audio + Librispeech	1.65 (0.14)	1.26 (0.07)	0.84 (0.013)
	Google audio + WOZ	1.68 (0.14)	1.23 (0.06)	0.82 (0.014)
	Librispeech + WOZ	**1.41 (0.14)**	**1.14 (0.08)**	**0.90 (0.017)**
	Google audio + Librispeech + WOZ	1.61 (0.13)	1.25 (0.06)	0.84 (0.012)
Baseline (Single-task learning)				
SincNet	-	2.06 (0.11)	1.47 (0.07)	0.78 (0.013)
Wav2Vec	-	1.92 (0.09)	1.45 (0.06)	0.80 (0.011)
VGGish	-	2.62 (0.13)	1.52 (0.08)	0.67 (0.024)

fitting, single-task learning and meta-learning do not improve much from gaining more data samples. However, an extrapolation of single-task learning using quadratic fitting gradually improves the BAC, which requires more than 125 samples to reach 90% BAC.

Meta-learning is a inner-outer iteration methods where the outer iteration aims to learn meta-weight that will need less optimization in the inner iteration of each task. This optimization mechanism makes few-shot learning (few inner-loop iteration) possible and provide a high detection accuracy compared to other learning methods that does not limit the optimization step or adaptation process in the target task.

Table 4.5: TBI classification results using meta-learning technique

Model	Source dataset	BAC	F1 score	AUC
SincNet	Google audio	66.27 (1.34)	71.11 (1.49)	70.51 (1.40)
	Librispeech	72.65 (1.32)	74.19 (1.47)	74.80 (1.26)
	WOZ	70.69 (1.35)	74.37 (1.33)	71.65 (1.30)
	Google audio + Librispeech	77.40 (1.35)	80.40 (1.62)	80.96 (1.35)
	Google audio + WOZ	73.89 (1.32)	77.63 (1.29)	78.56 (1.53)
	Librispeech + WOZ	**79.05 (1.16)**	81.09 (1.45)	81.25 (1.87)
	Google audio + Librispeech + WOZ	**79.32 (1.04)**	**82.36 (1.53)**	**81.41 (2.00)**
Wav2Vec	Google audio	64.28 (1.03)	70.25 (1.20)	69.53 (1.30)
	Librispeech	66.82 (0.94)	72.64 (0.88)	71.34 (0.99)
	WOZ	66.15 (1.04)	71.28 (1.11)	71.52 (1.24)
	Google audio + Librispeech	67.50 (1.35)	74.87 (1.31)	74.87 (1.34)
	Google audio + WOZ	66.72 (1.31)	71.08 (1.23)	71.25 (1.29)
	Librispeech + WOZ	**69.05 (1.22)**	**74.35 (1.19)**	77.32 (1.53)
	Google audio + Librispeech + WOZ	**69.29 (1.29)**	**74.38 (1.39)**	**78.75 (1.43)**
Baseline (Single-task learning)				
SincNet	-	62.74 (1.22)	69.26 (1.25)	65.62 (0.87)
Wav2Vec	-	61.99 (1.14)	65.60 (0.98)	68.35 (1.11)
VGGish	-	50.12 (0.48)	55.31 (0.60)	52.87 (0.63)

Table 4.6: TBI regression results using meta-learning technique

Model	Source dataset	RMSE	MAE	R-squared
SincNet	Google audio	1.93 (0.08)	1.43 (0.05)	0.78 (0.015)
	Librispeech	1.86 (0.08)	1.31 (0.06)	0.77 (0.014)
	WOZ	1.91 (0.09)	1.30 (0.07)	0.76 (0.015)
	Google audio + Librispeech	1.83 (0.07)	1.31 (0.07)	0.79 (0.014)
	Google audio + WOZ	1.88 (0.08)	1.29 (0.07)	0.80 (0.015)
	Librispeech + WOZ	**1.73 (0.08)**	**1.26 (0.08)**	**0.82 (0.014)**
	Google audio + Librispeech + WOZ	**1.71 (0.09)**	**1.24 (0.07)**	**0.83 (0.015)**
Wav2Vec	Google audio	1.88 (0.08)	1.37 (0.07)	0.79 (0.014)
	Librispeech	1.65 (0.08)	1.26 (0.06)	0.82 (0.013)
	WOZ	1.72 (0.08)	1.29 (0.06)	0.79 (0.015)
	Google audio + Librispeech	1.71 (0.09)	1.30 (0.07)	0.79 (0.016)
	Google audio + WOZ	1.78 (0.10)	1.34 (0.08)	0.79 (0.015)
	Librispeech + WOZ	**1.60 (0.10)**	**1.19 (0.07)**	**0.84 (0.015)**
	Google audio + Librispeech + WOZ	**1.58 (0.10)**	**1.16 (0.06)**	**0.84 (0.016)**
Baseline (Single-task learning)				
SincNet	-	2.06 (0.11)	1.47 (0.07)	0.78 (0.013)
Wav2Vec	-	1.92 (0.09)	1.45 (0.06)	0.80 (0.011)
VGGish	-	2.62 (0.13)	1.52 (0.08)	0.67 (0.024)

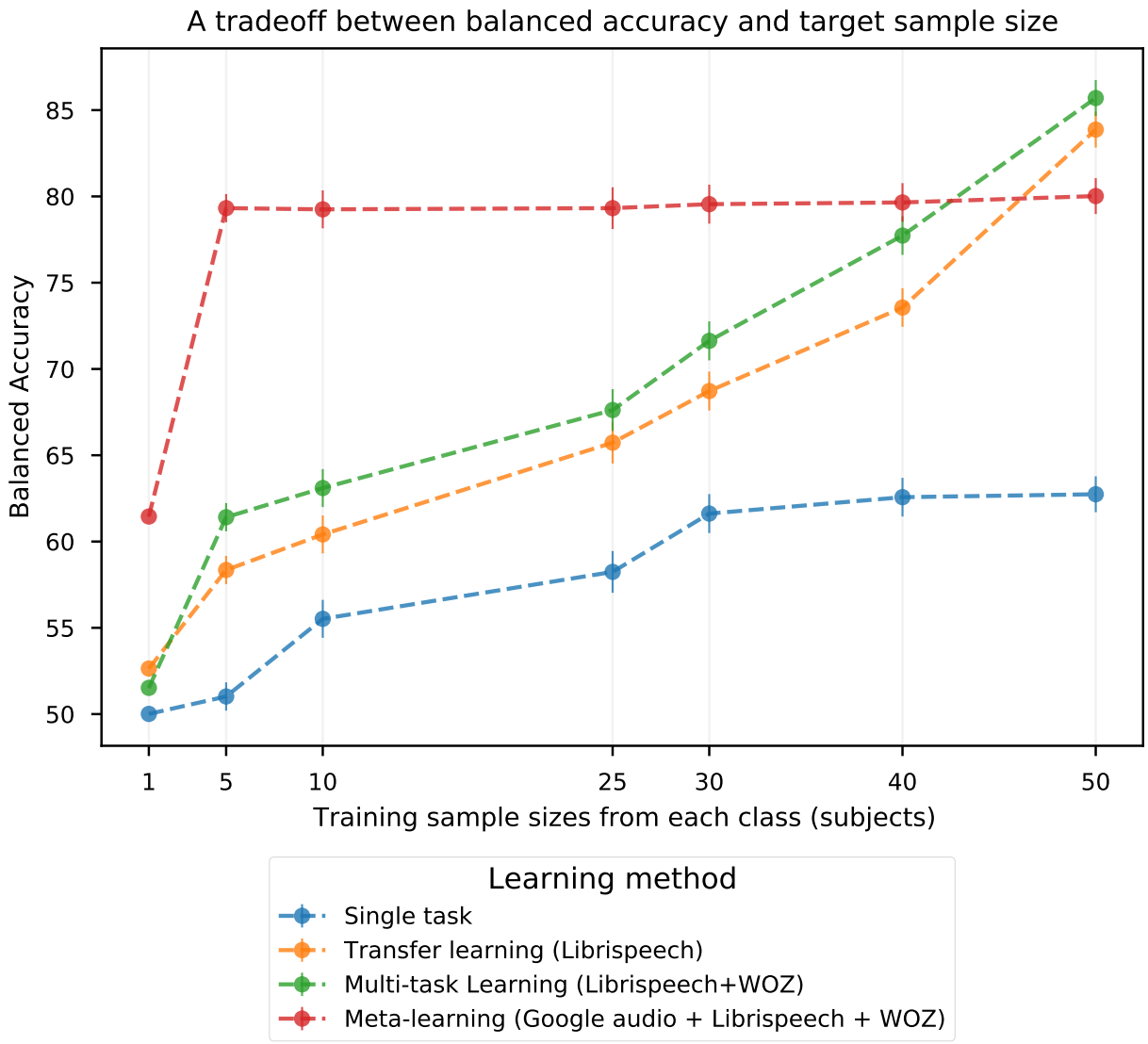

Figure 4.4: Few-shot learning: a trade-off between TBI detection accuracy and numbers of training sample

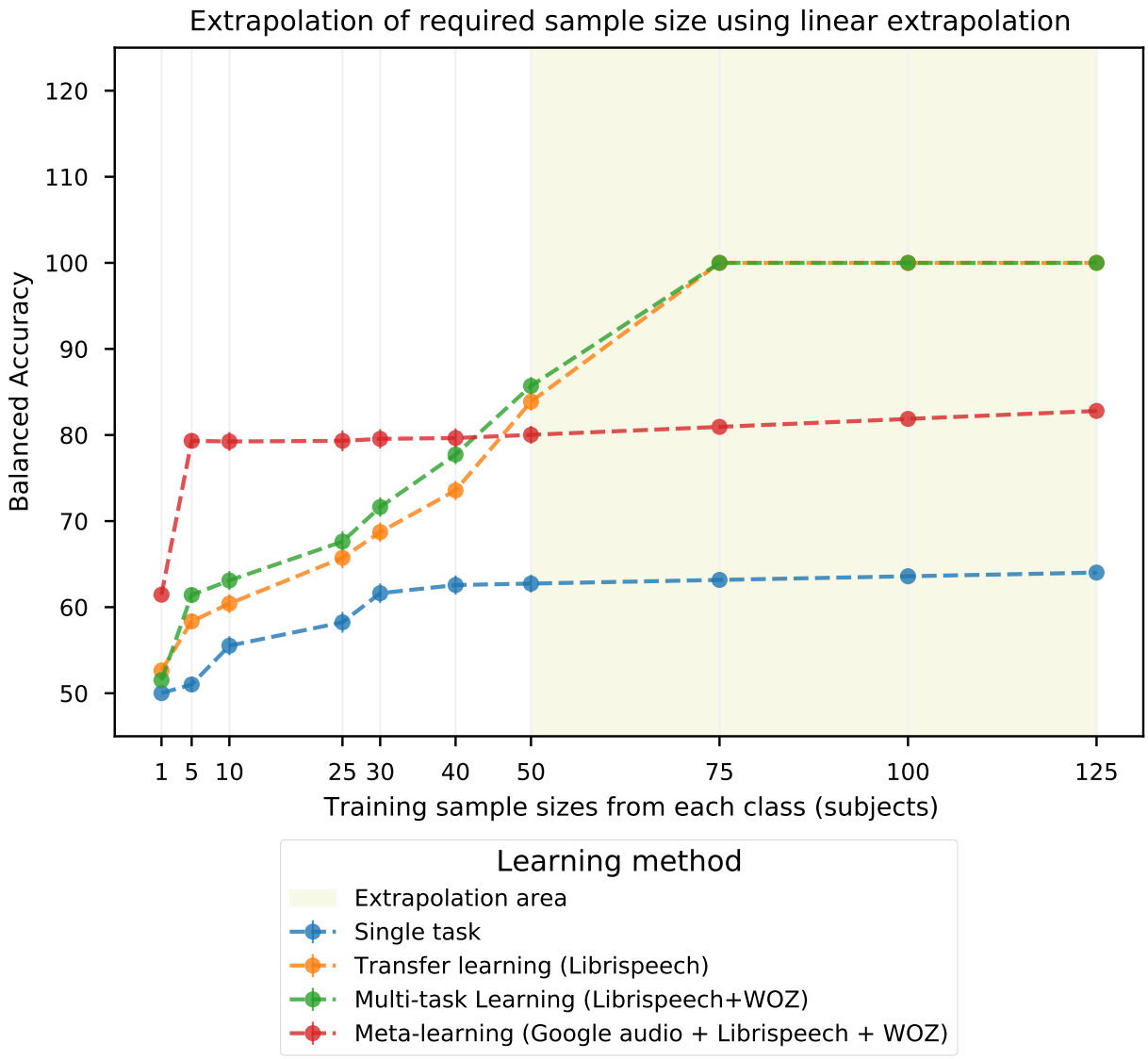

Figure 4.5: Improvement of TBI detection accuracy over an increasing size of training sample using linear fit: an extrapolation of TBI detection balanced accuracy on numbers of training samples from 50 to 125 subjects using linear fit.

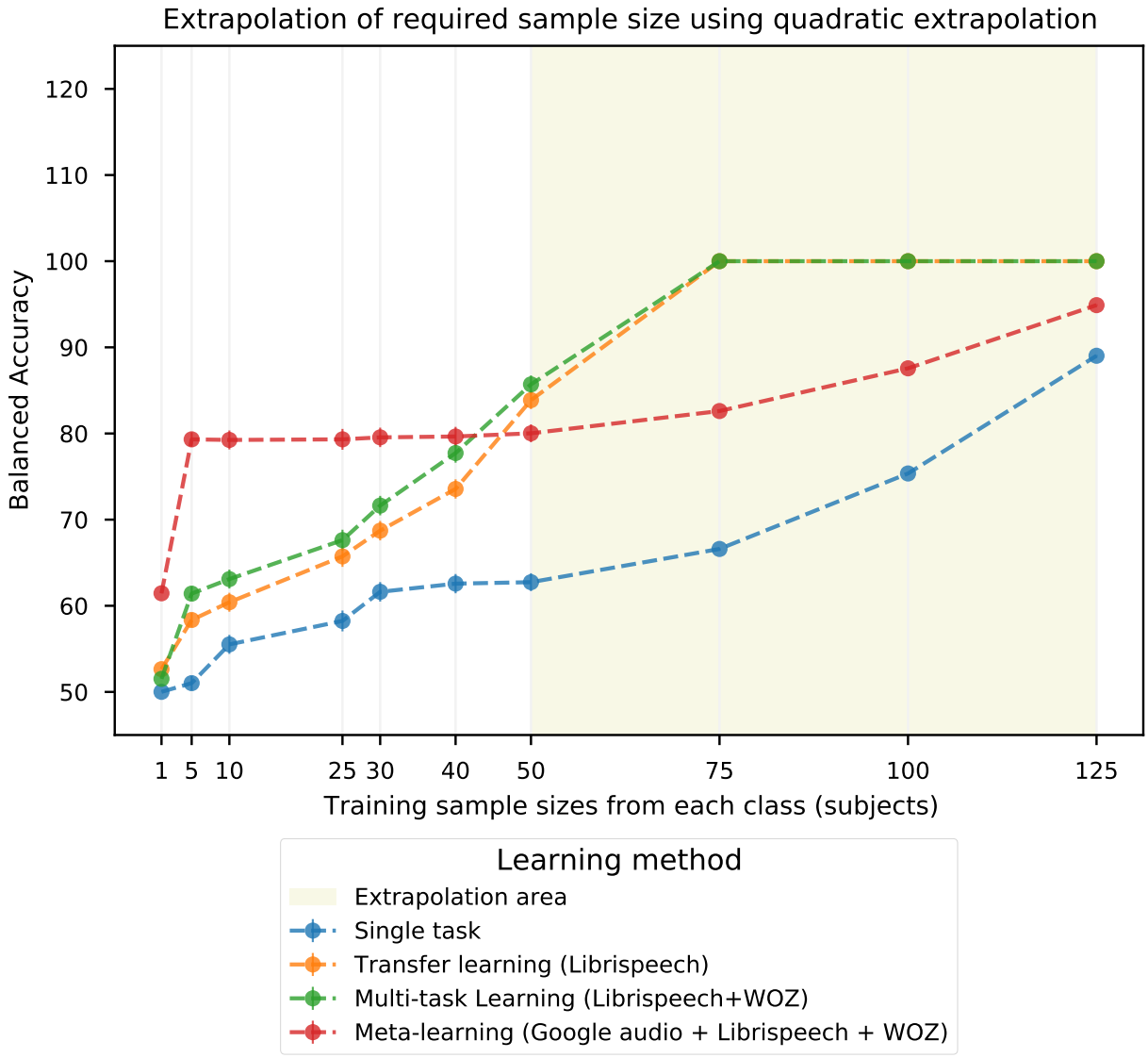

Figure 4.6: Improvement of TBI detection accuracy over an increasing size of training sample using quadratic fit: an extrapolation of TBI detection balanced accuracy on numbers of training samples from 50 to 125 subjects using quadratic fit.

5. DISCUSSION

This chapter gives answers to our research questions proposed in chapter 1 and concludes some experimental remarks.

Performance of deep learning in TBI detection compared to machine learning: In this thesis, we proposed a cascading DNN that detects TBI from features extracted by the backbone model over multiple time steps. We employ GRU as a core of the cascading mechanism because of its smaller number of trainable parameters compared to other gated RNN. From Table 5.1, the proposed cascading DNN outperforms COMPARE feature [64] but does not surpass TBI classification using Bag-Of-Audio-Word (BOAW) features [65]. After applying the learning methods for limited-data, it outperforms all baselines and archives a BAC of 85.70 using multi-task learning. Hence, deep learning with limited data outperforms traditional machine learning algorithms in TBI detection, agreeing with [10].

In regression task, transfer learning method outperforms multi-task learning, as shown in Table 5.2. It is worth to note that all source tasks are either binary or multi-class classifications and multi-task learning combines the loss of target task, which is RMSE, to the cross-entropy loss from source task, creating discrepancy in model optimizations. Similar to TBI classification, all learning methods outperform the baselines, whereas BOAW manifests the lowest RMSE among the baselines.

Best learning method for limited data: This study investigates three different learning methods that exploit external data to cope with overfitting problem. The decision to determine an appropriate learning method should be based on the available number of training samples. From Figure 4.5 and Figure 4.6, multi-task learning is predicted to reach a perfect classification score using fewer training examples than other methods, while meta-learning requires only a few examples (5-10) to reach its plateau. Hence, these two are the best learning methods depending on a trade-off between the desired detection accuracy and data collection cost.

Few-shot learning: The learning is called few-short learning if numbers of training sample is less

than ten; usually 1-shot and 5-shot leanings are usually used to validate the model in term of few-short learning. In both 1-shot and 5-shot learning, meta-learning using MAML algorithm gives the lowest TBI detection accuracy compared to transfer learning and multi-task learning. MAML demands less training samples of target task as it learns to perform a learning or adaptation to the new task as fast as possible using the second derivative of loss function. However, it is shown in the result that MAML stops improving although more training samples were provided. We suspect the reason behind this occurrence is due to the numbers of iteration in inner loop of MAML++ that dynamically changed during training, but fixed in the testing.

Effect of corpus size and specificity: All three learning methods get benefits from exploiting the external dataset, a source dataset that shares some knowledge with target dataset. The ideal source dataset is a corpus that is relatively large compared to target dataset. However, Google audio, which is the biggest corpus in this study with high generalization, does not improve the TBI classification and regression as well as the other two datasets. These surprising results indicate the specificity gap between source task and target task significantly impacts the TBI detection performance, which contradicts with the results in [29] reported that the adaptation from a source task with high generalization is more robust than from a source task with high specificity. In all learning methods, Librispeech corpus consistently outperforms other source datasets despite the fact that Librispeech corpus is not the biggest corpus but is closer to the TBI task. Hence, the Librispeech corpus could be used to represent a preferable size of the source dataset.

A setup for TBI detection: This thesis compares the TBI detection performances using different backbone networks and source-task datasets, which can be used as a guideline to construct a TBI detection system. The experiments in this study perform on clean speeches that prevents environmental noise and cross-talking. Employing this study to detect TBI in real-world would require a few preprocessing steps, such as [67], to ensure speech quality and trustworthy results.

Table 5.1: TBI classification improvement in SincNet using limited-data learning methods

Learning technique	BAC	F1	AUC
Transfer learning	83.87 (1.64)	85.09 (2.58)	87.44 (1.20)
Multi-task learning	**85.70 (1.18)**	**87.12 (1.51)**	**90.13 (1.43)**
Meta-learning	79.05 (1.16)	81.09 (1.45)	81.25 (1.87)
Baseline (Single-task learning)			
Single task learning	62.74 (1.22)	69.26 (1.25)	65.62 (0.87)
COMPARE feature			
SVM	56.37 (1.46)	60.14 (1.14)	59.62 (1.20)
RF	52.18 (0.96)	57.66 (1.30)	56.27 (1.23)
MLP	52.85 (1.21)	54.32 (1.43)	54.12 (1.15)
BOAW			
SVM	**66.05 (1.36)**	**74.44 (1.21)**	**71.41 (2.00)**
RF	50.14 (1.46)	53.72 (1.72)	52.17 (1.15)
MLP	62.97 (1.23)	66.74 (1.20)	65.28 (1.49)

Table 5.2: TBI regression improvement in Wav2Vec using limited-data learning methods

Learning technique	RMSE	MAE	R-squared
Transfer learning	**1.16 (0.06)**	**0.93 (0.05)**	**0.96 (0.012)**
Multi-task learning	1.41 (0.14)	1.14 (0.08)	0.90 (0.017)
Meta-learning	1.58 (0.10)	1.16 (0.06)	0.84 (0.016)
Baseline (Single-task learning)			
Single task learning	1.92 (0.09)	1.45 (0.06)	0.84 (0.016)
COMPARE feature			
SVM	2.61 (0.14)	1.45 (0.07)	0.70 (0.013)
RF	2.78 (0.13)	1.62 (0.07)	0.71 (0.013)
MLP	2.72 (0.13)	1.57 (0.07)	0.70 (0.014)
BOAW			
SVM	**1.86 (0.14)**	**1.34 (0.07)**	**0.80 (0.013)**
RF	2.14 (1.13)	1.49 (0.07)	0.78 (0.014)
MLP	2.08 (1.15)	1.46 (0.07)	0.78 (0.013)

6. CONCLUSION

Using DNN in a medical application, such as TBI detection, is challenging due to the data scarcity problem since collecting large amount of medical data is more expensive and more difficult in other domains. This thesis tackled an overfitting problem that commonly arises when training a deep learning model on limited labeled medical data. To take full-advantage of deep learning, we integrate learning method for limited data into deep learning-based TBI detection that inherently suffers a data scarcity problem. All three methods, namely, transfer learning, multi-task learning, and meta-learning, improved the TBI detection performances in both classification and regression tasks. Multi-task learning manifests the best improvement in TBI classification accuracy with no overfitting being observed. Transfer learning works best for the regression task, with an early stopping required to obtain its best performance before overfitting. In the scenario that audio samples are incredibly scarce, meta-learning is a promising method to perform few-shot learning in TBI detection tasks, which requires only a few subjects. Moreover, results in this study can estimate the amount of data needed for a task similar to TBI detection.

6.1 Future work

This study mainly focuses on the learning method for limited data, i.e., a method to train the model without modifying the network architecture. We do not include the data augmentation method, which synthesizes or mimics the actual data using traditional data processing or deep learning, such as Generative Adversarial Network (GAN). Previous work demonstrates the benefit and potential of incorporating GAN into speech assessment to solve overfitting problem [68, 69]. A future study will consider GAN and network architecture modification to mitigate the overfitting problem in a small dataset.

REFERENCES

[1] M. Faul, M. M. Wald, L. Xu, and V. G. Coronado, "Traumatic brain injury in the united states; emergency department visits, hospitalizations, and deaths, 2002-2006," 2010.

[2] F. M. Hammond, J. D. Corrigan, J. M. Ketchum, J. F. Malec, K. Dams-O'Connor, T. Hart, T. A. Novack, J. Bogner, M. N. Dahdah, and G. G. Whiteneck, "Prevalence of medical and psychiatric comorbidities following traumatic brain injury," The Journal of head trauma rehabilitation, vol. 34, no. 4, pp. E1–E10, 2019.

[3] M. A. Lindberg, S. A. Kiser, and E. M. M. Martin, "Mild tbi/concussion clinical tools for providers used within the department of defense and defense health agency," Federal practitioner, vol. 37, no. 9, p. 410, 2020.

[4] R. S. Norman, C. A. Jaramillo, M. Amuan, M. A. Wells, B. C. Eapen, and M. J. Pugh, "Traumatic brain injury in veterans of the wars in iraq and afghanistan: Communication disorders stratified by severity of brain injury," Brain injury, vol. 27, no. 13-14, pp. 1623–1630, 2013.

[5] M. McHenry, "Acoustic characteristics of voice after severe traumatic brain injury," The Laryngoscope, vol. 110, no. 7, pp. 1157–1161, 2000.

[6] M. Falcone, N. Yadav, C. Poellabauer, and P. Flynn, "Using isolated vowel sounds for classification of mild traumatic brain injury," in 2013 IEEE International Conference on Acoustics, Speech and Signal Processing, pp. 7577–7581, IEEE, 2013.

[7] B. S. Helfer, T. F. Quatieri, J. R. Williamson, L. Keyes, B. Evans, W. N. Greene, T. Vian, J. Lacirignola, T. Shenk, T. Talavage, et al., "Articulatory dynamics and coordination in classifying cognitive change with preclinical mtbi," in Fifteenth Annual Conference of the International Speech Communication Association, 2014.

[8] C. Poellabauer, N. Yadav, L. Daudet, S. L. Schneider, C. Busso, and P. J. Flynn, "Challenges in concussion detection using vocal acoustic biomarkers," IEEE Access, vol. 3, pp. 1143–1160, 2015.

[9] L. Daudet, N. Yadav, M. Perez, C. Poellabauer, S. Schneider, and A. Huebner, "Portable mtbi assessment using temporal and frequency analysis of speech," IEEE journal of biomedical and health informatics, vol. 21, no. 2, pp. 496–506, 2016.

[10] T. Talkar, S. Yuditskaya, J. R. Williamson, A. Lammert, H. Rao, D. Hannon, A. O'Brien, G. Vergara-Diaz, R. DeLaura, D. Sturim, et al., "Detection of subclinical mild traumatic brain injury (mtbi) through speech and gait," Proc. Interspeech 2020, pp. 135–139, 2020.

[11] B. E. Murdoch and D. G. Theodoros, "Traumatic brain injury: Associated speech, language, and swallowing disorders," 2001.

[12] C. Cortes and V. Vapnik, "Support-vector networks," Machine learning, vol. 20, no. 3, pp. 273–297, 1995.

[13] N. O'Mahony, S. Campbell, A. Carvalho, S. Harapanahalli, G. V. Hernandez, L. Krpalkova, D. Riordan, and J. Walsh, "Deep learning vs. traditional computer vision," in Science and Information Conference, pp. 128–144, Springer, 2019.

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

[15] X. Qiu, T. Sun, Y. Xu, Y. Shao, N. Dai, and X. Huang, "Pre-trained models for natural language processing: A survey," arXiv preprint arXiv:2003.08271, 2020.

[16] J. Wagner, D. Schiller, A. Seiderer, and E. André, "Deep learning in paralinguistic recognition tasks: Are hand-crafted features still relevant?," in Interspeech, pp. 147–151, 2018.

[17] D. Anderson and K. Burnham, "Model selection and multi-model inference," Second. NY: Springer-Verlag, vol. 63, no. 2020, p. 10, 2004.

[18] T. Poggio, K. Kawaguchi, Q. Liao, B. Miranda, L. Rosasco, X. Boix, J. Hidary, and H. Mhaskar, "Theory of deep learning iii: explaining the non-overfitting puzzle," arXiv preprint arXiv:1801.00173, 2017.

[19] C. Zhang, S. Bengio, M. Hardt, B. Recht, and O. Vinyals, "Understanding deep learning requires rethinking generalization," arXiv preprint arXiv:1611.03530, 2016.

[20] C. H. Lee and H.-J. Yoon, "Medical big data: promise and challenges," Kidney research and clinical practice, vol. 36, no. 1, p. 3, 2017.

[21] R. Wenke, M. Lawrie, T. Hobson, W. Comben, M. Romano, E. Ward, and E. Cardell, "Feasibility and cost analysis of implementing high intensity aphasia clinics within a sub-acute setting," International Journal of Speech-Language Pathology, vol. 16, no. 3, pp. 250–259, 2014.

[22] C. A. Coelho, K. M. Youse, and K. N. Le, "Conversational discourse in closed-head-injured and non-brain-injured adults," Aphasiology, vol. 16, no. 4-6, pp. 659–672, 2002.

[23] E. Elbourn, B. Kenny, E. Power, and L. Togher, "Psychosocial outcomes of severe traumatic brain injury in relation to discourse recovery: a longitudinal study up to 1 year post-injury," American journal of speech-language pathology, vol. 28, no. 4, pp. 1463–1478, 2019.

[24] I. Rebai, Y. BenAyed, W. Mahdi, and J.-P. Lorré, "Improving speech recognition using data augmentation and acoustic model fusion," Procedia Computer Science, vol. 112, pp. 316–322, 2017.

[25] L. Yang, D. Jiang, and H. Sahli, "Feature augmenting networks for improving depression severity estimation from speech signals," IEEE Access, vol. 8, pp. 24033–24045, 2020.

[26] S. Hershey, S. Chaudhuri, D. P. Ellis, J. F. Gemmeke, A. Jansen, R. C. Moore, M. Plakal, D. Platt, R. A. Saurous, B. Seybold, et al., "Cnn architectures for large-scale audio classification," in 2017 ieee international conference on acoustics, speech and signal processing (icassp), pp. 131–135, IEEE, 2017.

[27] S. Schneider, A. Baevski, R. Collobert, and M. Auli, "wav2vec: Unsupervised pre-training for speech recognition.," in INTERSPEECH, 2019.

[28] M. Ravanelli and Y. Bengio, "Speaker recognition from raw waveform with sincnet," in 2018 IEEE Spoken Language Technology Workshop (SLT), pp. 1021–1028, IEEE, 2018.

[29] J. Yosinski, J. Clune, Y. Bengio, and H. Lipson, "How transferable are features in deep neural networks?," arXiv preprint arXiv:1411.1792, 2014.

[30] Z. Zhao, Z. Bao, Z. Zhang, N. Cummins, H. Wang, and B. Schuller, "Hierarchical attention transfer networks for depression assessment from speech," in ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 7159–7163, IEEE, 2020.

[31] Z. Huang, J. Epps, D. Joachim, B. Stasak, J. R. Williamson, and T. F. Quatieri, "Domain adaptation for enhancing speech-based depression detection in natural environmental conditions using dilated cnns," Interspeech 2020, pp. 4561–4565, 2020.

[32] E. Rejaibi, A. Komaty, F. Meriaudeau, S. Agrebi, and A. Othmani, "Mfcc-based recurrent neural network for automatic clinical depression recognition and assessment from speech," arXiv preprint arXiv:1909.07208, 2019.

[33] S. A. Qureshi, S. Saha, M. Hasanuzzaman, and G. Dias, "Multitask representation learning for multimodal estimation of depression level," IEEE Intelligent Systems, vol. 34, no. 5, pp. 45–52, 2019.

[34] L. Chao, J. Tao, M. Yang, and Y. Li, "Multi task sequence learning for depression scale prediction from video," in 2015 International Conference on Affective Computing and Intelligent Interaction (ACII), pp. 526–531, IEEE, 2015.

[35] N. R. Koluguri, M. Kumar, S. H. Kim, C. Lord, and S. Narayanan, "Meta-learning for robust child-adult classification from speech," in ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 8094–8098, IEEE, 2020.

[36] Mar 2019.

[37] A. I. Schneiderman, E. R. Braver, and H. K. Kang, "Understanding sequelae of injury mechanisms and mild traumatic brain injury incurred during the conflicts in iraq and afghanistan: persistent postconcussive symptoms and posttraumatic stress disorder," American journal of epidemiology, vol. 167, no. 12, pp. 1446–1452, 2008.

[38] A. C. Lammert, J. R. Williamson, A. Hess, T. Patel, T. F. Quatieri, H. J. Liao, A. Lin, and K. J. Heaton, "Noninvasive estimation of cognitive status in mild traumatic brain injury using speech production and facial expression," in 2017 Seventh International Conference on Affective Computing and Intelligent Interaction (ACII), pp. 105–110, IEEE, 2017.

[39] I. Goodfellow, Y. Bengio, and A. Courville, Deep Learning. MIT Press, 2016. `http://www.deeplearningbook.org`.

[40] G. E. Hinton, N. Srivastava, A. Krizhevsky, I. Sutskever, and R. R. Salakhutdinov, "Improving neural networks by preventing co-adaptation of feature detectors," arXiv preprint arXiv:1207.0580, 2012.

[41] C. Szegedy, S. Ioffe, V. Vanhoucke, and A. Alemi, "Inception-v4, inception-resnet and the impact of residual connections on learning," arXiv preprint arXiv:1602.07261, 2016.

[42] D. Banerjee, K. Islam, K. Xue, G. Mei, L. Xiao, G. Zhang, R. Xu, C. Lei, S. Ji, and J. Li, "A deep transfer learning approach for improved post-traumatic stress disorder diagnosis," Knowledge and Information Systems, vol. 60, no. 3, pp. 1693–1724, 2019.

[43] S. Ghosh, E. Laksana, L.-P. Morency, and S. Scherer, "Representation learning for speech emotion recognition.," in Interspeech, pp. 3603–3607, 2016.

[44] K. Feng and T. Chaspari, "A review of generalizable transfer learning in automatic emotion recognition," Frontiers in Computer Science, vol. 2, p. 9, 2020.

[45] Y. Li, T. Zhao, and T. Kawahara, "Improved end-to-end speech emotion recognition using self attention mechanism and multitask learning.," in Interspeech, pp. 2803–2807, 2019.

[46] R. Lotfian and C. Busso, "Predicting categorical emotions by jointly learning primary and secondary emotions through multitask learning.," in INTERSPEECH, pp. 951–955, 2018.

[47] C. Finn, P. Abbeel, and S. Levine, "Model-agnostic meta-learning for fast adaptation of deep networks," arXiv preprint arXiv:1703.03400, 2017.

[48] G. Rizos and B. Schuller, "Modelling sample informativeness for deep affective computing," in IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 3482–3486, IEEE, 2019.

[49] G. I. Winata, S. Cahyawijaya, Z. Liu, Z. Lin, A. Madotto, P. Xu, and P. Fung, "Learning fast adaptation on cross-accented speech recognition," arXiv preprint arXiv:2003.01901, 2020.

[50] Y. LeCun, B. Boser, J. S. Denker, D. Henderson, R. E. Howard, W. Hubbard, and L. D. Jackel, "Backpropagation applied to handwritten zip code recognition," Neural computation, vol. 1, no. 4, pp. 541–551, 1989.

[51] Y. Li and Y. Yuan, "Convergence analysis of two-layer neural networks with relu activation," in NIPS, 2017.

[52] Y.-T. Zhou and R. Chellappa, "Computation of optical flow using a neural network.," in ICNN, pp. 71–78, 1988.

[53] J. Chung, C. Gulcehre, K. Cho, and Y. Bengio, "Empirical evaluation of gated recurrent neural networks on sequence modeling," arXiv preprint arXiv:1412.3555, 2014.

[54] J. F. Gemmeke, D. P. Ellis, D. Freedman, A. Jansen, W. Lawrence, R. C. Moore, M. Plakal, and M. Ritter, "Audio set: An ontology and human-labeled dataset for audio events," in 2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 776–780, IEEE, 2017.

[55] D. Povey, A. Ghoshal, G. Boulianne, L. Burget, O. Glembek, N. Goel, M. Hannemann, P. Motlicek, Y. Qian, P. Schwarz, et al., "The kaldi speech recognition toolkit," in IEEE 2011 workshop on automatic speech recognition and understanding, no. CONF, IEEE Signal Processing Society, 2011.

[56] V. Panayotov, G. Chen, D. Povey, and S. Khudanpur, "Librispeech: an asr corpus based on public domain audio books," in 2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 5206–5210, IEEE, 2015.

[57] J. Gratch, R. Artstein, G. M. Lucas, G. Stratou, S. Scherer, A. Nazarian, R. Wood, J. Boberg, D. DeVault, S. Marsella, et al., "The distress analysis interview corpus of human and computer interviews.," in LREC, pp. 3123–3128, 2014.

[58] L. Lee and R. C. Rose, "Speaker normalization using efficient frequency warping procedures," in 1996 IEEE International Conference on Acoustics, Speech, and Signal Processing Conference Proceedings, vol. 1, pp. 353–356, IEEE, 1996.

[59] F. Seide, G. Li, X. Chen, and D. Yu, "Feature engineering in context-dependent deep neural networks for conversational speech transcription," in 2011 IEEE Workshop on Automatic Speech Recognition & Understanding, pp. 24–29, IEEE, 2011.

[60] A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, Z. Lin, N. Gimelshein, L. Antiga, A. Desmaison, A. Kopf, E. Yang, Z. DeVito, M. Raison, A. Tejani, S. Chilamkurthy, B. Steiner, L. Fang, J. Bai, and S. Chintala, "Pytorch: An imperative style, high-performance deep learning library," in Advances in Neural Information Processing Systems 32 (H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, eds.), pp. 8024–8035, Curran Associates, Inc., 2019.

[61] K. Sechidis, G. Tsoumakas, and I. Vlahavas, "On the stratification of multi-label data," in Joint European Conference on Machine Learning and Knowledge Discovery in Databases, pp. 145–158, Springer, 2011.

[62] Z. Chen, V. Badrinarayanan, C.-Y. Lee, and A. Rabinovich, "Gradnorm: Gradient normalization for adaptive loss balancing in deep multitask networks," in International Conference on Machine Learning, pp. 794–803, PMLR, 2018.

[63] A. Antoniou, H. Edwards, and A. Storkey, "How to train your maml," arXiv preprint arXiv:1810.09502, 2018.

[64] B. Schuller, S. Steidl, A. Batliner, A. Vinciarelli, K. Scherer, F. Ringeval, M. Chetouani, F. Weninger, F. Eyben, E. Marchi, et al., "The interspeech 2013 computational paralinguistics challenge: Social signals, conflict, emotion, autism," in Proceedings INTERSPEECH 2013, 14th Annual Conference of the International Speech Communication Association, Lyon, France, 2013.

[65] M. Schmitt and B. Schuller, "Openxbow: introducing the passau open-source crossmodal bag-of-words toolkit," The Journal of Machine Learning Research, vol. 18, no. 1, pp. 3370–3374, 2017.

[66] C. Kohlschein, M. Schmitt, B. Schüller, S. Jeschke, and C. J. Werner, "A machine learning based system for the automatic evaluation of aphasia speech," in 2017 IEEE 19th International Conference on e-Health Networking, Applications and Services (Healthcom), pp. 1–6, IEEE, 2017.

[67] A. Ditthapron, E. O. Agu, and A. C. Lammert, "Privacy-preserving deep speaker separation for smartphone-based passive speech assessment," IEEE Open Journal of Engineering in Medicine and Biology, 2021.

[68] A. Chatziagapi, G. Paraskevopoulos, D. Sgouropoulos, G. Pantazopoulos, M. Nikandrou, T. Giannakopoulos, A. Katsamanis, A. Potamianos, and S. Narayanan, "Data augmentation using gans for speech emotion recognition.," in Interspeech, pp. 171–175, 2019.

[69] S. H. Yang and M. Chung, "Improving dysarthric speech intelligibility using cycle-consistent adversarial training," arXiv preprint arXiv:2001.04260, 2020.