Reconstruction of inner speech from Electroencephalogram (EEG) signal: A Quantitative Analysis of different Machine Learning and Deep Learning Models

I. Introduction

Speech is the vocalized form of communication through which humans convey thoughts, emotions and information in meaningful ways. It involves using several components of the human body like vocal cords, tongue, teeth, lips. However, there are individuals who suffer from speech impairment due to various conditions, e.g. neurodegenerative diseases, vocal abuse, mental disorder, accident and brain traumas. This type of impairment can significantly affect the individuals, decreasing their quality of life. However, humans possess an innate ability to produce speech internally. This is commonly referred to as inner speech. The internal dialogue that occurs within the human mind is called inner speech. Inner speech plays an important role in human communication as we use it for our thought process, problem solving and self- reflection. This speech operates entirely within our brain.

Advances in technology have allowed us to harness this inner speech. Brain-Computer interface, also known as BCI, is a technology that allows direct communication from the brain. BCIs operate by detecting electrical signals from the brain, which is then captured by some external devices. These external devices have both invasive and non-invasive methods. Invasive methods are operated by surgical means and non-invasive methods like Electroencephalogram (EEG) can be operated without surgery. It is less costly and requires significantly less time for operation. These electrical signals captured by the BCIs are then used to detect the neural activity associated with our inner speech i.e. thoughts, intentions and so on.

Different Machine Learning (ML) and Deep learning (ML) Models are used to process and interpret these signals. Support Vector Machines (SVM), Long Short-Term Memory networks (LSTM), Convolutional Neural Networks (CNN), Deep Neural Networks (DNN) and Bidirectional Long Short-Term Memory (BiLSTM)

are such type of machine learning and deep learning models. These models analyze signals differently, and as such have variation in accuracy, efficiency and usability.

This study aims to explore potentials of different ML and DL approaches in BCIs for inner speech reconstruction. By comparing the collected data of different models, we can identify the most effective model for enhancing BCIs and also improving the lives of individuals with speech impairment to regain their voice.

II. Background

Recent studies on Reconstructing internal speech have identified some key themes. People often evaluate it based on several aspects such as: intervention (imagined speech, covert speech, inner speech), datatype (electroencephalography, electropalatographic and electromyography). These are sufficient to retrieve all relevant studies.

EEG: In the last decade, numerous studies have explored to using the EEG data on inner speech. EEG is the most widely used non-invasive modality for practical use since it does not involve any surgical process and is relatively easy to access [1]. However, preliminary works were conducted with very few participants and syllables, where EEG waveform envelopes have been adopted to recognize EEG patterns.[2]

Bootstrap (BTS): It's a statistical technique often used in BeI for improving model reliability. Speech reconstruction from spoken speech or mimed speech brain signals, EMG Data has shown potential [3][4][5]

Electrocorticographic Recordings: Electrocorticography (ECOG) is used in patients with intractable epilepsy to localize the seizure onset zone, prior to brain tissue ablation. In this procedure, electrode grids, strips are depth electrodes are temporarily implanted onto the cortical surface, either above or below the dura mater. ECOG is suitable for neuroprosthetic and brain-computer interface applications [6]. ECOG activity contains different signal components that may related to different underlying physiological mechanisms [7]

Decoding Models: Decoding Models are algorithms or system designed to interpret brain signals (e.g., EEG, fMRI,or ECoG) and translate them into meaningful commands or outputs. Decoding models allow researchers to apply multivariate neural features to rich, complex and naturalistic stimuli or behavioral conditions.[8] [9] [10]. Holdgraf et al. (2017)[11] provide a review article that illustrates best- practices in conducting these analyses and included a small sample dataset along with several scripts in the form of 'Jupyter Notebooks'. The general framework is common to all methods.

Reserach Gaps: While there is extensive research work to which speech representations, such as acoustic processing, phonetic encoding and higher level of linguistic functions apply to inner speech, the lack of behavioral output during imagery, and inability to monitor the spectrotemporal structure of inner speech is still a major challenge.

Since imagined speech has no reference voice, voice samples during the spoken speech, which is recorded in the same sequence as imagined speech, were used as the target audio. To match the EEG to the voice of spoken speech, dynamic time warping. (DTW) was applied between the reconstructed mel-spectrogram from EEG and the mel- spectrogram of voice during spoken speech.

III. Datasets

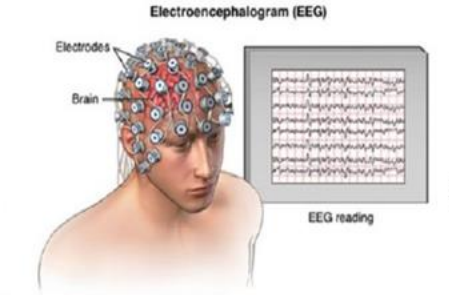
1. Participants: Ten adult male and female volunteers took part in this study (undergraduate and graduate students, aged 20-35 years old). All of them had no experience of psychophysiological examinations, but they were ensured proper instructions are followed. All of them confirmed participation consent and were adapted to use as data collection.

2. Paradigms: The participants were asked to respond to specific tasks, such as visual or auditory stimuli, motor tasks or resting state recordings. Specific symbols were shown in front of the participants for response. Participants were divided by a hundred trials of both spoken and imagined speech. Therefore, each participants had one-thousands trials for the spoken and imagined speech paradigm.

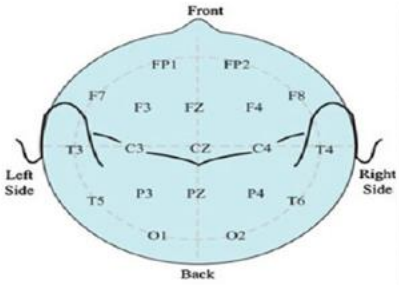
Protocol	Prompts	Total Number of Participants	IDs of the Participants
Long words	"independent" and "cooperate"	6	S2, S3, S6, S7, S9 and S11
Short words	"in", "out" and "up"	6	S1, S3, S5, S6, S8 and S12
Vowels	"/a/", "/i/" and "/u/"	8	S4, S5, S8, S9, S11, S12, S13 and S15
Short-long words	"in" and "cooperate"	6	S1, S5, S8, S9, S10 and S14

Figure 1: Specific symbols for the participants

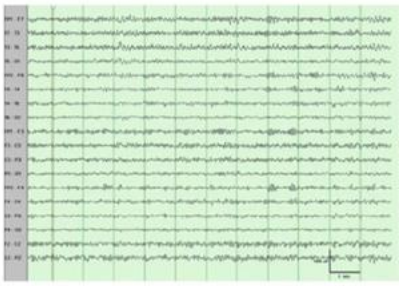
3. EEG recordings: The EEG system was set up, placing electrodes on participant's scalps according to the international 10-20 system. The recording was made with a 64- channel EEG amplifier from Neuro scan, using the left mastoid for reference and the right for using as ground. The EEG was sampled with 250 HZ, it was filtered between 1 and 50 Hz with Nitochfilter on 60 EEG channels were Recorded according the scheme.



An illustration of EEG recording



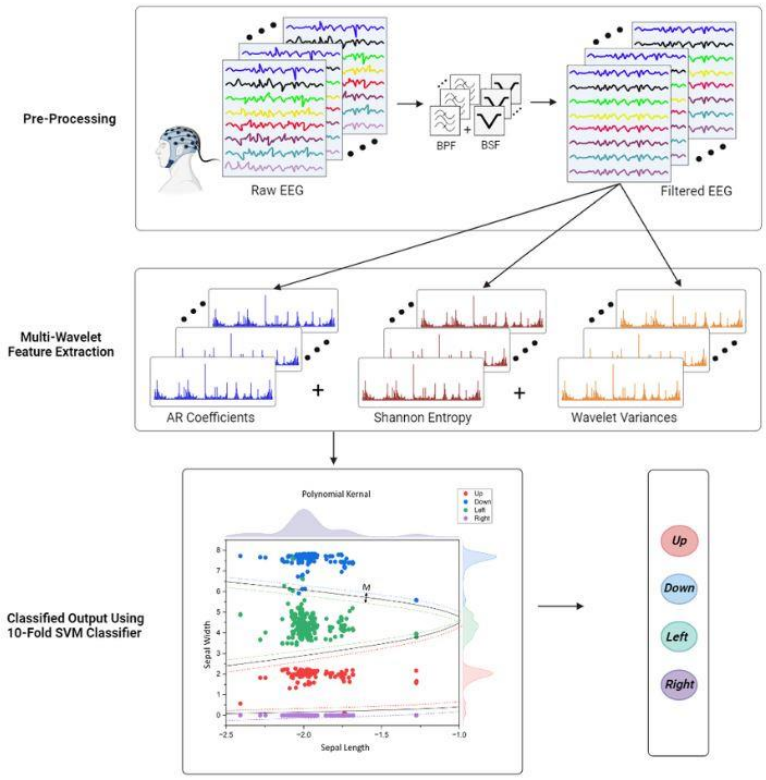
International 10-20 system of EEG Recording


Normal EEG signal

4. Data analysis: Data was analyzed using multiple regression analysis to identify the impact of various independent variables such as the participants own personal thoughts on the dependent variables, such as given tasks. Relevant features like power spectral density, event- related potentials, and connectivity potentials will be extracted from the EEG signals [12]

IV. ML and DL Models

SVMs: Support Vector Machine (SVM) is a widely used machine learning method in Brain-Computer Interface systems due to its effectiveness in classification tasks. It can classify neural signals associated with different words like ‘Yes’ and ‘No’ for inner speech. It is designed for high-dimensional datasets, but it works with low-dimensional datasets too. [13]



Architecture of proposed Support Vector Machine (SVM) model for classifying inner speech [14]

ELMs: Extreme Learning Machine (ELM) is a machine learning language designed for single-hidden layer forward neural networks (SLFNs). It assigns random weights and biases in the hidden layer and then determines the output weights analytically. With the use of Kernel-based ELM variations, its application has been extended to nonlinear signal mappings, which is significant for inner speech reconstruction [15]

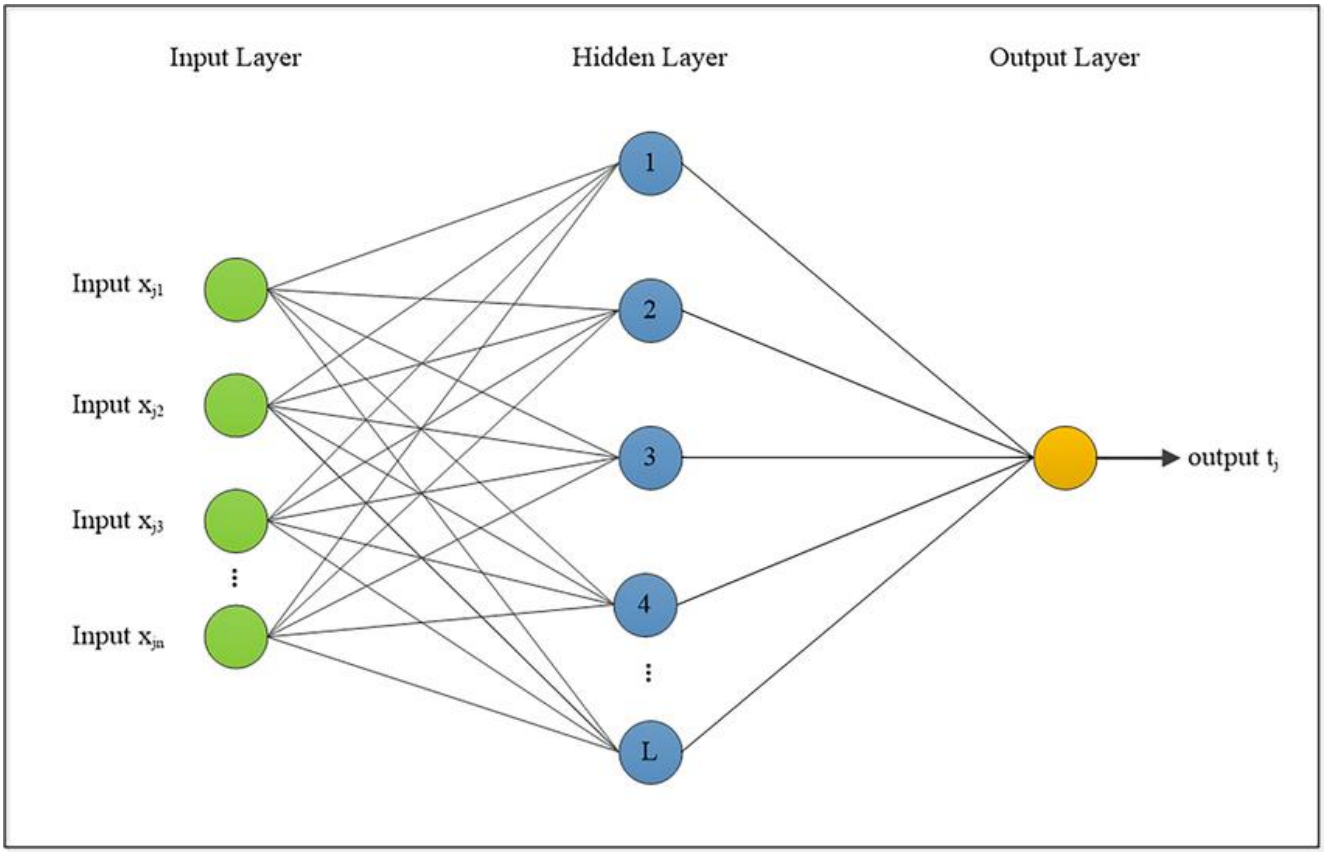


Diagram of the extreme learning machine [16]

DNNs: Deep Neural Network (DNN) is an artificial neural network that is made up of multiple interconnected neurons. It consists of input layer, multiple hidden layers and an output layer. The input layer receives data in form of image, text or in numerical values. Hidden layer extracts and process the received data by performing various mathematical transformation. The output is the last layer of DNN which generates predictions based on the data extracted from the hidden layers [17].

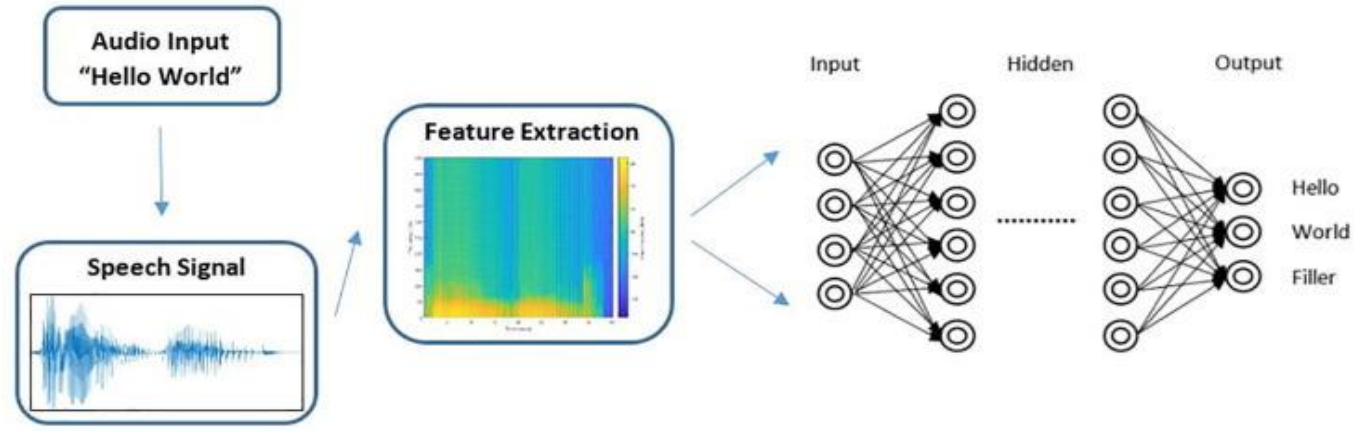


Diagram of the Deep Neural Network [18]

SNN: Shallow Neural Network is another artificial neural network that is similar to Deep Neural Network. It consists of Input Layer, Hidden Layer and Output Layer. However, compare to DNNs, it only has one hidden layer. So, it is less complex and requires less computational power and time to extract features from the input layer. As a result, it can output data much faster compare to DNNs. [19]

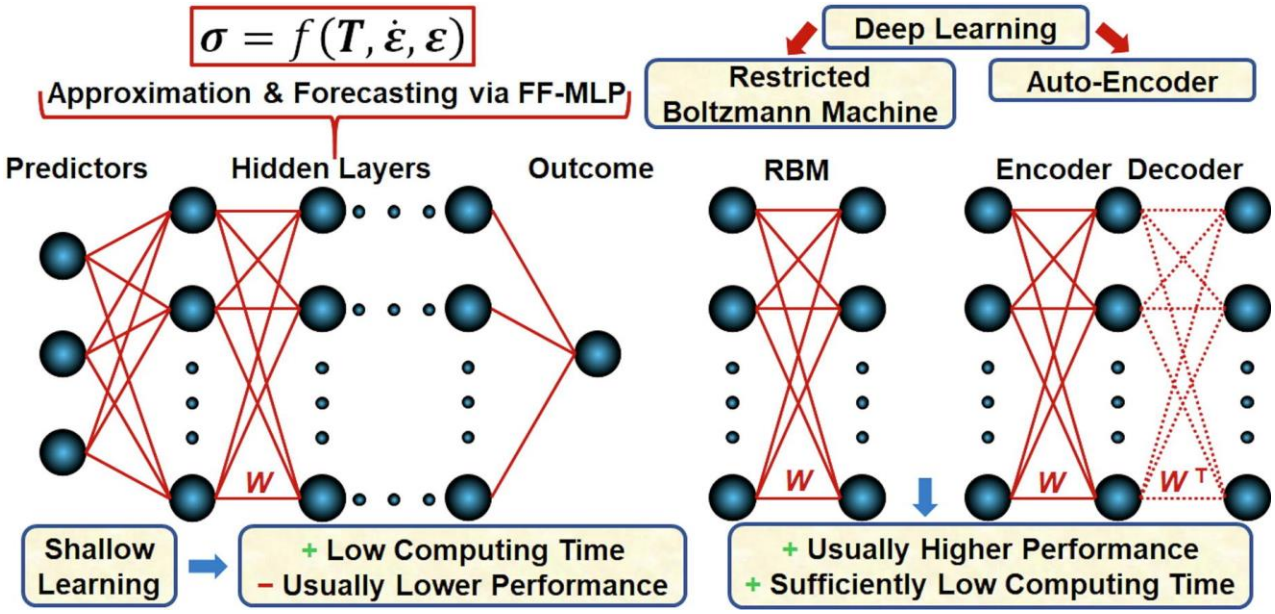


Diagram of the Shallow Neural Network **[20]**

CNN: Convolutional Neural Network (CNN) is a machine learning model that is primarily designed for image, video and spatial data processing. It is related with Deep Neural Network for its similar compositions e.g. Input layer, Hidden layer and Output Layer. However, CNN is much faster than DNN for data processing

because it has lesser parameters and usually focus on local features. CNNs extracts data from the feature space by using Convolutional layers. Then the data is down sampled using pooling layers and finally for output it flattens the data using Fully Connected Layers. Dropout layers are usually used to prevent overfitting by deactivating neurons randomly during trainings. [21]

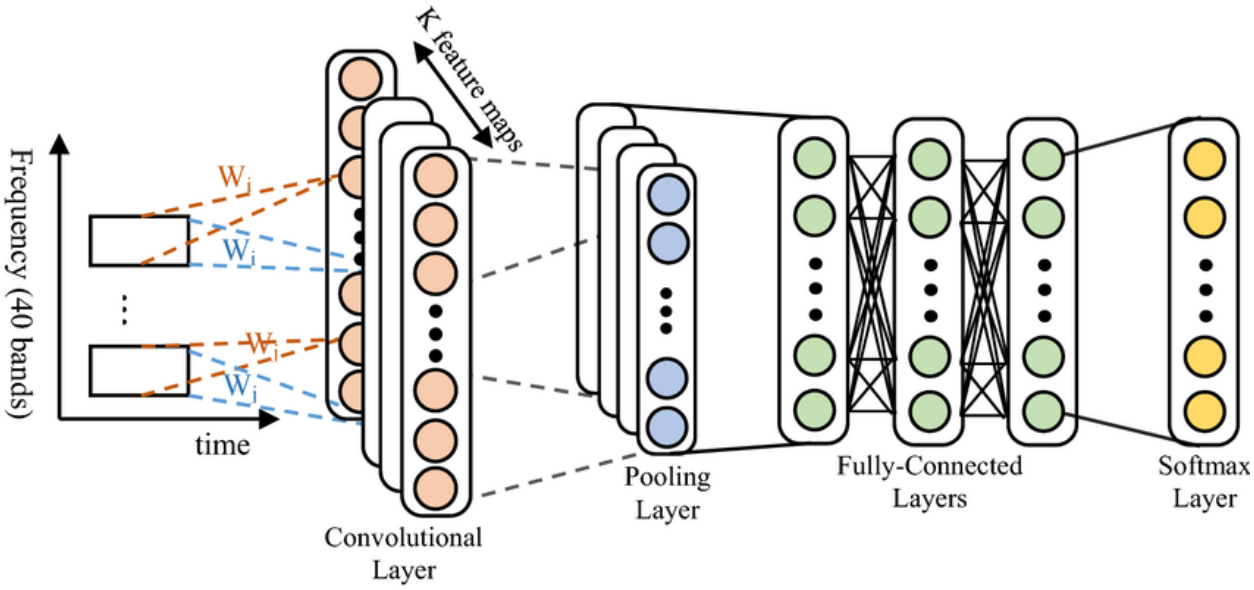


Diagram of the Convolutional Neural Network **[22]**

DBN: Deep Belief Network is a probabilistic generative machine learning model that is designed to learn hierarchical representations of data through unsupervised learning for pretraining. After pretraining, it is finely tuned with labeled data using supervised learning approach like backpropagation. It is composed of multiple layers of Restricted Boltzmann Machines (RBMs). Each RBM is made up of two layers e.g. Visible layer for inputting data and a Hidden layer for learning features. RBMs are the fundamental layers that make up architecture of DBN. Without them, DBN would lose its core mechanisms like learning hierarchical representations and unsupervised learning.[23]

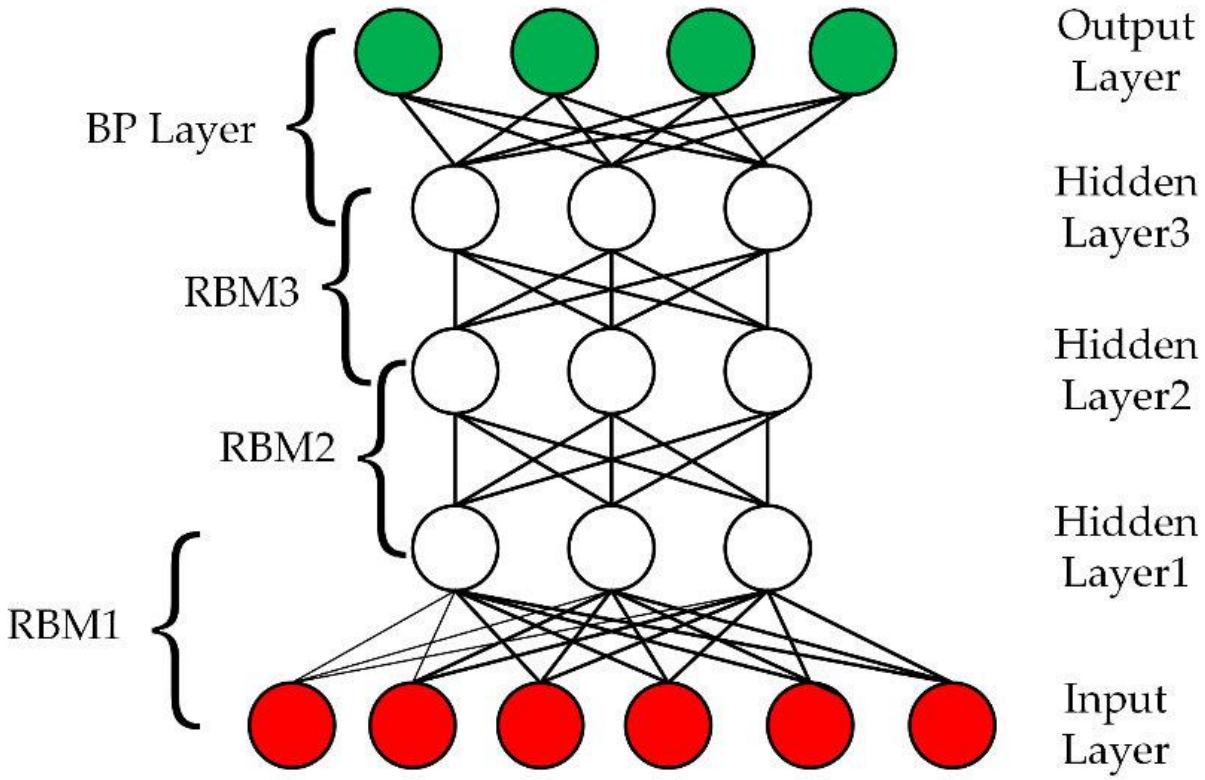


Diagram of the Deep Belief Network [24]

V. Comparison between Models

In this study we have used different types of Machine learning and Deep learning models. Each model has different characteristics and approaches. By analyzing the output data with the initial data, we could identify the strengths and weakness of different ML and DL models for inner speech reconstruction.

Name	Strengths	Weakness

Support Vector Machine	1. Worked well for binary classification 2. Found effective for low-dimensional datasets. [25]	1. Not ideal for complex non-linear patters. 2. Limited usage for high-dimensional datasets. [25]
Extreme Learning Machine	1. Very fast training speed 2. Found effective for small-scale problems and non-linear mappings. [26]	1. Shown problems for high-dimensional datasets. [26] 18
Deep Neural Network	1. Found suitable for high-dimensional datasets and complex non-linear mappings. [27]	1. Shown problems for low-dimensional datasets. [27]
Shallow Neural Network	1. Faster training speed than DNN 2. Required Less parameters. [28]	1. Had Limited capacity while modelling complex non-linear datasets. [28]
Convolutional Neural Network	1. Shown promising result while extracting spatial data. [29]	1. Required more effort in data augmentation. [29]
Deep Belief Network	1. Found effective for extracting datasets. [30]	1. Required much more time compare to previous datasets. [30]

VI. Discussion

After extracting data from the BCIs using EEG signals, we used six different Machine Learning and Deep Learning Models for processing. For criteria, we selected which model have better result in extracting data in noisy signal and output data accuracy. After rigorous analysis, we have come to conclusion that Convolutional Neural Network is the best model for inner speech reconstruction.

The F-scoreis a measure of a model's accuracy that is calculated by combining the precision and recall of the model. It is calculated by the following formula:

$$\text{F-score} = \frac{2 * \text{Precision} * \text{Recall}}{\text{Precision} + \text{Recall}}$$

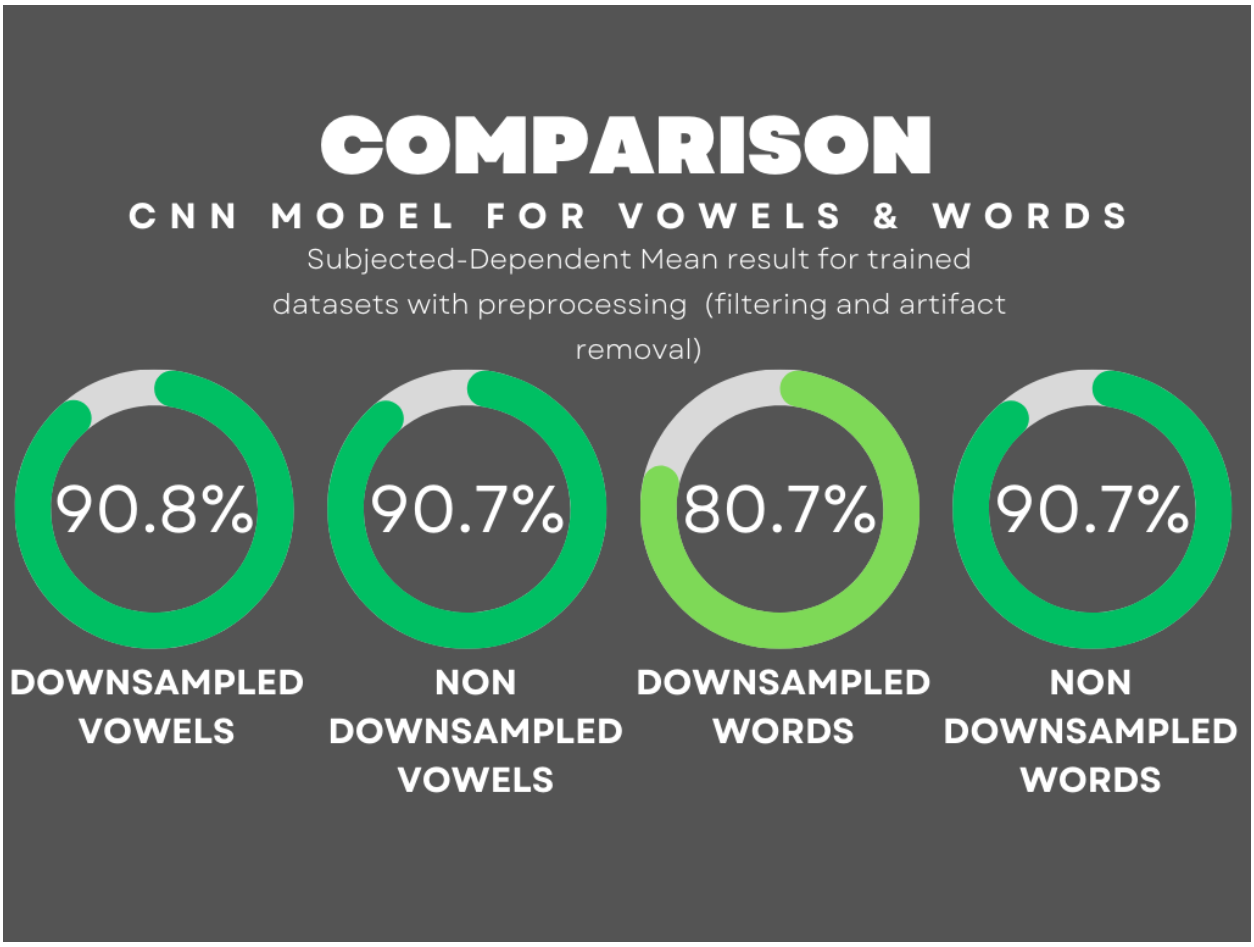


Diagram of Comparison of CNN models [31]

CNN is good in terms of accuracy of the result. One of the main advantages of CNNs is the ability to extract data even in noisy signal, which is a very common

problem found in other models. Moreover, CNNs can be used to process both low-dimensional and high dimensional signals. [32]

VII. Conclusion

In this study we have explored the idea of inner speech reconstruction using BCIs from EEG signals. We have discussed about the importance of inner speech reconstruction for speech impairment individuals. With the advancement of this technology, we can improve their living standards. For data processing, we have used six Machine Learning and Deep Learning Models. After analyzing the processed data, we concluded that Convolutional Neural Network is the best models in terms of data extraction and accuracy than the rest of the other models.

References

[1] P. Krishna, S. Kumar, and M. Ramesh, "Towards Voice Reconstruction from EEG during Imagined Speech," in *Proceedings of the Thirty-Seventh AAAI Conference on Artificial Intelligence*, 2020. [Online]. Available: https://aaai.org/. doi: 10.1609/aaai.v37i10.55000. [Accessed: Jan. 16, 2025].

[2] D. Lopez-Bernal, D. Balderas, P. Ponce, and A. Molina, "A state-of-the-art review of EEG-based imagined speech decoding," *Front. Neurosci.*, vol. 15, p. 695413, 2021. [Online]. Available: https://doi.org/10.3389/fnins.2021.695413. [Accessed: 10-Jan-2025].

[3] D. Gaddy and D. Klein, "Digital voicing of silent speech," *Proc. of the 2020 Conf. on Neural Information Processing Systems (NeurIPS)*, 2020. [Online]. Available: https://nlp.cs.berkeley.edu/pubs/Gaddy-Klein_2020_DigitalVoicing_paper.pdf. [Accessed: 10-Jan-2025].

[4] D. Gaddy and D. Klein, "An improved model for voicing silent speech," *Proc. of the 2021 Conf. on Neural Information Processing Systems (NeurIPS)*, 2021. [Online]. Available: https://nlp.cs.berkeley.edu/pubs/Gaddy-Klein_2021_ImprovedSilentSpeech_paper.pdf. [Accessed: 10-Jan-2025].

[5] J. A. Gonzalez, L. A. Cheah, A. M. Gomez, P. D. Green, J. M. Gilbert, S. R. Ell, R. K. Moore, and E. Holdsworth, "Direct speech reconstruction from articulatory sensor data by machine learning," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 25, no. 10, pp. 2072–2083, Oct. 2017. [Online]. Available: https://doi.org/10.1109/TASLP.2017.2757263. [Accessed: 10-Jan-2025]

[6] D. A. Moses, M. K. Leonard, and E. F. Chang, "Real-time classification of auditory sentences using evoked cortical activity in humans," *J. Neural Eng.*, vol. 15, no. 3, p. 036005, 2018, doi: 10.1088/1741-2552/aaab6f. [Accessed: 11-Jan-2025].

[7] A.-L. Giraud and D. Poeppel, "Cortical oscillations and speech processing: emerging computational principles and operations," *Nat. Neurosci.*, vol. 15, no. 4, pp. 511–517, 2012, doi: 10.1038/nn.3063. [Accessed: 11-Jan-2025].

[8] K. N. Kay, T. Naselaris, R. J. Prenger, and J. L. Gallant, "Identifying natural images from human brain activity," *Nature*, vol. 452, no. 7185, pp. 352–355, 2008, doi: 10.1038/nature06713. [Accessed: 11-Jan-2025].

[9] K. Kay and J. Gallant, "I can see what you see," *Nat. Neurosci.*, vol. 12, p. 245, 2009, doi: 10.1038/nn0309-245. [Accessed: 11-Jan-2025].

[10] T. Naselaris, K. N. Kay, S. Nishimoto, and J. L. Gallant, "Encoding and decoding in fMRI," *NeuroImage*, vol. 56, no. 2, pp. 400-410, May 15, 2011, doi: 10.1016/j.neuroimage.2010.07.073. [Accessed: 12-Jan-2025].

[11] C. R. Holdgraf, J. W. Rieger, C. Micheli, S. Martin, R. T. Knight, and F. E. Theunissen, "Encoding and decoding models in cognitive electrophysiology," *Front. Hum. Neurosci.*, vol. 7, p. 18, 2013, doi: 10.3389/fnhum.2013.00018. [Accessed: 12-Jan-2025].

[12] H. Zhang, Q. Q. Zhou, H. Chen, X. Q. Hu, W. G. Li, Y. Bai, J. X. Han, Y. Wang, Z. H. Liang, D. Chen, F. Y. Cong, J. Q. Yan, and X. L. Li, "The applied principles of EEG analysis methods in neuroscience and clinical neurology," *Mil. Med. Res.*, vol. 10, no. 1, p. 67, Dec. 19, 2023, doi: 10.1186/s40779-023-00502-7. [Accessed: 12-Jan-2025].

[13] F. Gasparini, E. Cazzaniga, and A. Saibene, "Inner speech recognition through electroencephalographic signals," *IEEE Access*, vol. 9, pp. 2100–2111, 2021, doi: 10.1109/ACCESS.2021.3054267. [Accessed: 12-Jan-2025].

[14] M. Abdulghani, W. Walters, and K. Abed, "Enhancing the classification accuracy of EEG-informed inner speech decoder using multi-wavelet feature and support vector machine," *IEEE Access*, vol. PP, no. 99, pp. 1-1, Jan. 2024, doi: 10.1109/ACCESS.2024.3474854. [Accessed: 16-Jan-2025].

[15] G.-B. Huang, Q.-Y. Zhu, and C.-K. Siew, "Extreme learning machine: Theory and applications," *Neurocomputing*, vol. 70, no. 1-3, pp. 489–501, 2006. [Accessed: 12-Jan-2025].

[16] R. S. Akinbo and O. A. Daramola, "Ensemble Machine Learning Algorithms for Prediction and Classification of Medical Images," *Artificial Intelligence*, IntechOpen, Dec. 22, 2021, doi: 10.5772/intechopen.100602. [Accessed: 18-Jan-2025].

[17] I. Goodfellow, Y. Bengio, and A. Courville, *Deep Learning*. Cambridge, MA: MIT Press, 2016. Available: https://www.deeplearningbook.org/. [Accessed: 12-Jan-2025].

[18] M. Alam, M. D. Samad, L. Vidyaratne, A. Glandon, and K. M. Iftekharuddin, "Survey on Deep Neural Networks in Speech and Vision Systems," *Neurocomputing*, vol. 417, pp. 302–321, 2020, doi: 10.1016/j.neucom.2020.07.092. [Accessed: 18-Jan-2025].

[19] A. Ajit, K. Acharya, and A. Samanta, "A review of convolutional neural networks," in *Proc. 2020 Int. Conf. Emerging Trends in Information Technology and Engineering (ic-ETITE)*, Vellore, India, 2020, pp. 1-5, doi: 10.1109/ic-ETITE47903.2020.049. [Accessed: 13-Jan-2025].

[20] P. Opěla, I. Schindler, P. Kawulok, R. Kawulok, S. Rusz, and M. Sauer, "Shallow and deep learning of an artificial neural network model describing a hot flow stress evolution: A comparative study," *Materials Today Communications*, vol. 31, p. 103491, 2022, doi: 10.1016/j.mtcomm.2022.103491. [Accessed: 18-Jan-2025].

[21] M. Zambra, A. Testolin, and M. Zorzi, "A developmental approach for training deep belief networks," *Cogn. Comput.*, vol. 15, pp. 103–120, 2023, doi: 10.1007/s12559-022-10085-5. [Accessed: 13-Jan-2025].

[22] V. Passricha and R. K. Aggarwal, "A comparative analysis of pooling strategies for convolutional neural network-based Hindi ASR," *Journal of

Ambient Intelligence and Humanized Computing*, vol. 11, pp. 675–691, 2020, doi: 10.1007/s12652-019-01325-y. [Accessed: 18-Jan-2025].

[23] M. Abdulghani, W. Walters, and K. Abed, "Enhancing the classification accuracy of EEG-informed inner speech decoder using multi-wavelet feature and support vector machine," *IEEE Access*, vol. PP, no. 99, pp. 1-1, Jan. 2024, doi: 10.1109/ACCESS.2024.3474854. [Accessed: 16-Jan-2025].

[24] G. E. Hinton, S. Osindero, and Y. W. Teh, "A fast learning algorithm for deep belief nets," *Neural Computation*, vol. 18, no. 7, pp. 1527–1554, Jul. 2006. DOI: 10.1162/neco.2006.18.7.1527.

[25] L. Xie, Y. Li, Z. Zhang, and L. Zhang, "Exploring extreme learning machine for decoding motor imagery EEG," *Front. Neurosci.*, vol. 15, p. 707826, 2021, doi: 10.3389/fnins.2021.707826. [Accessed: 16-Jan-2025].

[26] S. Anumanchipalli, J. S. Chartier, and E. J. Chang, "Speech synthesis from neural decoding of spoken sentences," *Nature*, vol. 568, pp. 493–498, 2019, doi: 10.1038/s41586-019-1119-1. [Accessed: 18-Jan-2025].

[27] Z. Zhang, X. Liu, L. Zhou, and Y. Zhao, "Shallow architectures for EEG decoding," *IEEE Trans. Neural Syst. Rehabil. Eng.*, vol. 27, no. 7, pp. 1444–1451, Jul. 2019, doi: 10.1109/TNSRE.2019.2912347. [Accessed: 18-Jan-2025].

[28] R. Schirrmeister, F. Springenberg, A. Fiederer, L. Wichmann, and J. Schultz, "Deep learning with convolutional neural networks for EEG decoding and visualization," *Front. Neurosci.*, vol. 11, p. 629, 2017, doi: 10.3389/fnins.2017.00629. [Accessed: 18-Jan-2025].

[29] A. Pudasaini, S. Lee, and M. Cho, "Deep belief networks for EEG-based emotion recognition: A survey," *IEEE Access*, vol. 8, pp. 110586–110597, 2020, doi: 10.1109/ACCESS.2020.3006905. [Accessed: 18-Jan-2025].

[30] X. Li, Y. Zhang, Z. Wang, and L. Xu, "Leveraging sinusoidal representation networks to predict fMRI signals from EEG," *IEEE Trans. Neural Syst. Rehabil. Eng.*, vol. 31, pp. 453–463, 2023, doi: 10.1109/TNSRE.2023.3054731. [Accessed: 18-Jan-2025].

[31] F. S. Liwicki, V. Gupta, R. Saini, K. De, and M. Liwicki, "Rethinking the Methods and Algorithms for Inner Speech Decoding and Making Them

Reproducible," *arXiv preprint arXiv:2103.16484*, 2021. [Online]. Available: https://arxiv.org/abs/2103.16484. [Accessed: 18-Jan-2025].

[32] M. Cooney, S. P. Kelly, and D. O'Neill, "Inner speech classification using EEG signals: A deep learning approach," *IEEE Access*, vol. 9, pp. 6578–6586, 2021, doi: 10.1109/ACCESS.2021.3053507. [Accessed: 18-Jan-2025].