A DEEP LEARNING METHOD FOR INNER SPEECH CLASSIFICATION USING EEG SIGNAL

By

MUHAMMAD AMEER HAMZA

NATIONAL UNIVERSITY OF MODERN LANGUAGES

ISLAMABAD

July, 2025

A Deep Learning Method For Inner Speech Classification Using EEG Signal

By

MUHAMMAD AMEER HAMZA

BSCS, National University of Modern Languages, Islamabad, 2019

A THESIS SUBMITTED IN PARTIAL FULFILLMENT OF

THE REQUIREMENTS FOR THE DEGREE OF

MASTER OF SCIENCE

In **Computer Science**

To

FACULTY OF ENGINEERING & COMPUTING

NATIONAL UNIVERSITY OF MODERN LANGUAGES ISLAMABAD

NATIONAL UNIVERSITY OF MODERN LANGUAGES FACUTY OF ENGINEERING & COMPUTING

THESIS AND DEFENSE APPROVAL FORM

The undersigned certify that they have read the following thesis, examined the defense, are satisfied with overall exam performance, and recommend the thesis to the Faculty of Engineering and Computing for acceptance.

Thesis Title: A Deep Learning Method for Inner Speech Classification Using EEG signal

Submitted By: Muhammad Ameer Hamza **Registration #: S22-29473**

Master of Science in Computer Science (MSCS) Computer Science

Degree Name in Full Name of Discipline

Dr. QuaratulAin Safdar _____

Research Supervisor Signature of Research Supervisor

Dr. Fazli Subhan _____

Head of Department (CS) Signature of HOD (CS)

Dr, M. Noman Malik _____

Name of Dean (FEC) Signature of Dean (FEC)

July 16th, 2025

AUTHOR'S DECLARATION

I Muhammad Ameer Hamza

Son of Malik Muhammad Tayyib

Registration # S22-29473

Discipline Computer Science

Candidate of **Master of Science in Computer Science (MSCS)** at the National University of Modern Languages do hereby declare that the thesis **A DEEP LEARNING METHOD FOR INNER SPEECH CLASSIFICATION USING EEG SIGNAL** submitted by me in partial fulfillment of MSCS degree, is my original work, and has not been submitted or published earlier. I also solemnly declare that it shall not, in future, be submitted by me for obtaining any other degree from this or any other university or institution. I also understand that if evidence of plagiarism is found in my thesis/dissertation at any stage, even after the award of a degree, the work may be cancelled and the degree revoked.

Signature of Candidate

Muhammad Ameer Hamza
Name of Candidate

16th July, 2025
Date

ABSTRACT

Title: A Deep Learning Method For Inner Speech Classification Using EEG Signal

This thesis studies utilizing electroencephalography (EEG) signals the difficulties of subject-independent inner speech classification. Particularly for those with severe motor disabilities, inner speech the act of silently communicating to oneself offers a potential modality for Brain-Computer Interfaces (BCIs). Low signal-to-noise ratios and great inter-subject variability make deciphering inner speech from EEG difficult, nevertheless. Based on EEG data, this thesis evaluates several machine learning and deep learning models for inner speech classification. Particularly, a deep learning model, a Convolutional Neural Network (CNN) with triplet loss, is contrasted against conventional machine learning methods including Linear Support Vector Machine (SVM), More general SVM with various kernels, and LightGBM. Subject-independent framework with leave-one-subject-out cross-valuation on the Thinking Out Loud (TOL) dataset evaluates the models. Performance is evaluated with reference to accuracy, F1-score, precision, and recall. The CNN-based triplet network achieves the best average accuracy among other models, so the results show the promise of deep learning for subject-independent inner speech classification. Although the results imply that deep learning presents a viable path for future research, especially with bigger and more diverse datasets and advanced architectures, benefits over conventional approaches are minor. This work advances knowledge of the difficulties and possible solutions for creating strong, generally applicable inner voice BCIs.

TABLE OF CONTENT

LIST OF TABLES

LIST OF FIGURES

LIST OF ABBREVIATIONS

TOL	-	Thinking out loud dataset
EEG	-	Electroencephalography
BCI	-	Brain Computer interface

ACKNOWLEDGEMENT

First and most importantly, I would be grateful to Almighty Allah for enabling this study and for its success. This study would not have been finished without the help of many people to whom I would want to sincerely thank you. My accomplishment was mostly the result of several people, and I am appreciative of their help particularly of my research supervisors, who regularly directed me during my study process.

DEDICATION

This thesis is dedicated to my parents, whose unwavering love has been my pillar, and to my professors over the years, whose direction has molded my academic path and taught me the need of diligence.

CHAPTER 1

INTRODUCTION

1.1 Overview

Brain-Computer Interface (BCIs) are new techniques circumventing conventional neuromuscular paths by means of a direct communication link between the human brain and outside equipment. This technology allows humans to control robotic limbs, computers, or other equipment just with their brain activity. BCIs change human-computer interaction and provide new opportunities for people with impairments since they basically provide a new communication channel from the brain to the outer environment. Thus, the BCI offers great opportunities for persons who have lost physical or language abilities from neurological illnesses or other constraints. By translating brain signals into executable commands, BCIs provide tools to reestablish communication, control assistive technology, and finally increase quality of life [1]. A case study for instance showed the possibility of BCIs for communication, self-expression, and social connection when a patient with locked-in ALS effectively long-term utilized a BCI for autonomous painting at home [2].

Usually fitting either invasive or non-invasive categories are BCIs. By physically implanting electrodes precisely into the brain, invasive BCIs generate a high-density signal for recording and stimulation of neural activity. Although it offers improved signal quality and accuracy, this approach reveals natural risks connected to surgery including infection, bleeding, and probably tissue damage. On the other hand, non-invasive BCIs evaluate brain activity from the scalp or surface of the head using external sensors, such as those used in electroencephalography (EEG) or functional near-infrared spectroscopy (fNIRS). These methods are more prone to noise and artifacts even if they have superior safety and simplicity of use than invasive BCIs. Their spatial resolution is really poor generally. Using non-invasive

BCI method electroencephalography (EEG), the current thesis explores the subtleties of inner speech classification.

1.1.2 Non-Invasive BCI

Unlike non-invasive techniques, invasive BCIs surgically put electrodes right into the brain. This method naturally carries hazards including brain damage, infection, and surgical problems even if it offers acceptable signal quality [3]. Usually depending on electroencephalography (EEG) to capture brain waves from the scalp, non-invasive techniques usually have certain benefits instead of these hazards. Examining several brain activities including sleep, emotion, and cognition, this safe and non-invasive method is EEG especially suitable for studies on inner speech since EEG can record the minute changes in brain activity resulting from inner speech. Non-invasive BCIs such as EEG offer accessibility, simplicity of use, reduced expense, and most likely application in both research and pragmatic contexts. Moreover, non-invasive methods are not useful for BCI development because of ethical issues and patient comfort they entail [3]. Particularly, electroencephalogram (EEG) measurements, non-invasive BCIs have shown good efficacy in treating severely and partially disabled people. These BCIs restore basic control tools by allowing individuals to recover prosthetic limbs and wheelchairs as well as communication skills. Researchers have created BCI systems, for example, allowing paraplegic people to type words on a computer screen, maneuver a wheelchair or robotic arm merely using their thoughts.

1.1.3 Inner Speech: A Natural BCI Modality

Development of EEG-based systems has much benefited from common BCI methods as those using P300 waves, SSVP visual stimuli, or motor imagery. These methods, meanwhile, are sometimes slow or demanding for consumers, which limits their practicality for daily and extended usage. Researchers are looking at speech-related techniques such silent, imagined, or inner speech in BCI systems as a more simple approach to manage devices in order to meet these difficulties. It is well known that speech generation is a complicated process including

auditory processing, semantics, syntax, and articulatory motions. Though these speech-related techniques have great promise, their particular definitions and differences in the literature remain very vague. [3}

Among these techniques, inner speech—also known as verbal thinking or internal monologue offers a particularly interesting path for BCI control because of its covert character and absence of reliance on overt speech articulation or muscle movements. For those with severe movement disabilities, BCIs can offer a more natural and simple way of communication and control by directly decoding the brain correlates of inner speech. It is easier and more natural for humans.

Inner speech has a long history, beginning in the writings of Plato and Socrates. Still, scientific study of inner speech started only in the early 20th century. According to Soviet psychologist Lev Vygotsky, a process starting in childhood leads from exterior speech to internal speech. Young children commonly express their ideas and behaviors loudly as they engage with their surroundings. This "self-talk" gradually becomes absorbed and turns into the quiet inner speech we know from adulthood. Vygotsky claimed that complex reasoning, planning, and problem-solving all depend on this internalization of language, therefore supporting cognitive development.

More recently, neuroimaging methods including fMRI and EEG have been used to explore the brain foundations of inner speech with conventional focus on the left hemisphere; these investigations have shed light on the complex network of brain areas engaged in this cognitive process. Although areas like the frontal lobe, temporal lobe, and parietal lobe all help to process inner speech, a studies challenge the notion of the left hemisphere being solely dominant even if their results also strongly support the significance of particular areas within the *left* hemisphere. For example, although some research employing EEG have showed greater activity in the left hemisphere during imagined speech classification, this result should be seen in light of the complex and multifarious network supporting inner speech *within that hemisphere. [16]

Using lesion analysis, a method able to pinpoint areas vital for a certain function rather than those just engaged, a study looked at the neurological correlates of inner speech [16]. Using

voxel-based lesion-symptom mapping to link lesion locations with performance on inner speech tasks (such as rhyme and homophone judgments), overt speech tasks (like reading aloud), and working memory tasks (including sentence repetition), their study comprised 17 patients with chronic post-stroke aphasia. This method enabled a comparison of neural correlates especially important for inner speech, separating them from those required for overt speech and working memory. Lesions to the left pars opercularis in the inferior frontal gyrus (BA 44) and to the white matter close to the left supramarginal gyrus (BA 40) clearly influenced inner speech abilities, the study showed. These findings underline the function of several non-motor cognitive processes in inner speech within the left hemisphere and imply that inner speech is not only overt speech minus a motor component.

The research also emphasizes the need of discriminating between various kinds of inner speech and suggests that activities requiring active monitoring, such rhyming and homophone judgments, can depend on "conscious inner speech," whilst other tasks might generate a less aware version. The study emphasizes how strongly the left inferior frontal gyrus—especially the pars opercularis—appears to be linked to the more conscious kind of inner speech. Moreover, the participation of white matter close to the left supramarginal gyrus points to a vital function for the dorsal language route in the processing of inner speech, generally connected with speech production and repetition. This path could move phonological codes from frontal to posterior brain locations of the left hemisphere.

1.2 Motivation

P300 (brain wave reaction to unexpected events), SSVEP (Steady-state visual evoked potentials), and motor imagery—thinking about physical actions—are the main foundations of current BCI systems. Although these methods have produced important progress in EEG-based BCIs, their practical relevance is clearly hampered. These approaches are often limited by poor response times, heavy user training requirements, and lack of long-term practicality. Researchers have started researching the speech-related paradigms—such as silent speech, imagined speech, and inner speech—that provide a more natural and straightforward way of engaging with gadgets in order to handle these problems.

A major obstacle in creating efficient inner speech decoding systems is obtaining accuracy and generalizability, particularly in situations where BCIs have to adjust to new users without much retraining. Many current techniques for inner speech decoding involve subject-dependent approaches, customizing models to match individual brain patterns. This method limits the scalability and utility of BCIs in more general, real-world environments even when it can produce reasonable accuracy in a controlled environment.

Given its ability to transform communication and control for people with extreme disabilities, a strong, subject-independent inner speech decoding system is especially needed. By means of developing models that enable BCIs to serve a larger population, therefore reducing the need for specific training and enabling more pragmatic uses. By means of analysis and comparison of several machine learning techniques meant to enhance subject-independent performance, this work intends to remove these constraints, opening the path for more efficient and user-friendly BCI systems.

1.3 Problem Background

The present research on inner speech decoding via EEG signals face several major complexities:

i. **Low Signal-to-Noise Ratio:** Separating the weak brain signals linked with inner speech from the ambient noise complicates internal speech decoding. Low signal-to-noise ratio (SNR) occurs from the intrinsically weak character of these signals combined with interference from muscle activity, ocular motions, ambient influences, and the limits of existing EEG technology. This low SNR causes inaccurate detection and extraction of core speech patterns, therefore impeding the development of consistent decoding models.

ii. **Individual Variability:** The inherent variations in how each person's brain produces these internal ideas create a major challenge in decoding inner speech. The brain processes connected with inner speech essentially differ greatly among individuals. Therefore, a model developed to identify the inner speech patterns of one individual

could not be useful for another. Inner speech generates such delicate brain signals, which makes it difficult to differentiate from the many other electrical processes recorded by EEG. Among the elements causing this difficulty are the naturally weak signals connected with inner speech, interference from many noise sources—including muscle and eye movements and ambient noise—and the intrinsic limits of EEG technology. Therefore, identifying the particular brain patterns connected to inner speech becomes quite difficult and slows down the creation of consistent decoding models capable of functioning among various people.

iii. **Limited Datasets:** The significant variability in brain patterns and cognitive processes linked with inner speech makes constructing inner speech decoding models that consistently perform across multiple individuals problematic. This variation is produced by differences in brain shape, cognitive style, linguistic experience, and other personal characteristics. Therefore, a model performing well on one person's data could not be able to efficiently decode inner speech from another person, so restricting the general use of inner speech BCIs.

Although several machine learning and deep learning techniques have been investigated for inner speech decoding, they usually find it difficult to reach both great accuracy and wide applicability among users. Most current research depends on subject-dependent methods, in which models are taught on data from particular users. Subject-dependent models frequently perform poorly and lack the capacity to generalize successfully to new users, therefore restricting their practical relevance even if they can occasionally attain modest accuracy inside controlled conditions.

Development of subject-independent techniques that attain both great accuracy and generalizability is desperately needed to solve these difficulties. Subject-independent methods would improve the feasibility and scalability of inner speech BCIs by allowing models to decipher inner speech across varied persons without considerable retraining, hence opening new paths for real-world uses.

1.4 Problem Statement

Current techniques for deciphering inner speech have significant limitations including their sensitivity to noise in EEG data, significant variation between individuals, and general lack of stability. These problems impede the development of dependable Brain-Computer Interface (BCI) systems that are easily flexible for use by a broad spectrum of people with communication or motor problems. Deep learning is a possibly strong solution to the difficulties of inner speech decoding since it has shown proven ability to extract complex patterns from raw data. Comparative analysis is required to evaluate whether deep learning models provide benefits over other machine learning approaches for subject-independent inner speech categorization, hence transcending the emphasis on deep learning architectures. The development of practical BCI systems depends on this kind of approach since it could minimize the need for laborious calibration for every user, so increasing the availability of BCIs to a greater population, particularly in cases of severe motor disabilities who might be unable to generate enough data for individualized model training. In an attempt to identify those with great accuracy and resilience in detecting inner speech from EEG data over a heterogeneous population, the present work examines numerous Machine learning models. This study will provide perceptive analysis to enable more readily available and practical BCI systems. This work intends to eventually impact the field of assistive technology by developing BCI systems that can empower more people without demanding large individual customizing.

1.5 Research Questions

1. *RQ1:* In a subject-independent setting, what machine learning and deep learning models and algorithms show the best accuracy in classifying inner speech from EEG signals?
2. *RQ2:* What are the main determinants of the generalizability across participants between several methodologies of inner speech decoding?

1.6 Aim of the Research

The main goal of this work is to investigate and compare many machine learning techniques for inner speech classification employing EEG data. This work aims to not only implement current machine learning methods but also objectively evaluate and compare a variety of machine learning approaches to identify which ones best fit to remove the difficulties in comprehending inner speech from EEG data. Researching a range of architectures, optimization techniques, and learning paradigms encompassed inside the framework of machine learning helps one to find approaches that are more accurate and resilient. By means of this comparison study, the research aims to not only pinpoint the most accurate models but also the most generalizable and strong tactics across a range of situations. This work suggests to create easily adaptable solutions using the non-invasive character of EEG in order to overcome some of the ethical and practical obstacles related with invasive techniques. The aim of this work is to identify different inner speech patterns by means of a range of mental commands manifested by different patterns of brain activity. Based on the results of the research, it is of great importance to reach subject-independent generalizability as both of which are absolutely required for the evolution of efficient BCI systems. To be sure that brain-computer interfaces (BCI) offer consistent communication and control interfaces, great accuracy is needed. Furthermore, it is essential to give subject-independent generalizability a priority if we are to enable these systems to be quickly adopted and successful over an extensive population. Though the quest of both high accuracy and subject-independent generalizability poses a difficult situation, the development of BCI technology that is both relevant and accessible is vital. This work seeks to offer answers to both of these issues. This work offers fresh approaches and insights on inner speech decoding as well as a road map for next BCIs, helping to expand brain-computer interface technology. The application of effective techniques for decoding the complexity of the human mind would not only help to increase BCI system performance but also lead to the development of more sensible and user-friendly interfaces. By improving the availability of brain-computer interfaces (BCIs), our effort helps to close the difference between technical innovations and the people most likely to gain from them. The focus on the design of user-friendly interfaces highlights the need of users being able to control devices by using natural mental instructions instead of needing a great degree of mental effort or training. Ultimately, this study helps to create more accessible and understandable communication and control interfaces for those with severe disabilities as well as for others.

1.7 Research Objectives

The following objectives will be pursued to achieve the research aim:

i. Review various machine learning models against a deep learning model, focusing on methods that could achieve subject-independent generalizability.
ii. Evaluate the performance of selected models using key metrics—such as accuracy, F1-score, precision, and recall—within a subject-independent framework to assess their generalization capabilities.

1.8 Scope of Research Work

Since it lowers the requirement for time-consuming calibration of individual users, this study is especially targeted on subject-independent classification of inner speech, a vital feature for the practical implementation of BCI systems. Unlike subject dependent models, which might perform well in controlled situations but badly on new users, this work intends to develop approaches that can efficiently generalize among different people. Selected as a publicly available dataset especially for inner speech research, the Thinking Out Loud (TOL) dataset offers a useful tool for validating the proposed models. The carefully specified experimental methods and availability of the dataset make it perfect for a comparison of several classification techniques. The choice of machine learning and deep learning models was predicated on their general applicability in BCI research and their capacity for subject-independence performance. This study contrasts the ability of these models to efficiently decode inner speech signals in a subject independent way and their generalizing capacity across subjects, therefore determining their relative strengths and shortcomings. A strong assessment of the models' capacity to extend to unseen data from new subjects is obtained by use of a leave-one-subject-out cross-valuation method. This thorough testing approach guarantees that performance criteria fairly represent the efficiency of these approaches in a real-world environment and are not distorted by subject-specific artifacts. This work mostly addresses model creation and evaluation; real-time BCI implementation is seen outside the purview of this work. This lets one concentrate on thorough offline study before entering more intricate real-time surroundings. Before advancing on to other datasets in the future, this stage of the research depends on a controlled environment for

comparing several models, which the restriction of one dataset offers. Beyond only evaluating model performance, this work seeks to offer understanding for next investigations on the possibilities of deep learning methods for subject-independent inner voice decoding. The study aims to further the field of inner voice BCIS and open the path for pragmatic uses of this technology by identifying exciting directions for next research.

1.9 Thesis Organization

Chapter 2: Focusing mostly on non-invasive approaches using electroencephalography (EEG), this chapter offers a thorough assessment of the body of current research related to inner speech decoding. The chapter will review the body of current studies on brain-computer interfaces (BCIs) and more especially, how inner speech is decoded using EEG. It will explore studies on imagined speech since these relate to inner speech and have similar underlying neurological mechanisms, so providing insightful analysis of decoding mental language processes. This evaluation will not only show the successes but also critically evaluate the limits of present methods and point up the research gaps this thesis seeks to solve.

Chapter 3: Methodology: The methodological approach followed to create and evaluate models for inner speech classification is described in this chapter. It starts by going through the data collection methods including specifics on the experimental paradigm employed to produce inner speech and the EEG recording configuration. The chapter next covers data preparation methods including signal filtering, noise reduction, and artifact removal therefore clarifying the methods for cleaning the raw EEG data and getting it ready for analysis. It explains the justification for selecting particular preprocessing methods as well as the anticipated effect on the performance of the model. Moreover, this chapter will address the machine learning model design applied in this work. It offers precise architectural details including algorithm choice for the models. At last, it addresses the training process and evaluation criteria for measuring model performance thereby guaranteeing a strong, repeatable, and unambiguous technique.

Chapter 4: Model Development: This chapter reviews the models meant for classification of inner speech. This addresses details on the layers, parameters, and activation mechanisms of every model. It also covers every model's training method including the loss

function, the optimization method, and strategies to raise the training quality. Moreover, what we covered in this chapter is the hyperparameter tuning method to offer every model the ideal configuration. It clarifies the ways of choosing the optimal model structure and the parameters. This all-encompassing method of creating models helps to clarify the process and promotes its repetition.

Chapter 5: Performance Evaluation: With an especially focus on their performance in subject-independent scenarios, this chapter presents and analyzes the outcomes of assessing the presented models. It starts by summarizing the performance measures applied to assess the models, including accuracy, precision, recall and F1-score. It then offers a thorough examination of the findings backed by striking tables and figures. Moreover, in the framework of inner speech categorization, the chapter comprises comparison studies of the several models applied in this study to grasp their advantages and shortcomings. The generalizability of the model across several subjects will be the main emphasis of this assessment as well as whether the method can be applied to categorize inner speech across many people. The relevance of the findings and consequences for the field of BCI and inner speech decoding will also be discussed together with any difficulties that happened throughout the experiment, with particular attention to their effect on model performance.

Chapter 6: Conclusion and Future Work: This chapter repeating the main contributions of the research to the field of inner speech decoding summarizes and contextualizes the significant conclusions of the thesis. It addresses the consequences of the findings, stressing their possible influence on the evolution of advanced BCI systems especially for those with speech problems. It then points out the limits of the study and talks on the need for more research in the field, suggesting particular areas where the present approach could be strengthened or improved. At last, it describes the possible future paths for this study, pointing up fresh directions to investigate and advice on next actions to advance the discipline of inner speech decoding. With the ultimate goal of creating strong, real-world applicable BCI systems for communication based on inner speech, these recommendations are meant to inspire additional research and growth in this field.

CHAPTER 2

LITERATURE REVIEW

2.1 Overview

For Brain-Computer Interventions (BCIs), inner speech, the silent internal discourse we go through in our brain, has great natural and intuitive communication and control power. Accurately deciphering inner speech from the brain does, however, provide great difficulty. Low signal-to-noise ratios in recordings resulting from the delicate and complicated character of inner speech make it challenging to separate pertinent brain activity from background noise. The tiny amplitude of the brain signals connected with inner speech and the other types of noise present in EEG recordings such as muscular activity and ambient interference combine to produce this low signal to noise ratio. Moreover, individual variation in cognitive processes and brain patterns gives still another level of difficulty to the decoding work. Many elements might be blamed for this diversity, including variations in brain architecture, cognitive styles, and language experiences, which greatly hamper the evolution of generalizable models. Complicating these difficulties is the lack of publicly available datasets created especially for study on inner speech. There are just now two such datasets, both of which are somewhat recent. Lack of data hinders the possibility to build strong and generalizable models, thereby hindering field developments.

With an emphasis on non-invasive methods employing electroencephalography (EEG), this literature review seeks to give an overview of the body of current study on inner speech decoding. The aim to make the technology more practical and accessible drives this focus on non-invasive approaches since it avoids the ethical and safety issues related with invasive techniques. The review will also cover pertinent studies on imagined speech, which shares similar neural mechanisms and can provide insightful analysis since inner speech datasets are

so rare. Particularly, imagined speech stimulates many of the same brain circuits as inner speech, so it is a useful arena for research providing complementing data and analysis. The review will start with looking at studies especially aimed at inner speech decoding with EEG. It will next look at research on imagined speech, stressing the relationships and its ramifications for developing inner voice decoding technologies. This method enables a more complete knowledge of the state of the art and aids in the identification of typical hazards requiring attention in next research projects. By means of this study, the review aims to pinpoint main obstacles, limits of current methods, and interesting future paths of research. This covers methodological problems with data collecting and evaluation in addition to technical ones related to inner speech.

2.2 Inner speech decoding

Inner speech decoding using EEG shows tremendous challenges because to the low signal-to-noise ratio, individual variation in brain pattern, and lack of publically available datasets. Notwithstanding these challenges, a number of studies have examined many approaches to classify inner speech using EEG data. Researchers in [5] investigated the implementation of compact convolutional neural network architecture, EEGNet, and obtained an average accuracy of 29% on the Thinking Out Loud dataset using a subject-dependent strategy. This result stresses the difficulty of correctly classifying inner speech even if it exceeds chance level skill. Another work [8] looked at using recurrent neural networks (RNNs), more notably Long Short-Term Memory (LSTM) and Bidirectional LSTM (BiLSTM) networks, for inner speech classification. Once more using a subject-dependent approach on the Thinking Out Loud dataset, they obtained accuracy of 30.4% with LSTM and 36.1% with BiLSTM. RNNs could help with inner speech decoding since they can find temporal links in EEG data. This implies that RNNs, with their ability to capture temporal dependencies in EEG data, may be beneficial for inner speech decoding. Furthermore, a study comparing different machine learning algorithms [7] found that a linear Support Vector Machine (SVM) achieved an average accuracy of 35% for subject-dependent approaches on the same dataset. These findings suggest that while traditional machine learning methods like SVMs can achieve reasonable accuracy.

Recently, new papers have emerged that utilize subject-independent approaches for EEG-based classification. Ng and Guan (2024) [18] made a noteworthy contribution with their meta-learning framework, which significantly increased subject-independent accuracy to 31.15%. In 2025, Radwan et al. (2025) [19] proposed another innovative method that used an Extra Tree-based approach and showed a comparable subject-independent accuracy of 32%.

Table 1: Comparison of different methods for inner speech decoding using EEG signals

Id,Year	Classes	Dataset	Classifier	Indep	Results (Accuracy)
[5],2021	4	TOL	EEGNet	No	29%
[6],2021	6	ISD	iSpeech(CNN)	Yes	35% (vowels), 29% (words)
[7],2022	4	TOL	SVM	No	35%
[8], 2022	4	TOL	BiLSTM	No	36.1%
[18], 2024	4	TOL	DeepConvNet	Yes	31.15%
[19],2025	4	TOL	BruteExtraTree	Yes	32%

TOL = Thinking out loud dataset. ISD = Imagined speech dataset. Indep = Subject-independent approach.

2.3 Imagine speech

A paper proposes a similar CNN-based iSpeech architecture and tests their model on imagined words and vowels. They observe that increasing the number of filters improves performance, but adding more layers does not help. They conduct extensive experiments and report an average accuracy of 35% for vowels and 29% for words classification for a subject-independent approach on ISD dataset. They also perform a one-tailed paired t-test and claim that transfer learning does not make a statistically significant difference [6]. Cooney et al. [9] applied transfer learning to classify imagined speech. Using a subject-dependent approach, they achieved 35% accuracy for vowel classification. In a later paper [10], they used a CNN model for imagined speech and obtained 24.46% accuracy with a subject-independent approach. They also demonstrated the importance of hyperparameter tuning for training CNNs for imagined speech. A paper applied a deep belief network for imagined speech classification and achieved 90% accuracy for consonant binary classification [11]. RF was also applied for imagined speech by [12]. They reported an accuracy of 18.5 for six word classification and 22.3 for vowel classification.

A paper proposed the Relevance Vector Machine (RVM) model. They tested the model on their own dataset and obtained 70% accuracy for three-class classification and 95% accuracy for binary classification using a subject-dependent approach [13]. The authors of [13] proposed SVM for imagined speech and tested their model using a subject-dependent approach. They reported an accuracy of 58.41 ± 11.45%. Recently, Agarwal et al. [15] proposed deep LSTM for imagined speech classification. They used their own dataset and trained their model on different frequency bands separately. They showed that the alpha band was able to recognize imagined speech better than other bands, followed by theta and delta bands. They reported an accuracy of 73% on five-class classification using a subject-independent approach.

Table 2: Comparison of different methods for imagined speech decoding using EEG signals

Id	Classes	Dataset	Classifier	Indep	Results
[9],2020	6	ISD	CNN	Yes	24%
[10],2019	6	ISD	CNN	No	35.68%
[12],2015	2	Other	DBN	Yes	90% (consonant)
[12],2017	5	ISD	RF	No	22.3% (vowels), 18.5% (words)
[13],2017	3	Other	RVM	No	95% (2), 70% (3)
[14],2016	5	Other	SVM	No	58.41 ± 11.45%
[15],2022	5	Other	LSTM	Yes	73%

2.4 Research Gap and Directions

The investigated deep learning models, CNNs and RNNs among others, show promise for classifying inner speech from EEG signals. Their capacity to understand intricate patterns from brain input clearly shows this promise; nonetheless, present uses have not completely exploited these features. For practical BCI uses, the stated accuracies—often ranging around 30–35%—for subject dependent are far from perfect. Such accuracy degrees limit the utility of inner speech BCIs in real-world situations since they are insufficient for practical uses requiring dependable and exact decoding. Furthermore, most research has concentrated on subject-dependent methods, therefore restricting the useful relevance of these models to new users. These approaches are challenging to scale for higher populations since each user depends on subject-dependent models and needs rigorous and time-consuming calibration. Thus, advancing inner speech decoding technology depends on the development of techniques that may attain better accuracy and generalize well among many individuals. The development of strong and

useful BCI systems that can be implemented in real-world environments depends first on this demand for both improved accuracy and generalizability among various users. In this thesis, we develop machine learning and deep learning models using a subject-independent approach.

2.5 Summary

The present state of research on inner speech decoding has been explored in this survey of the literature, therefore stressing both the possibilities and difficulties of this technology. While challenges such low signal-to-noise ratios and high inter-subject variability make inner speech decoding a difficult and challenging task, the ability to use inner speech as a natural and intuitive control mechanism for BCIs offers an alternative to traditional methods that may be less user friendly. This motivates us to explore inner speech further to enhance lives of countless people. Although several deep learning and conventional machine learning approaches have shown encouraging outcomes especially in controlled environments achieving great accuracy and generalizability remains a major challenge. Often lacking performance levels required for real-world applications, current BCI systems based on inner speech often fail for new users as well. Further complicating the decoding effort are the lack of publicly accessible information and the natural variability of EEG signals. Training models that can broadly generalize well are quite challenging given the small scale and scope of current datasets. Furthermore, complicating models' capacity to capture generalizable features across users is the great variability in EEG signals. Nonetheless, continuous research with an eye toward fresh deep learning architectures, robust feature extraction methods, and contrastive learning approaches provide exciting paths to increase the performance and practicality of inner speech decoding systems. These novel deep learning architectures, feature extraction, and learning strategies give some promise that the performance of inner speech decoding systems will enhance in the future and are thus a focus of research.

CHAPTER 3

METHODOLOGY

3.1 Overview

A major restriction of current inner speech decoding models, as the literature review emphasizes, is their limited generalizability over subjects. This restricts the applicability of these models greatly since models trained on data from one subject generally perform badly when applied to data from another. Since it would force new users to go through a very long and difficult model training or adaptation phase, this lack of generalization impedes the development of really user-friendly Brain-Computer Interface (BCIs). Moreover, this emphasizes a significant obstacle in the area to create more general models instead of subject dependent models. Especially, no previous study has concentrated on assessing these models with a strictly subject-independent method. Most research has either focused on data analysis inside a single subject or used subject-dependent approaches, therefore leaving a notable knowledge vacuum on how well these approaches operate in a more general environment. Given real-world BCI applications would demand models to perform effectively on new users with minimal prior data from these new users, this gap is very crucial to solve. For a long period, BCI research has been hampered by this lack of research in subject independent environments. This work fills up this void by means of a comparative study of several classification models especially inside a subject independent framework for inner speech decoding. This work intends to provide more robust and also more scalable approaches by concentrating on subject-independent techniques. This study intends to offer a thorough and exhaustive study of several models by applying a strict evaluation technique. This exacting method will enable a better knowledge of several models and hence be beneficial in developing new and more accurate BCI systems. Understanding the relative performance and fit of several models for generalization across individuals is more important than attaining state-of- the-art accuracy.

Although great accuracy is crucial, the primary goal of this work is to present a clear comparison of several models and identify which ones show the best possibilities for practical application. By use of this unambiguous comparison, BCI practitioners will be able to select the optimum model based on the requirements of their particular application. Deep learning methods and conventional machine learning approaches coexist among the models chosen for this study. This spectrum of models has been chosen to provide a more comprehensive evaluation of many approaches, each with possible advantages and drawbacks. This wide spectrum of several machine learning and deep learning models provides a thorough comparison of conventional methods with state-of-the-art deep learning methodologies.

Section 3.2 will go into great length on the dataset used for this study. This covers the data source, the features of the subjects, and the framework of the trials inside the dataset. It will also highlight the particular features of the dataset that fit for assessing subject independent approaches. Section 3.3 will go over the ideas and justification behind channel choice. The selection of a specific group of EEG channels and their significance in collecting pertinent brain activity for inner speech will be clarified in this section. It will also discuss additional channels that are less relevant for the categorization of inner speech and clarify why the selected channels are more so. Section 3.4 will explore triplet network architecture and contrastive learning technique development. This will clarify the special features of this network, including the way triplets are built and the training use of contrastive loss. The particular features of this network and the training procedure will be discussed in this part so that we have a quite clear knowledge of the model. At last, Section 3.5 will list the several classifiers together with the evaluation criteria applied to evaluate their performance especially in the subject-independent environment. This last part lists the models applied in this study together with the metrics utilized to give a fair comparison between the models. The particular measures being utilized will be the main emphasis of this part, which will also argue why these metrics fit for comparing models of inner speech decoding.

3.2 Dataset

The inner speech dataset Thinking Out Loud (TOL) [3] is the one utilized in this work. This dataset offers a special chance to investigate the brain activity linked with silently thinking

words, especially intended for inner speech research. It consists of a mean age of 34 ± 10 years, electroencephalography (EEG) recordings from 10 healthy participants—6 male and 4 female. Most importantly, none of these subjects had any prior knowledge of Brain Computer Interfaces (BCIs), so the dataset is perfect for analyzing the brain reactions of people fresh to such paradigms. Given that all participants spoke native Spanish, this helps to guarantee consistency in the language-related brain processes under investigation.

Three separate phases, known as sessions, comprised data collecting; each included several experimental runs under several conditions. Participants in each session were given four Spanish words—arriba (up), "abajo" (down), "derecha" (right), and "izquierda" (left). These terms were selected not only for their unambiguous spatial connotations but also as a collection of commands that might find use in actual BCI systems. The participants were directed to silently picture pronouncing these words and to use their inner voice without any obvious articulation. Beginning each session with a 15-second baseline recording, participants were guided to relax and reduce movement, therefore serving a benchmark for comparison with task-related activity.

Every session consisted in three main runs covering three important criteria: a visualized condition, inner speech, and pronounced speech. The participants in the pronounced speech condition really said the words out, which let the examination of overt speech production possible. The main focus of this study is the inner speech condition, which had the participants silently picture using the target words. Participants in the visualized condition were asked to picture the word visually, including the appropriate arrow pointing in a given direction. Within a session, the sequence of runs was always the same: one run of pronounced speech then two runs of inner speech then two runs of the imagined condition. Between runs came a one-minute respite. Every participant finished over 200 trials in the first and second sessions; but, the number of trials in the third session changed depending on participants' level of exhaustion and will.

The inner speech trials used a particular framework meant to produce the desired cognitive engagement. Originally showing at the middle of the screen, a white circle functioned as a cue for the beginning of the experiment and a fixation point. There was a 0.5 second presentation of this visual cue. Then a white triangle pointing in one of four directions—up,

down, right, or left—that matched the target phrase was shown for half a second. Participants were asked to silently visualize uttering the word again in their minds until the white circle turned blue just after the triangle vanished. Over this 2.5 second action period, the participants were supposed to do their designated assignment. The color shift of the circle marked the end of the task time; the participants were then advised to discontinue all task-related activities but to remain still with limited eye movement until the circle vanished to mark the conclusion of the trial. Eye blinks were to be controlled until the circle disappeared. The inter-trial rest interval ran from 1.5 to 2 seconds. A focus control system was used to guarantee participants' involvement and attention. Participants were prompted periodically to remember the direction of the last given signal for both visualized conditions and inner speech. They answered using keyboard arrows, and feedback was given following each answer to support task involvement and aid to reduce mistakes.

Specifically selected for this study was the TOL dataset since it directly relates to inner speech decoding and is the only publicly accessible dataset created especially for this kind of investigation utilizing electroencephalography. Together with the baseline, the controlled experimental design of the dataset offers a useful and complete tool for assessing the performance of inner speech decoding models. Furthermore included are displayed conditions and the marked speech. Furthermore, the design of the trials—with their exact timings and inclusion of the attention-monitoring task—makes it a viable dataset for researching the subtle variations in inner speech and its possible use in BCI applications, particularly among native Spanish speakers.

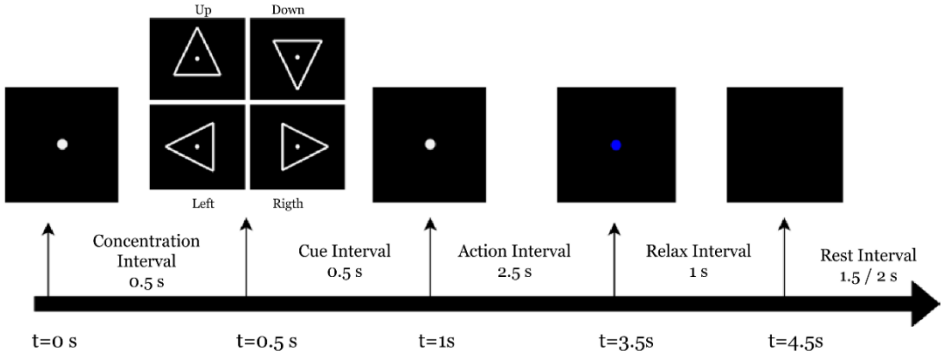

Figure 3.1 Thinking out loud dataset

3.3 Research Design and Development

Evaluating the generalizability of many machine learning and deep learning models for inner speech categorization inside a subject-independent framework forms the main focus of the research design. A fundamental need for real-world BCI applications, this concentration on generalizability assesses the capacity of models to work effectively on new, unseen subjects. Given that many earlier research has depended on subject-specific models which are not relevant in real-world environments, that makes exploring the independent approach more crucial. The Thinking Out Loud (TOL) dataset has complete recordings from 128 channels, hence it was the source of the EEG data for this work. This dataset provides a useful, publicly available source for inner speech research, which helps us to evaluate our findings in line with others. This comprehensive and sophisticated dataset is ideal for doing research in this field since it features recordings of many inner speech circumstances. We chose a group of channels especially connected to language processing in the left hemisphere since not all channels equally support internal speech decoding. Existing research underlined in this choice that neurological correlates of language and speech are more strongly confined to parts of the left hemisphere. Furthermore, this channel choice lowers the number of input features, so lowering the computational cost as well. Based on earlier studies showing their role in language and inner speech processing, this channel choice concentrated on EEG channels identified with "D". These channels, typically located on the left hemisphere, are situated above areas of the brain known to be important for language processing, including regions such as Broca's and Wernicke's areas; this set lets us concentrate on particular domains most likely to be important for inner speech decoding.

A bandpass filter was used to separate the alpha frequency band (8–12 Hz), which has been connected to cognitive activity including inner speech, therefore improving the data quality. Studies demonstrating that the alpha frequency band is more firmly linked with cognitive processes more especially, with regard to inner speech processing and language-related activity led to the choice of the band. Less susceptibility of the alpha band to noise than other bands will also aid with model performance. This filtering concentrated the study on the most pertinent signal components and helped to quiet noise. We aim to reduce noise and highlight brain activity especially related to inner speech by restricting the data to a given range of frequencies, therefore enabling the models to more precisely detect relevant patterns.

Statistical features (mean, variance, skewness, kurtosis etc.) were obtained from the filtered signals for conventional machine learning models. These statistical qualities offer a complete picture of the time-domain attributes of the signal and can draw attention to significant traits the models can subsequently utilize to set apart different inner speech circumstances.

3.3.1 Data Preprocessing

Because of their great dimensionality and low signal-to- noise ratio, EEG signals must be handled by means of data preparation. The preprocessing actions carried out in this study build on the first preprocessing done by the authors of the Thinking Out Loud (TOL) dataset [3], which comprise essential measures to improve the signal quality and ready it for analysis. These phases, together with the additional preprocessing done in this work, comprise channel selection, frequency filtering, statistical feature extraction, and other pertinent processes described below.

3.3.1 Initial Preprocessing by the TOL Dataset Authors

The data examined in this work derived from the preparation pipeline created by the TOL dataset creators. This pipeline consisted in several important phases:

i. *Data Loading:* Stored as.bdf files, the raw data was imported containing continuous EEG, external electrode signals, and related event markers.
ii. *Event Verification and Adjustment*: The raw data underwent a review for correct event tagging. The authors found missing tags and suggested a fix to guarantee exact event markers in the data and a whole flow of events.
iii. *Re-referencing:* Re-referencing the data to channels EXG1 and EXG2, respectively, placed on the left and right earlobes, respectively. This method minimized common-mode voltage and line noise by first generating a virtual channel from the average of the two EXG channels then subtracting from all other channels.

iv. *Digital Filtering*: The data was fed a zero-phase bandpass finite impulse response (FIR) filter. The lower and upper cut frequencies were set to 0.5 and 100 Hz, respectively; a notch filter at 50 Hz was used to eliminate line noise.

v. *Epoching and Downsampling:* A factor of four down sampled the continuous data to provide a final sampling rate of 256 Hz. The data were then split into 4.5-second epochs matching every trial, running from the start of the concentration period to the end of the relaxation period.

vi. *Independent Component Analysis (ICA*): ICA helped to reduce artifacts on the EEG channels. Using correlation with the exterior (EXG) channels, the scientists found and cut out sources related to blinks, eye movements, and mouth motions. The final dataset was produced by rebuilding the data after artifact elimination.

vii. *Electromyography (EMG) Monitoring*: ICA worked to clear artifacts from the EEG channels. The researchers identified and removed sources connected to blinks, eye movements, and mouth motions by means of correlation with the outside (EXG) channels. Rebuilding the data following artifact removal generated the final dataset.

viii. *Ad-hoc Corrections:* Based on participant comments, the writers additionally made ad hoc corrections— for example one person's condition markers were changed.

3.3.2 Further Preprocessing for this Research

Following the TOL dataset authors' initial preparation, the following actions were taken especially for this study:

3.3.1.1 Channel Selection

Analysis focused on EEG channels mostly derived from the left hemisphere, especially those beginning with the label "D. The selected channels are:

[D5, D6, D7, D8, D9, D10, D11, D12, D13, D14, D15, D16, D17, D18, D19, D20,

D21, D22, D23, D24, D25, D26, D27, D28, D29, D30, D31, D32]

Based on their claimed participation in language processing more especially, inner speech these channels were selected in line with past studies [16].

3.3.1.2 Frequency Filtering

The alpha frequency band (8–12 Hz) was isolated using a bandpass filter; this band is well-known to be connected with cognitive processes including inner speech [15]. Lower and upper frequency cutoffs at 8 Hz and 12 Hz respectively were used in filtering. This frequency spectrum was chosen to reduce pointless background noise and improve pertinent signal components.

3.3.1.3 Feature Extraction

A complete collection of statistical features was obtained from the filtered EEG signals for conventional machine learning models. Computed for every channel separately, these features behave as condensed depictions of the EEG signal properties. We extracted the following:

Basic Statistical Measures

i. *Mean:* (central tendency) The mean of an EEG signal is its average amplitude across a specific time interval. It is sensitive to variations in total signal power and offers a baseline level estimate of the signal.
ii. *Standard deviation:* (signal variability) The standard deviation gauges the dispersion or spread of the EEG signal about its mean. It shows the degree of variation in the signal and represents its frequency of fluctuations.
iii. *Variance:* (spread of the signal) The square of the standard deviation is variance. It likewise gauges signal dispersion, but unlike standard deviation it is more sensitive to outliers. Variance helps one measure data signal variability.

iv. *Minimum and maximum values*: These characteristics record within a time window the lowest and maximum amplitude values. They help to identify excessive signal changes or artifacts and give knowledge of the signal limits inside every window.

v. *Range:* (peak-to-peak amplitude) The range of the EEG signal is its difference from maximum to least value. This provides the signal's overall amplitude fluctuation over a given time interval.

Signal Shape Descriptors

i. *Skewness*: (measure of signal asymmetry) Skewness estimates the signal distribution's asymmetry around its mean. A positive skew shows a longer tail towards higher amplitudes; a negative skew shows a longer tail towards lower amplitudes. One can evaluate signal asymmetry by means of skewness.

ii. *Kurtosis:* (measure of signal peakedness) Kurtosis gauges the signal distribution's "peakedness." Reduced kurtosis denotes a flatter distribution; higher kurtosis predicts a sharp peak and heavy tails. Kurtosis aids in signal change identification in the tail and peak areas.

iii. *Root Mean Square (RMS)*: (signal magnitude) RMS is the mean's square root of the squared signal values. Reflecting both its average and fluctuation, it is a gauge of the total signal magnitude and hence of signal power.

Temporal Features

i. *Mean absolute difference:* (signal roughness) In a time series, the mean absolute difference—the average absolute difference between successive signal values—is It offers a gauging of the signal's roughness or instantaneous variations across time.

ii. *Indices of minimum and maximum values:* (temporal localization of extrema) These characteristics give the time indices within a certain period where the minimum and highest signal values occur. This metric can show particular moments in the EEG signal when significant events or changes occur.

Wavelet Features

The Discrete Wavelet Transform (DWT) is a mathematical technique that analyzes the EEG signal in both the time and frequency domains simultaneously. Unlike traditional methods, it decomposes the signal into different frequency bands at various resolutions, which is ideal for capturing the transient, non-stationary characteristics of brain signals. The DWT provides a detailed view of how signal properties change over time, capturing features that are often missed by other methods.

For each channel and at each level of decomposition, we compute the following statistical features from the wavelet coefficients:

i. Mean, Standard Deviation, and Variance: These measure the central tendency, spread, and overall energy of the signal within a specific frequency band. They provide insights into the power and variability of different EEG rhythms (e.g., Alpha, Beta).
ii. Minimum, Maximum, and Range: These values identify the most extreme amplitudes within a frequency band, helping to detect significant, brief events or spikes in the signal.
iii. Mean Absolute Value: This acts as a robust measure of the average energy of the signal coefficients, which is useful for quantifying the overall magnitude of the signal within a given frequency range.
iv. Skewness and Kurtosis: These describe the shape of the signal's distribution within each frequency band. Skewness shows if the signal's power is concentrated towards higher or lower amplitudes, while kurtosis indicates the presence of sharp, "peaky" events.
v. Root Mean Square (RMS): This is a key measure of the overall signal magnitude or power within a specific frequency band. It is a more robust indicator of signal strength than the simple mean.

By extracting these features from the wavelet coefficients, we create a rich, multi-scale representation of the EEG data that captures subtle, time-localized patterns crucial for accurate inner speech classification.

These properties were computed over all EEG channels to produce a rich feature vector spanning both temporal dynamics and amplitude traits of the signals. These characteristics

taken together offer a multi-dimensional representation of the EEG data, which lets the machine learning models discover discriminative patterns for inner speech classification. Adding other features, such as wavelets, could increase the feature dimension and potentially decrease model accuracy, making it more likely to overfit to noise.

3.3.2 Models For comparison

Four models were selected under a subject-independent framework to evaluate their decoding of inner speech from EEG signals.

3.3.2.1 Linear SVM with Statistical Features

Inner speech categorization started from a baseline model, a Support Vector Machine (SVM) with a linear kernel. Aimed to identify an ideal decision boundary for classification problems, SVMs are a kind of supervised learning method. Regarding a linear SVM, the decision boundary is a hyperplane. Statistical features are input for this model. Extracted from the preprocessed EEG data, these characteristics form the input for the SVM and help the model to identify several inner speech states. The efficiency of a simple, interpretable method in separating inner speech categories among participants was assessed using a linear SVM model. Starting with a linear SVM offers a clear, understandable approach for comparison with more intricate models. Since it lets one better grasp the extra advantages of more complicated models, its simplicity makes it an excellent benchmark for more advanced methods. Relying on statistical features collected from EEG data to describe the signal patterns connected with inner speech, the linear kernel helps the model to classify data by determining the ideal hyperplane separating the classes in a high-dimensional space. Though its name suggests otherwise, the linear kernel will not be useful if the classes cannot be separated in the input space by a straight line (in 2D) or a plane (in higher dimensions). Generally speaking, the SVM will next look for the best hyperplane in a higher-dimensional feature space where the data might be linearly separable. Particularly the linear kernel only performs as expected when an initial input space allows a linear separation. The SVM seeks to identify the plane with the biggest margin,

therefore the greatest distance from the closest point to the decision plane. Whether basic linear bounds are sufficient for subject-independent generalization in inner speech categorization is judged against this baseline. Building alternative models starts from the fact that a linear SVM can reach decent performance for subject-independent categorization. Furthermore, this can imply that subject-independent interior speech decoding might not call for highly sophisticated models.

3.3.2.2 SVM with Different Kernels

We investigated whether a nonlinear decision boundary might improve model performance using an SVM with several kernels. Although a linear kernel performs well for linearly separable data, it might not be sufficient in more complicated relationships between classes. For these non-linear interactions, non-linear kernels can perhaps find improved decision bounds. Apart from the linear kernel, many nonlinear kernels—such as radial basis function (RBF), polyn (poly), and sigmoid—were tested to evaluate their capacity to represent the more complicated, maybe nonlinear interactions inside EEG data. These kernels are chosen since their various qualities could be useful in the classification of EEG signals. This is so since simple linear decision limits cannot readily separate brain activity in EEG signals. Non-linear interactions in these signals are rather widespread. These nonlinear kernels let the SVM convert the data into a higher-dimensional feature space in which separating classes according to intricate patterns could be simpler. This mapping to a higher-dimensional space is implicit and accomplished by means of a "kernel trick," therefore enabling us to employ higher dimensional mappings without explicitly computing them. This is advantageous computationally since other computing costs would be much higher. Modeling complex, non-linear interactions in the data calls for the RBF kernel. Datasets marked by polynomial relationships between variables suited for the polynomial kernel. Sometimes the sigmoid kernel is used since it quite resembles the activation function of a neural network.

3.3.2.3 LightGBM

Using its advanced tree-based ensemble learning powers, LightGBM was applied as a gradient boosting framework for inner speech classification. It use decision trees, gradient boosting is a machine learning technique whereby an ensemble of models is created whereby each next model is taught to correct the mistakes committed by its predecessors. This model was selected especially for its capacity to keep computational efficiency while managing the high-dimensional, noisy character of EEG inputs. This qualifies for EEG signal processing where a fast-running time is advantageous considering the high dimensionality and intrinsic noise commonly found in EEG signals. The framework uses a chain of decision trees, with each next tree concentrated on fixing the prediction mistakes of its predecessor, hence generating a strong ensemble model. Gradient boosting is based mostly on this sequential method. Every new tree aims to fix the mistakes of the past to enhance general performance.

Two key technical innovations of LightGBM make it particularly suitable for EEG-based inner speech classification:

Leaf-wise tree growth strategy, This, in the statistical data obtained from EEG signals, helps to find patterns more quickly. Particularly with high-dimensional data, LightGBM produces trees by prioritizing the leaf split that results in the biggest reduction in the loss function, unlike conventional approaches that grow trees level by level, therefore improving the learning process.

Histogram-based splitting mechanism, therefore, when processing high-dimensional EEG data, greatly lowers computing cost and memory use. This method approximates splits during training by grouping numerical feature values into bins, therefore enabling significantly quicker and more memory efficient training.

Although LightGBM's sophisticated tree structure might enable it to detect more complicated patterns than linear models, its gradient boosting framework features built-in regularization algorithms to assist reduce overfitting to subject-specific noise. Particularly crucial when working with noisy data, like in the case of EEG recordings, regularization is a method used to guarantee that intricate models do not overfit the training data and generalize well on new and unexplored data. This balance between model complexity and regularization makes it especially pertinent for subject-independent classification problems, in which the

objective is to find consistent inner speech patterns among many people. Real-world BCI applications depend on the models being employed across fresh subjects, hence this regularity helps them to be utilized.

Using LightGBM's ability to effectively manage different feature distributions and scales, training of the model made use of the retrieved statistical characteristics from the preprocessed EEGs. Given that EEG features vary widely, this effective handling of various feature distributions aids with noisy data. This method offers a good framework for strong cross-subject inner speech classification by combining the interpretability of conventional statistical features with the strong pattern recognition capacity of gradient boosting. Therefore, we obtain a computationally effective and strong approach to classify interior speech over several individuals by aggregating gradient boosting with feature extraction techniques.

3.3.2.4 CNN with Triplet Loss (Triplet Network)

Leveraging its capacity to automatically extract hierarchical spatiotemporal features from EEG signals, a convolutional neural network (CNN) architecture was applied as a deep learning approach for inner speech classification. CNNs often struggle to detect subtle class differences, especially in EEG data where high individual variability and low signal-to-noise ratios complicate feature extraction. Although their convolutional layers excel in pattern recognition.

We improved CNN architecture using a triplet network structure to get beyond these constraints. This method runs three parallel instances of the same CNN to process triplets of input samples: an anchor sample, a positive sample (same class as anchor), and a negative sample (different class). Triplet loss defined as is used for network training.

$$L = \sum_{i=1}^{N} \quad \left(0, \left|\left|f(x_i^a) - f\left(x_i^p\right)\right|\right|^2 - \left|\left|f(x_i^a) - f(x_i^n)\right|\right|^2 + \alpha\right) \quad (3.1)$$

f(x) is the CNN embedding function; alphaα is a margin value deciding the necessary separation between positive and negative pairings. By reducing the distance between samples of the same class and hence optimizing the embedding space, this loss function maximizes the distance between samples of different classes. Through relative distances instead of absolute feature values, the triplet network emphasizes learning class-discriminative patterns that generalize well across subjects.

Emphasizing relational learning, the triplet network solves two important problems in EEG-based inner speech classification. First of all, it improves subject independence by concentrating on the links between samples, therefore enabling the network to capture generalizable properties not unique to any one person. Second, by focusing on the most pertinent discriminative features and therefore lowering the effect of subject-specific fluctuations, it maximizes the embedding space for class separation. Consequently, the embedding space of the network gets more strong, allowing one to keep good classification over several subjects.

Over traditional CNNs, this architecture has significant benefits. The relationship-based learning method of the triplet network not only increases generalization over subjects but also strengthens resistance to individual variances in EEG patterns. Moreover, the organized embedding space offers more flexibility, therefore facilitating the inclusion of new classes or subjects without considerable retraining. Focusing on relative distances helps the triplet network also better manage the high noise levels inherent in EEG data, therefore supporting more accurate inner speech classification. All things considered, the resultant embedding space proves flexible for a variety of BCI uses and offers a strong basis for subject-independent inner speech classification, hence separating different mental states. The capacity of the model to learn generalizable features makes it especially appropriate for real-world situations, when consistent cross-subject performance is crucial.

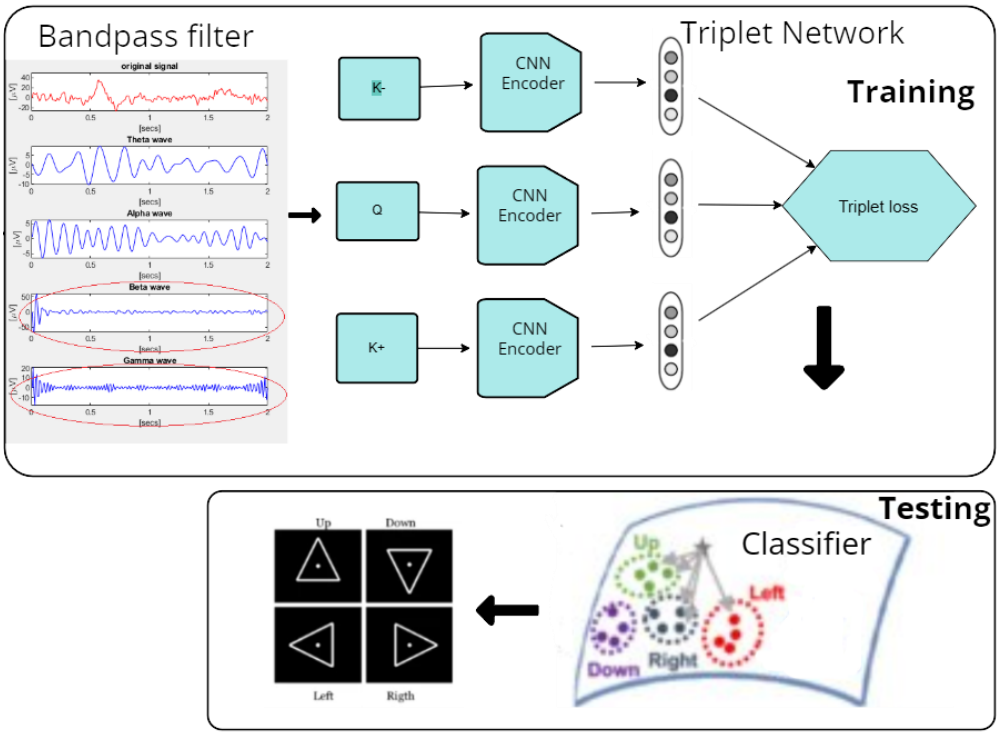

Figure 3.2 Overview of the proposed EEG signal classification model.

3.4 Evaluation Metrics and Experimental Setup

Every model's performance was evaluated by means of accuracy, F1-score, precision, and recall. These measures offer a whole picture of the models' classification performance by considering both the general accuracy and the balance of accurate classifications among categories.

3.4.1 Subject-Independent Evaluation

A subject independent evaluation system was used to objectively assess the generalizability of every model across people. This method shows that the models are not only picking subject-specific patterns but rather are catching consistent and representative traits among many distinct people. In this protocol, data from one subject is held out exclusively for testing, while data from all other subjects is used for training. This procedure is repeated for

each subject in a leave-one-subject-out cross-validation framework, providing a comprehensive measure of the model's ability to generalize to new subjects.

The great inter-subject heterogeneity in EEG signals makes the leave-one-subject-out approach especially appropriate for EEG-based inner speech classification. Each model is tested on entirely unprocessed data by training on all but one topic, therefore imitating a genuine situation whereby the model would have to extend to new users. This assessment system offers a precise evaluation of every model's resilience and emphasizes its advantages and drawbacks in managing the intricate and subtle changes in inner speech EEG signals among individuals.

To quantify performance, several key evaluation metrics were employed:

i. *Accuracy:* Measures the overall correctness of classifications, providing a straightforward assessment of each model's effectiveness. However, accuracy alone may not be enough if classes are not balanced.
ii. *F1-Score*: Combining recall with accuracy into a single figure this statistic offers a fair evaluation of performance. It is well appropriate for examining EEG data, where datasets could be skewed or classification limits may be subtle. The F1-score highlights the precision-recall balance of the model, which is vital for handling EEG inputs possibly contaminated with noise or artifacts.
iii. *Precision*: This statistic shows among all the cases the model finds that belong to a given class the proportion of correctly classified ones. When reducing false positives is a top goal in EEG classification, precision is very crucial.
iv. *Recall:* This statistic compares the proportion of accurately found instances of a class to all the actual instances of that class. Even with some noise present, high recall is necessary to guarantee that the model detects as many actual cases of inner speech as feasible.

These measures offer a multi-dimensional assessment of every model, therefore enabling a sophisticated comparison across subject-independent conditions. Analyzing the trade-offs between precision, recall, and F1-score helps the research to find models that provide consistent and dependable performance in the demanding setting of cross-subject inner speech decoding.

3.4.2 Experimental Setup

This work used a combination of Python's powerful machine learning and deep learning tools in order to implement and assess every model. Python was selected for research and development mostly because of its simplicity and broad ecosystem of scientific and machine learning libraries, which make it a usually used tool. Important instruments and sources included:

i. *scikit-learn*: A cornerstone library for machine learning tasks in Python, scikit-learn was utilized for implementing traditional machine learning models (e.g., SVM, LightGBM) and performing tasks such as cross-validation and grid search. It offers a complete collection of methods for dimensionality reduction, preprocessing, classification, regression, clustering, and model selection. scikit-learn's broad functionality facilitated efficient model training, evaluation, and comparison. Its user-friendly interface makes it easy to train, test, and compare various machine learning models, which facilitated the comparison of traditional machine learning methods.

ii. *PyTorch*: This library is used extensively for deep learning research and development and has great flexibility. Employed for the deep learning model (CNN with triplet loss) due to its flexibility and extensive support for custom architectures. Unlike many other libraries, Pytorch makes it easy to define custom models with various loss functions, and has the ability to use dynamic computational graphs. Triplet network structure implementation was made possible by PyTorch, together with GPU acceleration to maximize training times. Pytorch's adaptability and processing capacity fit for developing and training deep learning models with intricate architectures and loss functions. Models can thus be taught faster as well.

iii. *Optuna*: Applied for hyperparameter tuning across all models, Optuna's capacity to do automated, effective optimization searches enabled the identification of the optimal parameters for every model, hence improving their subject-independent performance. Designed to quickly optimize even vast search areas of hyperparameters, Optuna's optimization algorithm makes the process of hyperparameter optimization fast, efficient, and simple to use.

iv. *MNE-Python*: Designed especially for neurophysiological data including EEG and MEG data, this open source package MNE was first utilized for data preparation including channel selection and filtering in an open-source toolkit for processing EEG

data. It provides various tools unique for EEG processing that simplify processing. MNE's customized tools provide preprocessing capabilities catered especially for neuroimaging data and helped simplify EEG data management. MNE's pre-processing tools are particularly meant for neurophysiological data and simplify analysis on this kind of data.

3.5 Summary

Emphasizing cross-subject generalizability, this chapter addressed the method for a comparative analysis of models aiming at reaching subject-independent inner speech classification using EEG signals. The method covered in this chapter provides a rigorous framework to assess multiple classification models and their ability to generalize to new and unseen individuals. Preprocessing steps included selecting left-hemisphere EEG channels, bandpass filtering the alpha frequency band, and compiling statistical data. The choice of channel was motivated by the established language processing role of the left hemisphere. Bandpass filtering helped the study to focus the alpha band. These methods enable to reduce noise and assist in feature extraction and categorization. Four models, Linear SVM, SVM with RBF kernel, LightGBM, and CNN with triplet loss were investigated using a leave-one-subject-out cross-valuation technique, which tested each model's performance over unseen individuals. Most importantly for practical uses, this testing method was chosen particularly to assess each model's ability to detect EEG signals across individuals and provides a way to measure how successfully the trained model may be transferred to a new person. Using accuracy, F1-score, precision, recall, and other performance indicators, every model's performance in identifying inner speech was completely assessed. These values provide a full picture of the model; accuracy indicates the general efficacy of the model; F1-score is a harmonic mean of precision and recall; and precision and recall disclose the balance between positive and negative classes. This approach reveals the relative advantages and limitations of any strategy for BCI applications since it offers a robust platform for assessing and comparing model performance in upcoming chapters. By precisely applying the method covered in this chapter, this study prepares the stage for deeper investigation of every model. Furthermore, these results can offer perceptive study of the trade-offs inherent in every approach and which would be better suitable for pragmatic BCI applications.

CHAPTER 4

MODEL DEVELOPMENT

4.1 Overview

The evolution and use of the models assessed for inner speech classification inside a subject-independent framework is discussed in this chapter. Since it explores in great detail how the models were constructed, applied, and trained for this research, this chapter is an essential component of the thesis. With both conventional machine learning techniques and a deep learning model included in the comparison, every model was selected for their capacity to generalize across individuals. Particularly in subject independent settings, the models in this study are meant to evaluate the variations in performance among several kinds of models. Starting with an examination of the architecture and training method for every model including the Linear SVM, SVM with RBF kernel, LightGBM, and CNN with triplet loss in this chapter. While the training procedure is crucial for how these models will learn, the architecture details how each model is built is also important. This work mostly uses CNN with triplet loss built in PyTorch as the main deep learning method. PyTorch is applied since it offers the adaptability to define and train the deep learning models. The CNN with triplet loss is selected since it offers a better way to learn features across many participants and it is hypothesised to be able to better capture non-linear correlations in EEG signals. Furthermore, hyperparameter tuning methods are applied for performance optimization of every model, including Optuna. Model development depends on hyperparameter tweaking since the outcomes of a model can be much influenced by different model parameters.

4.2 Model Architectures

Different models used for comparison in explained below.

4.2.1 Linear SVM

Linear Support Vector Machines (SVMs) offer a powerful approach to binary classification by identifying an optimal hyperplane that effectively segregates data points in a high-dimensional feature space. The fundamental principle behind linear SVMs lies in their ability to maximize the margin between classes while maintaining accurate classification.

In the basic linear SVM setup, each data point $x_i \in R^d$ is represented as a d-dimensional feature vector, with corresponding labels $y_i \in \{-1, +1\}$. The SVM aims to discover a hyperplane defined by the equation:

$$w^T x + b = 0 \tag{4.1}$$

where w is the normal vector to the hyperplane and *b* is the bias term. The classification decision is made based on the sign of the linear function:

$$f(x) = w^T x + b \tag{4.2}$$

To find this optimal hyperplane, the following optimization problem is solved:

$$\frac{1}{2} |w|^2 \tag{4.3}$$

subject to:

$$y_i(w^T x_i + b) \geq 1, \quad \forall i \tag{4.4}$$

This formulation ensures maximum margin separation between classes while correctly classifying the training data. In practice, to handle non-perfectly separable data, the soft-margin SVM introduces slack variables ξ_i and a regularization parameter C:

$$i\,mize_{w,b,\xi}(1/2)\left||w|\right|^2 + C * \Sigma_{i=1}^{n}\xi_i \tag{4.5}$$

subject to:

$$y_i(w^T x_i + b) \geq 1 - \xi_i, \quad \xi_i \geq 0, \quad \forall i \tag{4.6}$$

The statistical features used as input to linear SVMs typically undergo preprocessing steps such as:

i. *Standardization*: Scaling features to zero mean and unit variance
ii. *Normalization*: Scaling feature vectors to unit norm

Linear SVMs excel particularly in high-dimensional spaces where data is linearly separable or nearly linearly separable. This property has made them especially valuable in text classification, bioinformatics, and image recognition tasks where the input space is naturally high-dimensional. The generalization error bound for SVMs depends on the margin rather than the dimensionality of the feature space, making them particularly resistant to overfitting in high-dimensional scenarios.

The computational efficiency of linear SVMs, coupled with their strong theoretical foundations in statistical learning theory, has led to their widespread adoption in various applications. Their performance is particularly noteworthy when:

i. The feature space dimensionality is high relative to the number of training samples
ii. The classes exhibit approximately linear separation in the feature space
iii. Robust generalization is required for unseen data

Modern implementations of linear SVMs often employ efficient optimization techniques such as sequential minimal optimization (SMO) or coordinate descent methods, making them practical for large-scale learning tasks while maintaining their theoretical guarantees of finding the globally optimal solution.

4.2.2 SVM with RBF Kernel

The RBF kernel expands the capabilities of SVMs by implicitly mapping input data into a higher-dimensional, and theoretically infinite, feature space. The RBF kernel between two points x and x' is defined as:

$$K(x, x') = exp\ exp\ \left(-\gamma \left|\left|x - x'\right|\right|^2\right) \tag{4.7}$$

Where γ is a parameter that controls the width of the Gaussian function. Through this mapping, the SVM can effectively learn nonlinear decision boundaries in the original input space, with the kernel trick ensuring computational feasibility.

Because brain activity patterns are naturally nonlinear, the RBF kernel should be more helpful for classification of EEG signals. Nonlinear decision limits help to better capture the complicated temporal and spatial correlations that can show in EEG signals. Particularly when handling motor imagery and event related potential classification tasks.

The effectiveness of RBF-SVMs in EEG classification stems from their ability to:

i. Model complex, non-linear patterns within the EEG features.
ii. Deal with the inherent inconsistency and variability common in EEG data.
iii. Perform well and generalize to new data even with small EEG training datasets.

Besides the RBF kernel, the study also explored other non-linear kernel options.

i. *Polynomial Kernel*: The polynomial kernel is defined as K$K(x, x') = (x^T x' + c)^d$, where c is a constant and d is the degree of the polynomial. This kernel enables the support vector machine to learn polynomial form non-linear relationships. An essential hyperparameter that must be maximized is the degree, d, of the polynomial. The linear kernel replaces the polynomial kernel when d= 1.

ii. *Sigmoid Kernel*: The sigmoid kernel is defined as K(x, x') = tanh(αxTx' + c), where α is a scaling factor and c is a constant offset. This kernel can also be interpreted as a two-layer neural network. However, its performance varies a lot and is highly dependent on data distribution.

With their particular mathematical formulations, each of these kernels enables the SVM to map the data to several high dimensional spaces where classes may be linearly separated in that new space. Computed efficiently, this is a substitute for explicitly determining a non-linear decision boundary.

4.2.3 LightGBM

LightGBM, introduced by Microsoft Research (Ke et al., 2017) [17], is a high-performance gradient boosting framework specifically designed to address the computational and memory limitations of traditional gradient boosting methods, particularly when dealing with large, high-dimensional datasets [17]. At its core, LightGBM integrates two novel techniques: "Gradient-based One-Side Sampling" (GOSS) and "Exclusive Feature Bundling" (EFB), which work synergistically to drastically reduce training times while maintaining, and often improving, model accuracy.

LightGBM operates as an ensemble method, constructing decision trees sequentially. Each tree in the sequence attempts to correct the errors of its predecessors by fitting the negative gradients (also known as residuals) of the loss function. Typical of gradient boosting, this iterative method lets a powerful predictive model be developed. LightGBM is unique in that it uses a leaf-wise, best-first tree development method. LightGBM expands the tree by splitting the leaf node that produces the maximum information gain, unlike conventional level-wise development when all nodes at a particular level are split before advancing to the next level.

This method produces asymmetric trees that can efficiently capture complicated patterns and nonlinearities in the data, hence maybe producing a more accurate and efficient model. It also lets LightGBM create deeper trees without raising memory consumption as much as in a level-wise method.

By selecting data instances depending on their gradients, the Gradient-based One-Side picking (GOSS) method maximizes the training process even more. Data examples with higher gradients, that is, under-trained instances have a more important influence on the computation of information gain, as Ke et al. (2017) underline [17]. GOSS solves this by maintaining all data instances with higher gradients and doing random sampling on instances with modest gradients, hence lowering the number of data examples needed for training without appreciably compromising the accuracy of information gain estimate. GOSS uses a constant multiplier to these sampled examples when computing the information gain in order to minimize the bias produced by the smaller gradient reduced sampling of instances. This ensures that these instances still have influence on the model, preserving the original data distribution to a large extent.

Apart from GOSS, LightGBM uses the Exclusive Feature Bundling (EFB) method, which bundles mutually exclusive features into single composite features thereby addressing the problem of high feature dimensionality. Ke et al. (2017) [17] claim that many of the elements used in practical applications are sparse that is, they hardly simultaneously take non-zero values. EFB essentially lowers the dimensionality of the data without appreciable information loss by integrating these unique properties. This is done by building the bundles such that features can live in several bins inside the bundle. For instance, a unified feature bundle can be produced whereby the values of each original feature remain identifiable by adding an offset to the original values of the features. The NP-hardness of optimally dividing features into the least number of bundles is emphasized in this work together with a greedy method to generate the feature bundles. LightGBM is especially suited for handling sparse datasets, including those typically encountered in EEG data, since this decrease in dimensionality results in a notable decrease in memory consumption.

LightGBM is well fit for EEG classification applications where high-dimensional and often sparse data is the norm thanks to these efficiency gains. Ke et al. (2017) show that

LightGBM has great advantages; it uses less memory because of its optimal data management and use of GOSS and EFB, and it provides notably faster training speeds than conventional GBDT approaches. Moreover, it efficiently uses the natural sparsity of EEG characteristics by means of EFB, therefore minimizing needless processing. Additionally supported by LightGBM are parallel and GPU computation, offering a scalable and memory-efficient approach fit for subject-independent inner speech categorization. LightGBM particularly fits high-dimensional, large-scale datasets like the EEG recordings utilized in this work since the tree-learning and split-finding procedure is also quite highly optimized.

4.2.4 CNN with Triplet Loss

Building upon the foundational concepts introduced in Section 3.3.2.4, this section details our implementation of a CNN-based triplet network for EEG signal classification. The triplet network was chosen as a deep learning approach to help classify inner speech as it is particularly good at learning embeddings that can separate different classes. Unlike many other methods which classify classes directly, triplet networks learn a mapping function that can map input signals into an embedding space that can then be used for classification. The architecture employs three identical CNN branches that share weights, processing triplets of input samples simultaneously: an anchor sample, a positive sample (same class as anchor), and a negative sample (different class). Each branch of the network has the same architecture, which means the parameters are shared which makes training much more efficient. The three inputs to the network are called the "anchor", "positive" and "negative" sample.

The CNN architecture begins with an input layer that accepts EEG signals formatted as 2D arrays (channels × time points), applying batch normalization to standardize the input distribution. This batch normalization step is useful for standardizing inputs and results in faster training times. The feature extraction stage consists of three consecutive convolutional blocks. Each block implements a 2D convolution with a 3x3 kernel size, followed by batch normalization and ReLU activation. This feature extraction stage extracts information from the input data using filters. Batch normalization and ReLU activation are common techniques in deep learning to ensure that networks are trained better. A MaxPooling operation (2x2) reduces the spatial dimensions, while a dropout rate of 0.5 helps prevent overfitting. The max pooling

operation is important for reducing the size of each feature map, which decreases computational costs. Dropout is a way to help prevent overfitting, a common problem in deep learning. The final embedding layer flattens the extracted features and processes them through a dense layer to produce 128-dimensional embeddings, which are then L2-normalized to ensure uniform scaling. This dense layer maps the features from the convolutional layers to a final 128 dimensional embedding space, and L2-normalization ensures that the values are of unit length, which improves model performance.

The network is trained using triplet loss, defined mathematically as

$$L_{triplet} = \left(\left|\left| f(x_a) - f(x_p) \right|\right|_2^2 - \left|\left| f(x_a) - f(x_n) \right|\right|_2^2 + \alpha, 0 \right) \tag{4.8}$$

where f(x) represents the CNN embedding function, α is the margin parameter (set to 0.2), and ||·||2 denotes the L2 norm. The triplet loss aims to ensure that the distance between embeddings of samples from the same class (anchor and positive) is small, and the distance between embeddings of samples from different classes (anchor and negative) is large. This loss function optimizes the embedding space by enforcing the constraint effectively pushing same-class samples closer together while separating different-class samples. The α parameter is important because it provides a separation between the classes and ensures that model is not overconfident. The goal is to ensure the same class samples are closer than a margin, α, to the different class samples.

To enhance the network's learning efficiency, we implement hard triplet mining. This approach selects the most challenging triplets within each batch by identifying positive samples with maximum distance from the anchor and negative samples with minimum distance from the anchor. This approach is better than random selection of the positive and negative samples because it allows the model to learn from the hardest examples which significantly increases the learning speed of the model. This strategy ensures that the network focuses on the most informative examples during training, leading to more robust feature learning.

For EEG-based inner speech classification, the produced embedding space exhibits numerous important features. Independent of subject identification. Learning a certain embedding space helps the network to better cluster related inner speech patterns while keeping constant distances between many patterns. Moreover, subject-specific features are eliminated from the embedding space by learning the embeddings using the triplet loss, hence enabling improved generalization among subjects. Particularly useful for cross-subject generalization, the architecture may acquire discriminative features via relative distances instead of absolute patterns. This is a significant benefit over conventional classification methods since the emphasis is on the relative class similarities rather than the absolute characteristics of the samples. This increases the model's resilience against personal fluctuations in EEG signals.

4.3 Training Process

Every model's training schedule was intended to guarantee strong performance and subject-independent generalization. This comprised a robust subject-independent testing framework, cross-valuation inside the training set, and hyperparameter optimization with Optuna. The particular training protocols followed for every model will be covered in the sections that follow.

4.3.1 Linear SVM Training Process

A leave-one-subject-out cross-validation method and Optuna, a tool for hyperparameter adjustment, were used to train the Linear SVM. For hyperparameter tuning, Optuna was employed to optimize the regularization parameter C, the penalty type (l1 or l2), and the dual formulation (when using the l2 penalty). The C parameter was explored in the range of 1e-3 to 1e3, as this parameter controls the regularization strength, and a higher value means the classifier will aim to minimize misclassification, while a lower value means the classifier will aim to maximize the margin. The penalty parameter specified the norm used in the penalization, where the L1 norm produces sparse parameters, and the L2 norm produces less sparse solutions. The dual parameter specifies whether to use the dual form or the primal form, and this parameter

is only relevant when the penalty parameter is l2. The hyperparameter search was performed using an inner stratified k-fold cross-validation strategy within the training set, using five folds to ensure the model's ability to generalize well on unseen data. The objective function was set to maximize the F1 score on the inner validation set. For the training process, for each subject in the dataset, the data from that subject was held out for testing, while the data from all remaining subjects was used for training. The training data was standardized using StandardScaler. After hyperparameter optimization, the best hyperparameters were used to train a final model on the entire training set, which was then evaluated on the held-out test set. The trained Linear SVM was evaluated using the following metrics: accuracy, precision, recall and F1 score.

4.3.2 SVM with Non-Linear Kernels Training Process

The SVM with non-linear kernels was trained using a combination of Optuna for hyperparameter tuning and a leave-one-subject-out cross-validation approach. This section also uses a stratified k-fold cross-validation strategy, and the performance metrics are the same as in the Linear SVM section above. For hyperparameter tuning, Optuna was used to optimize the hyperparameters of the SVM with non-linear kernels, namely, the regularization parameter C, the kernel (linear, poly, RBF, or sigmoid), the kernel coefficient gamma, the polynomial kernel degree degree and the coefficient for poly or sigmoid kernels, coef0. The C parameter was explored with a log scale between 1e-3 and 10, and this parameter controls the regularization strength, with smaller values meaning higher regularization. The gamma parameter was explored with a log scale between 1e-4 and 1, and it determines the kernel width or the influence of each data point. The kernel parameter specifies the kernel to use, with linear, poly, rbf, and sigmoid all being tested. The degree parameter is relevant only for polynomial kernels and specifies the degree of the polynomial. The coef0 parameter is relevant for the poly and sigmoid kernels and determines the bias of the kernel. The hyperparameter optimization strategy is the same as that used for the Linear SVM. The training and testing process was exactly the same as described for the Linear SVM method, where the data from each subject was held out for testing, while the remaining subjects were used for training.

4.3.3 LightGBM Training Process

The LightGBM model was trained using Optuna for hyperparameter optimization and the same leave-one-subject-out cross-validation approach used for the other models. For hyperparameter tuning, the following key parameters were optimized using Optuna: the number of leaves, num_leaves for the tree which were optimized with values between 10 and 150; the learning rate, learning_rate which was optimized with a log scale from 1e-4 to 1; the feature_fraction and bagging_fraction which controlled the number of features and samples to be used to prevent overfitting; and the lambda_l1 and lambda_l2 parameters which controlled L1 and L2 regularization. The hyperparameter tuning procedure is the same as described above. For the training process, the process was exactly the same as described for the SVM methods above, where the data from each subject was held out for testing, while the remaining subjects were used for training. This detailed training process allows a comprehensive and fair comparison between the different models, while also ensuring optimal performance from each.

4.3.4 CNN with Triplet Loss Training Process

The CNN with triplet loss was trained using a combination of Optuna for hyperparameter tuning, a leave-one-subject-out cross-validation approach, and a hard triplet mining strategy. For hyperparameter tuning, Optuna was used to optimize the hyperparameters of the CNN with triplet loss, which includes the number of filters, num_filters for the convolution layers which were tested from 50 to 200, the kernel size of convolutional layer, kernel_size which was tested between 3 and 7, the size of the fully connected layer, fc_size, which was tested between 30 and 120, the dropout rate which was explored between 0.1 and 0.5, the learning rate, lr, for the model which was explored using a log scale from 1e-5 to 1e-3, the margin for the triplet loss which was explored from 0.5 to 2.0 and the parameter p for the loss function which was explored from 1 to 5. Furthermore, the regularization parameter C for the final SVM classifier was optimized using Optuna, exploring a range from 0.1 to 50.0 with log scaling. The model was trained using a Stratified K-Fold Cross-validation strategy (with k=3) on the training set, leaving out the data for the test subject, and the optimization process used the F1 score as the primary metric, as is the same for the other methods. For the training procedure, for each subject in the dataset, the data from that subject was held out for testing,

while the data from all remaining subjects was used for training. The EEG data was converted to PyTorch tensors and loaded onto the GPU. Hard triplet mining was performed on the data using a HardestTripletSelector in order to only pick the hardest triplets for training, using a margin. Then the model was trained using a custom train_triplet_network2 function. After the triplet network was trained, features were extracted using the embedding network, and these features were then used to train an SVM classifier with a RBF kernel. The SVM classifier was trained using the optimal regularization parameter C found during hyperparameter tuning. The model's performance on the test set was evaluated using the same metrics as the other models: accuracy, precision, recall, and F1 score.

4.4 Summary

This part described the training technique used to maximize and assess the performance of every model in a framework independent of subjects. Optuna was used to tweak hyperparameters for every model using LightGBM, SVM, and CNN models gaining from tailored parameter searches to improve classification accuracy. Cross-valuation was done inside the training set, one-fold reserved for validation during the tuning phase, therefore preserving a strict and objective approach. Using data from n−1 individuals, each model was trained and evaluated on the remaining subject in a leave-one-subject-out framework therefore providing a strong estimate of subject-independent performance. This method enabled a thorough assessment of the efficacy of every model in inner speech classification, therefore stressing their respective advantages and drawbacks for different people.

CHAPTER 5

Results and Discussion

5.1 Overview

The results of a comparison study of the models created for inner speech categorization from EEG data are covered in Chapter 5. To evaluate generalizability, performance assessment was done under a suite of metrics accuracy, F1-score, precision, and recall all within a subject-independent perspective. An explanation of the quantitative results is offered in a discussion section after their presentation. This part emphasizes the advantages, drawbacks, and possible consequences of every model especially for applications in subject-independent BCI. As discussed in Chapter 2, other papers used subject-dependent approaches, so accuracy for a subject-independent approach will be lower.

5.2 Results and Analysis

Figure 5.1 presents the subject-wise classification results using a Linear Support Vector Machine (SVM) with statistical features extracted from inner speech EEG data. The results show considerable variation in accuracy across subjects when evaluated using leave-one-subject-out cross-validation, reflecting the inherent inter-subject variability in EEG patterns during inner speech production. The model achieved an average accuracy of **26.5%** across all subjects.

The exploration of different kernel functions in SVM, as shown in Figure 5.2, yielded an average accuracy of **25.50%**, performing marginally worse than the linear SVM. Despite

the theoretical advantage of non-linear kernels in capturing complex decision boundaries, their performance did not surpass that of the linear kernel in this context. This outcome may be due to the high noise-to-signal ratio typical in EEG data, which can hinder the ability of non-linear kernels to generalize effectively and may lead to overfitting. These results suggest that the benefits of non-linear kernels are not realized in this application, possibly because the noise in the data masks the underlying non-linear relationships.

LightGBM achieved an average accuracy of **26.5%**, exactly the same as linear SVM. This indicates that tree-based ensemble methods may be somewhat more effective at capturing complex patterns in the data than kernelize SVM. LightGBM's gradient boosting framework allows it to handle non-linear relationships and may offer some robustness to noise. However, given the small increase in accuracy, we must be cautious in drawing strong conclusions about its advantage over SVM methods.

With an average accuracy of 26.8%, the triplet network design using a CNN was the most effective. When compared to the alternative approaches, this is an incremental step forward. This shows that deep learning methods, which construct representations from raw EEG data, could be better able to capture discriminative features than conventional feature engineering methods. One possible explanation for its performance could be the combination of the CNN's capability to learn hierarchical features and the triplet network's ability to learn an embedding space that highlights similarities and differences across classes. To completely evaluate deep learning's potential in this setting, more research with more datasets may be required, but the gain is noticeable.

Table 5.1: Linear SVM results

Subject	Accuracy	Precision	Recall	F1 Score
1	0.210	0.210	0.201	0.200
2	0.292	0.292	0.288	0.287
3	0.267	0.267	0.269	0.263
4	0.287	0.287	0.283	0.279
5	0.254	0.254	0.255	0.251
6	0.310	0.310	0.308	0.304
7	0.242	0.242	0.240	0.238
8	0.280	0.280	0.279	0.273
9	0.288	0.287	0.279	0.279
10	0.221	0.221	0.217	0.221

Average	0.265	0.265	0.262	0.259

Table 5.2: SVM with RBF results

Subject	Accuracy	Precision	Recall	F1 Score
1	0.285	0.285	0.290	0.278
2	0.258	0.258	0.260	0.247
3	0.283	0.283	0.283	0.278
4	0.242	0.242	0.237	0.234
5	0.254	0.254	0.243	0.240
6	0.301	0.302	0.318	0.292
7	0.233	0.233	0.238	0.229
8	0.205	0.205	0.191	0.195
9	0.250	0.250	0.250	0.238
10	0.233	0.233	0.237	0.231
Average	0.255	0.255	0.255	0.246

Table 5.3: LightGBM results

Subject	Accuracy	Precision	Recall	F1 Score
1	0.280	0.280	0.279	0.277
2	0.308	0.308	0.309	0.307
3	0.288	0.288	0.273	0.272
4	0.271	0.271	0.266	0.268
5	0.246	0.246	0.245	0.245
6	0.282	0.282	0.286	0.282
7	0.254	0.254	0.254	0.254
8	0.240	0.240	0.261	0.231
9	0.250	0.250	0.252	0.250
10	0.225	0.225	0.227	0.220
Average	0.265	0.265	0.265	0.261

Table 5.4: Triplet network results

Subject	Accuracy	Precision	Recall	F1 Score
1	0.280	0.280	0.307	0.254
2	0.300	0.300	0.315	0.293
3	0.244	0.244	0.241	0.239
4	0.237	0.237	0.237	0.195
5	0.308	0.308	0.307	0.304
6	0.245	0.245	0.239	0.240

7	0.283	0.283	0.274	0.273
8	0.280	0.280	0.277	0.266
9	0.258	0.258	0.287	0.211
10	0.250	0.250	0.264	0.232
Average	0.268	0.268	0.274	0.2507

Table 5.5: Comparison of models accuracy

Subject	Linear SVC	SVM	LightGBM	Triplet Network
1	0.210	0.285	0.280	0.280
2	0.292	0.258	0.308	0.300
3	0.267	0.283	0.288	0.244
4	0.287	0.242	0.271	0.237
5	0.254	0.254	0.246	0.308
6	0.310	0.301	0.282	0.245
7	0.242	0.233	0.254	0.283
8	0.280	0.205	0.240	0.280
9	0.288	0.250	0.250	0.258
10	0.221	0.233	0.225	0.250
Average	0.265	0.255	0.265	0.268

The confusion matrix for Subject 1 is shown in Figure 6.1. This matrix's tendency to classify instances as "down" more often than other categories indicates a blatant bias in the model's predictions. On the other hand, the confusion matrix for Subject 4 in Figure 6.2 exhibits a distinct predictive pattern. The model shows a significant bias toward "up" predictions for this subject. The sharp disparity in these results between Subjects 1 and 4 reveals a high level of inter-subject variability, suggesting that the model's functionality and particular biases vary from person to person.

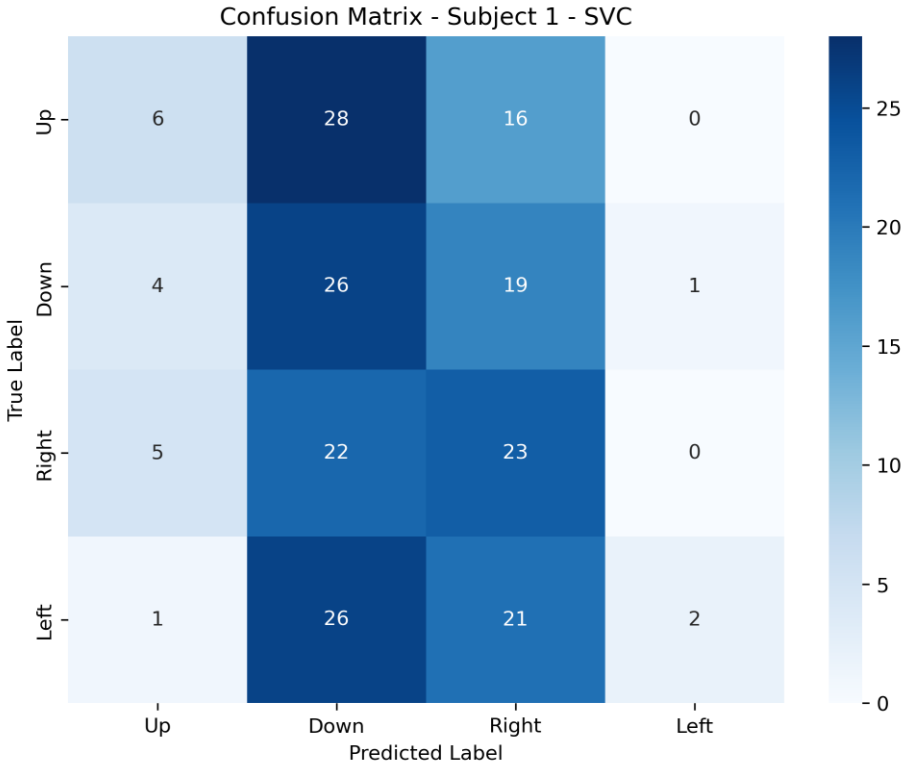

Figure 5.1 Confusion matrix for subject 1 when predicted using triplet embedding with SVC classifier

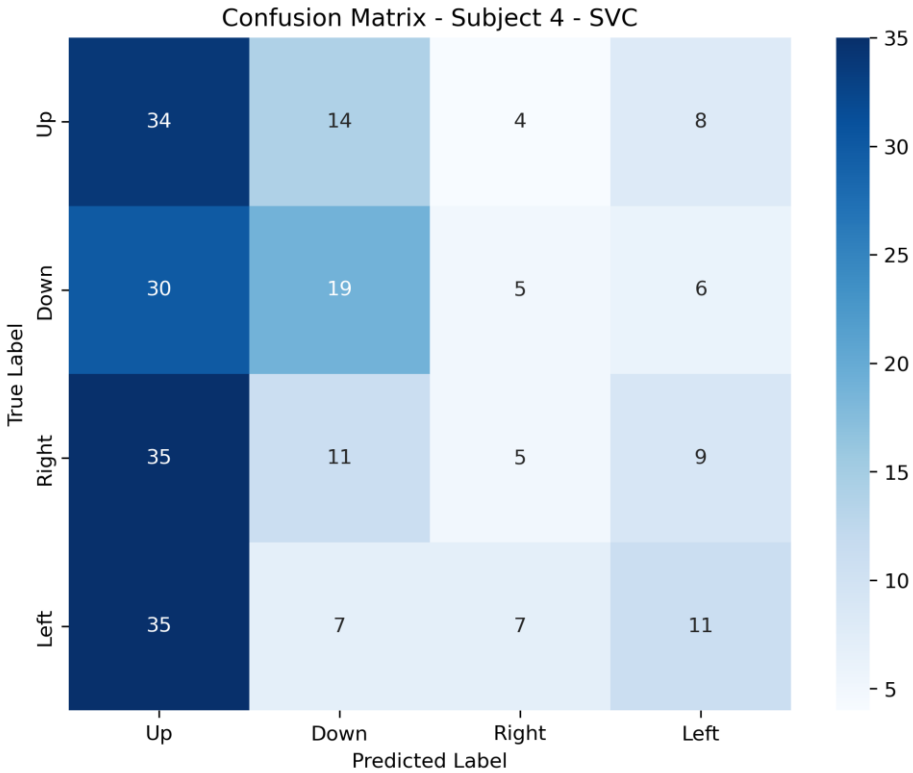

Figure 5.2 Confusion matrix for subject 4 when predicted using triplet embedding with SVC classifier

5.3 Discussion

In contrast to earlier research that mostly focused on subject-dependent approaches in order to test the models' capacity for generalization, this study evaluated models for a subject-independent approach.. Deep learning techniques improved accuracy from 25.5% to 26.8% compared to typical machine learning methods, suggesting that representation learning and non-linear modeling may play a significant role in inner speech EEG classification. The advantages are small, but they show that more advanced models could be able to provide even smaller ones. Based on these findings, it seems that expanding our focus beyond classic feature engineering techniques and investigating more complex neural network topologies could lead to better classification performance in the future. Another possible solution is to make use of bigger and more varied datasets.

5.3.1 Determinants of Generalizability

Our subject-independent evaluation's findings point to two main issues that restrict the generalizability of the model for inner speech classification: the low signal-to-noise ratio (SNR) present in EEG data and high inter-subject variability.

The comparatively low average accuracy across all models is primarily caused by the substantial inter-subject variability. Each model's performance varied significantly between subjects, as Table 7 illustrates. For instance, the accuracy of the SVM with the RBF kernel was 0.205 for Subject 8, but 0.285 for Subject 1. The clear predictive biases seen in the confusion matrices for Subjects 1 and 4 (Figures 6.1 and 6.2) further support this, showing that a model's particular failure modes can differ significantly from person to person. This implies that learning a common, reliable representation of inner speech signals across a diverse population is difficult for a single, one-size-fits-all model. Furthermore, the low SNR typical of non-invasive EEG recordings proved to be a significant obstacle. It can be challenging for models to capture discriminative features because this noise can obscure the subtle, underlying brain signals associated with inner speech. Our discovery that non-linear SVM kernels did not perform better than the linear kernel raises the possibility that the noise-to-signal ratio of the data may limit these intricate models' capacity to generalize, possibly resulting in overfitting on noise instead of the actual signal. The slight improvements observed with more sophisticated models, such as the triplet network, demonstrate that although deep learning can start to address this issue, it still remains a critical hurdle.

5.4 Summary

In this section, we examined various models for subject-agnostic EEG-based inner speech classification. An average accuracy of 26.8% was achieved by the convolutional neural network (CNN) that utilized triplet loss, highlighting the promise of deep learning and representation learning in the field of inner voice classification. With accuracies of 25.5% and 26.5%, respectively, traditional machine learning methods like Linear SVM and LightGBM demonstrated their effectiveness in managing noise and variability in EEG data. Despite LightGBM's minor performance advantage in capturing complicated patterns, non-linear SVMs were not able to surpass the linear kernel owing to the high noise-to-signal ratio. Further

investigation into deep learning methods and data gathering tactics for BCI applications is necessary, as the results indicate that more complex structures and bigger, more varied datasets may lead to slight gains in classification performance.

CHAPTER 6

CONCLUSION AND FUTURE WORK

6.1 Overview

Classification of subject-independent inner speech using EEG signals was the main focus of this research. The goal was to compare the generalizability of deep learning approaches, especially a CNN with triplet loss, against that of more conventional machine learning models, such as SVM and LightGBM, using a subject-independent approach. The study's overarching goal was to promote the development of reliable brain-computer interface (BCI) systems capable of decoding inner speech, so as to address the challenges posed by inter-subject variability and the large noise-to-signal ratio that characterize EEG data.

6.2 Summary

This study has made the following contributions:

i. Compared a deep learning model (CNN with triplet loss) against more conventional machine learning models (Linear SVM, SVM with RBF, and LightGBM) for inner speech categorization that is independent of the subject.
ii. Demonstrated the marginal superiority of deep learning approaches in generalizing across subjects, with the CNN and triplet loss architecture achieving the highest average accuracy of 26.8% across subjects.
iii. Highlighted the limitations of traditional feature engineering and non-linear kernels for noisy and highly variable EEG data, providing evidence for the potential of representation learning in BCI applications.

6.3 Future Work

Drawing from the insights gained in this study, future research should emphasize deep learning approaches, with a particular focus on investigating methods for learning improved representations and testing out new architectures. Collecting more extensive and varied datasets should be a priority, as this would facilitate the use of advanced techniques like transfer learning and the training of cutting-edge models such as transformers. Furthermore, tackling the inherent noise in EEG signals is still a significant hurdle; future endeavors should concentrate on creating effective noise-reduction techniques to enhance signal quality and boost model performance. These steps will not only bolster subject-independent inner speech classification but also progress the field of brain-computer interfaces more generally.

REFERENCES

1. Alonso, L. F. Nicolas, & Gomez-Gil, J. (2012). Brain computer interfaces, a review. Sensors (Basel). 12(2), 1211-1279.
2. Holz, E. M., Botrel, L., Kaufmann, T., & Kübler, A. (2015). Long-term independent brain-computer interface home use improves quality of life of a patient in the locked-in state: a case study. Arch Phys Med Rehabil. 96(3), S16-S26.
3. Nieto, N., Peterson, V., Rufiner, H. L., et al. (2022). Thinking out loud, an open-access EEG-based BCI dataset for inner speech recognition. Sci Data. 9, 52.
4. Mudgal, S. K., Sharma, S. K., Chaturvedi, J., & Sharma, A. (2020). Brain computer interface advancement in neurosciences: Applications and issues. Interdisciplinary Neurosurgery. 20, 100694.
5. Berg, B. van den, Donkelaar, S. van, & Alimardani, M. (2021). Inner speech classification using EEG signals: A deep learning approach. 2021 IEEE 2nd International Conference on Human-Machine Systems (ICHMS). 8-10 September 2021. Magdeburg, Germany: IEEE, 1–4.
6. Liwicki, F. S., Gupta, V., Saini, R., De, K., & Liwicki, M. (2022). Rethinking the methods and algorithms for inner speech decoding and making them reproducible. NeuroSci. 3(2), 226–244.
7. Jonsson, L. (2022). Using machine learning to analyse EEG brain signals for inner speech detection. Master's Thesis, Luleå University of Technology, Luleå, Sweden.
8. Gasparini, F., Cazzaniga, E., & Saibene, A. (2022). Inner speech recognition through electroencephalographic signals. arXiv preprint arXiv:2210.06472.
9. Cooney, C., Folli, R., & Coyle, D. (2019). Optimizing layers improves CNN generalization and transfer learning for imagined speech decoding from EEG. 2019 IEEE International Conference on Systems, Man and Cybernetics (SMC). 6-9 October 2019. Bari, Italy: IEEE, 1311–1316.
10. Cooney, C., Korik, A., Folli, R., & Coyle, D. (2020). Evaluation of hyperparameter optimization in machine and deep learning methods for decoding imagined speech EEG. Sensors. 20(16), 4629.

11. Zhao, S., & Rudzicz, F. (2015). Classifying phonological categories in imagined and articulated speech. 2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). 19-24 April 2015. South Brisbane, QLD, Australia: IEEE, 992–996.
12. Coretto, G. A. P., Gareis, I. E., & Rufiner, H. L. (2017). Open access database of EEG signals recorded during imagined speech. Proceedings of the SPIE, Volume 10160: 12th International Symposium on Medical Information Processing and Analysis. 8-10 December 2016. Tandil, Argentina: SPIE, 1016002.
13. Nguyen, C. H., Karavas, G. K., & Artemiadis, P. (2017). Inferring imagined speech using EEG signals: A new approach using Riemannian manifold features. Journal of Neural Engineering. 15(1), 016002.
14. Gonz´alez-Casta˜neda, E. F., Torres-Garc´ıa, A. A., Reyes-Garc´ıa, C. A., & Villase˜nor-Pineda, L. (2017). Sonification and textification: Proposing methods for classifying unspoken words from EEG signals. Biomedical Signal Processing and Control. 37, 82–91.
15. Agarwal, P., & Kumar, S. (2022). Electroencephalography-based imagined speech recognition using deep long short-term memory network. ETRI Journal. 44(1), 672-685.
16. Geva, S., Jones, P. S., Crinion, J. T., Price, C. J., Baron, J. C., & Warburton, E. A. (2011). The neural correlates of inner speech defined by voxel-based lesion-symptom mapping. Brain. 134(10), 3071-3082.
17. Ke, G., Meng, Q., Finley, T., Wang, T., Chen, W., Ma, W., Ye, Q., & Liu, T.-Y. (2017). LightGBM: A highly efficient gradient boosting decision tree. Advances in Neural Information Processing Systems. 30, 4-9 December 2017. Long Beach, California, USA: Curran Associates, Inc.
18. Ng, H. W., & Guan, C. (2024). Subject-independent meta-learning framework towards optimal training of EEG-based classifiers. Neural Networks. 172, 106108.
19. Radwan, Y. A., Mohamed, E. A., Metwalli, D., Barakat, M., Ahmed, A., Kiroles, A. E., & Selim, S. (2025). Stochasticity as a solution for overfitting—A new model and comparative study on non-invasive EEG prospects. Frontiers in Human Neuroscience. 19, 1484470.