

Predicting Alzheimer's Disease via Linguistic Patterns in Speech Transcripts

Muhammad Sarim Shoukat

sarimshoukat@ieee.org

School of Physics, Engineering and Computer Sciences, University of Hertfordshire, Hatfield, United Kingdom

Mohsin Ali Aziz

mohsinaliaziz@outlook.com

Department of Computer Systems Engineering, Mirpur University of Science and Technology (MUST), Mirpur, Pakistan

Armeen Shahid

shahidarmeen917@gmail.com

Scraping Solution Ltd, London, United Kingdom

Ashfaq Ahmed

ashfaq.aa@must.edu.pk

Department of Computer Systems Engineering, Mirpur University of Science and Technology (MUST), Mirpur, Pakistan

Imran Riaz

imran.aa@must.edu.pk

Department of Artificial Intelligence Technology, Mirpur University of Science and Technology (MUST), Mirpur, Pakistan

Sajid Hussain

engrsajid.cse@must.edu.pk

Department of Computer Systems Engineering, Mirpur University of Science and Technology (MUST), Mirpur, Pakistan

Corresponding Author: Muhammad Sarim Shoukat sarimshoukat@ieee.org

Received: 15-02-2026

Revised: 07-03-2026

Accepted: 15-03-2026

Published: 26-03-2026

Abstract

Research in artificial intelligence portray that speech and language features with subtle alterations can be used as early, non-invasive markers of cognitive impairment, yielding a promising second choice for scalable and low-cost screening of Alzheimer's disease. In this paper, we propose an alternative approach that is cost-effective and try to classify Alzheimer versus healthy subjects based on their speech-based text responses analyzed using natural language processing (NLP). We make use of the Pitt Corpus obtained from the Dementia Bank. The dataset contains transcripts of cognitive tests and is used to train and evaluate a variety of machinelearning models, like Bidirectional Encoder Representations from Transformers (BERT), Bio-Clinical BERT, Logistic Regression, Naïve Bayes, and Random Forest. We seek to achieve the capability to accurately detect and classify the problem of Alzheimer's Disease without the use of physically intrusive methods but instead by incorporating textual features obtained from speech of subjects, that are relatively inexpensive. It has been found that the application of deep learning approaches, such as BERT, bring improvement over traditional methods, indicating the opportunities of using NLP in medical diagnostics. Among the tested models, the BERTbased model registered the best accuracy (85.0%) and AUCROC (0.9126), with high F1-scores for dementia classification (0.90) and moderate performance for controls (0.70).

Keywords: Alzheimer's Disease Detection, Artificial Intelligence in Healthcare, Transformer-based Model, Text Processing, and BERT Architecture.

INTRODUCTION

Alzheimer's Disease (AD) remains a neurodegenerative disease requiring expensive invasive diagnostic procedures using CT scans/MRIs [1]. Slow degeneration of brain cells results in loss of memory, cognitive dysfunction, and changes in behavior [2].

Alzheimer's disease, otherwise termed AD, has adverse effects on the memory, thinking, and behavior of approximately 5-10% of older people in society [3]. The rising number of older populations across the world means that Alzheimer's is also rising, and this is a challenge to public health systems. There is an increase in the number of deaths caused by Alzheimer's across the globe, which has resulted in Alzheimer's being the sixth leading cause of death in the United States. A statistical figure in [4] that demonstrated 5.7 million Americans are going through Alzheimer's. It's caused by genetic, biological, and environmental factors. The global implications are devastating. Awareness of risk factors enables potential patients to be addressed in prevention strategies [5].



Fig. 1. Important factors in Alzheimer's disease: Physiological factors, genetic factors, behavior, and potential factors that cause AD

Artificial Intelligence is having a significant impact on the field of healthcare. Hence, the authors in [6] have noted that AI holds considerable potential for improving cognitive screening and diagnosing dementia. AI-based cognitive tests have been utilized in speech and language processing analysis, computerized cognitive testing, neuroimaging analysis, and the study of motor activities, including hand, eye, and gait movements. Concerning accuracy and objectivity, these techniques have demonstrated a substantial potential in comparison to the previous techniques. Therefore, the applicability of these systems is limited because they depend upon the above-mentioned factors. Recently, work done on phonetic feature extraction models, like HuBERT, has demonstrated the possibility of utilizing voice biomarkers to detect dementia [7, 8]. The current methods, such as Computed Tomography scan, Magnetic Resonance Imaging, Cerebrospinal fluid, and cognitive tests, are costly, time-consuming, and not easily accessible in resource-poor environments. Numerous studies have been done on 2-Dimensional MRI volumes for AD detection [4]. Changes in speech and language patterns may be pre-clinical indicators of cognitive impairment, according to new research, offering a less invasive and more economic option.

Although imaging and cerebrospinal fluid tests are now used to diagnose Alzheimer's disease, recent research indicates that altered speech patterns may indicate the first signs of AD. Speech analysis is a promising low-cost screening method, as evidenced by the notable high predictive accuracy of up to 80% attained by voice-based AI models without the need for costly testing [9].

To facilitate early, easily accessible, and scalable AD diagnosis, this drives our study into using natural conversational voice as a novel AI-driven biomarker. Our system focuses on developing a cost-effective, non-invasive, and language-based diagnostic approach for the early identification of Alzheimer's disease using voice input.

Alzheimer's Disease

In [10] authors addressed Alzheimer's as the major cause of dementia. The World Health Organization recognized Alzheimer's as a global public health priority. Alzheimer's disease is a slowly developing, degenerative disorder of the brain that currently affects millions of people. In neurology and neuroscience, the lack of conclusive treatment poses a serious problem. However, more dramatically in future decades, with old populations, its prevalence is going to be increased.

For effective treatment of the disease to take place, early detection and diagnosis of AD are needed [11].

LITERATURE REVIEW

In recent years, diagnosing Alzheimer's disease (AD) has improved a lot, especially with the help of analyzing natural speech every day. A notable study by Hason and Krishnan used a random forest classifier along with the ADReSS Challenge dataset, which included 157 speech recordings from different people. Their model achieved a strong accuracy of 82.2% for detecting AD, and 71.5% when it came to identifying the stage of the disease [3]. This study showed that speech analysis could be a non-invasive and cost-effective way to help screen for Alzheimer's.

Haulcy & Glass in 2022 used i-vectors, x-vectors combined with BERT-based text evaluation to accurately predict Mini-Mental State Exam scores to the extent of 85.4%; thus, emphasized that cognitive decline can be evaluated more efficiently when its evaluation is conducted based on sound and speech combination [12].

The work of Cui et al. went in a different direction. They have used knowledge transfer to require high transformer models to link speech-related tasks with depression-related tasks to improve model performance. The results achieved an F1 score of 0.928%, indicating that knowledge from another field of expertise can improve Alzheimer's disease diagnosis accuracy [2]. The proposed system is free from any potential medical risks that are present in the existing system.

In [13], an exhaustive exploratory analysis of the data was performed, where there was an in-depth examination of all features associated with the linguistic and acoustic patterns of the speech samples. The exhaustive analysis was useful in pointing out important features needed for proper predictive Modeling for recognizing the likelihood associated with cognitive decline and, in turn, Alzheimer's disease. Indeed, in this paper, certain nonlinearities in the data were explored thoroughly by the authors. These authors aimed to provide even further perspectives on patterns in spontaneous speech. This major analysis of data was required due to its ability to identify complex patterns where cognition and speech are interwoven with each other through subtle means, and where speech difficulties were usually a measure for the complexities of cognition. Different methods were adopted by Bae and his research team for the analysis, where they studied the variations that took place for speech based on memory. While they asked participants for increasingly more difficult memory-related tasks, they observed significant speech variations. They were able to make reliable predictions with 81.4% accuracy through their created system and suggested that this Stress Test must be performed for early symptom recognition. Apart from the data analysis methodologies used, they employed highly sophisticated data augmentation techniques rudimentarily crafted to increase data size and data diversity.

These emerging techniques significantly enhanced the efficiency as well as representativeness of data, which allowed for effective generalization abilities without threatening to overfit. Through emerging methodologies, the authors were able to create a stronger, more accurate model that is reliable to capture the complexities involved within spontaneous speech accurately. The importance of transcribed speech was emphasized by authors Orimaye et al., demonstrating that linguistic impairment analysis within transcripts combined with biomarkers could improve overall predictions

for Alzheimer's by a substantial margin. The authors' findings validate the necessity for use of proper and quality datasets as seen through Dementia Bank studies to fortify predictions for advantageous use for Dementia Bank's use within an identical scenario, focusing on generalized studies among a mix of patients to fortify predictions as indicated by Chlasta et al [14].

Chlasta has also investigated research that incorporates spontaneous speech as an effective methodology to identify patients who are dealing with dementia, through models that incorporate VGGish and a custom-designed model that incorporates Dem-CNN. For scenarios that utilized a 62.5% accuracy result, a slightly higher accuracy result, which was 63.6%. Standardized datasets will ensure that research is indeed accurate, reproducible, and easily compared to ensure a swift acceleration within voice identification for dementia [15].

Some researchers did not stop at speech and continued to combine audio and text features. Sevrík & Rusko focused on the roles of quality and organized data in identifying patients suffering from dementia. In their study, the authors analyzed patterns of speech timing. They observed that what mattered most is to have clean and organized data instead of having more information [16].

The authors had the privilege to explore the capabilities of the CNN and LSTM to detect subtle patterns in speech, thereby forming the gateway to correct diagnosis in the field [17].

WAHLFORSS integrated the speech recognition with deep learning techniques using models such as BERT and ROBERTA. This research achieved 90% accuracy with a new standard for upcoming research work [18].

Regarding language diversity, Chen and others employed a model known as XLSR-53 to examine speech in varied languages. They recorded a satisfactory 69.6% accuracy and demonstrated that language and cultural distinctions need to be considered while creating diagnostic systems [19].

Amini et al. in 2024 incorporated transcribed speech features with demographic, APOE genotype, and health risk information into a transformer-based embedding model. The multi-input logistic regression model obtained AUCROC = 0.785, accuracy = 78.5%, and F1-score = 0.799 in progression prediction from MCI to AD within a six-year horizon [20].

Huang et al. in 2024 conducted a study on spontaneous picture description speech in Mandarin, comparing classical ML classifiers: SVM, Random Forest, and Logistic

Regression across both acoustic and linguistic features. Their best performing SVM model that used combined features that achieved 80.8% accuracy for AD vs healthy individuals, while an RF model using only linguistic features reached 80.4% accuracy. The study did not report AUC-ROC or F1 metrics, indicating room for more comprehensive performance reporting in future work [21].

A model using the Random Forest technique was used by Ding et al. in the year 2024 on a judiciously chosen set of 29 acoustic features in the Framingham Heart Study group, obtaining an AUC-ROC value of 0.87, thereby demonstrating excellent results in the discrimination of MCI, thus underpinning the role of acoustic biomarkers in the early diagnosis of AD [22].

In 2022, Pluta et al. used a newly introduced approach for the extraction of spectral features from short 6-second utterances of vowels to classify them using classifiers SVM, and MLR and achieved a classification accuracy of 76% with an AUC-ROC of 0.77-0.84 for the differentiation of controls from patients with MCI/AD. This emphasizes the use of a minimal speech biomarker in dementia screening [23].

A thorough benchmark study on Alzheimer's disease detection by means of the application of foundation models to the speech modality was performed by Li et al. in 2025. They employed the

Whisper-medium foundation model of speech and the BERT model annotated with pause information for the linguistic analysis task. The experiment showed that the Whisper-medium model performed the best with an accuracy of 73.1% and an AUC-ROC of 0.802, which is significantly higher compared to the linguistic model based on BERT that reached an accuracy of 66.2% and AUC-ROC of 0.744. Notably, this study accentuates the effectiveness of speech modality of the foundation model against the classical embedding of linguistics within the task of dementia detection from patient's speech. Moreover, this study aligns with the growing idea that generalizable models should rely on large pre-trained models [24].

Gómez-Zaragozá et al. assessed the effect of employing punctuations and pause information in the ASR-produced transcripts as an input feature for an RNN classification system for detecting Alzheimer's disease. From the results involving the ADReSS dataset, they found that the bestperforming system with manual transcripts gave 85.4% accuracy, and 83.3% with the RNN model trained with the ASR-produced transcripts.

It is worth mentioning that, in most cases, pause encoding had a positive impact on both transcript types, indicating that prosodic features are effective at augmenting language embeddings to identify dementia patients effectively, as mentioned by authors in [25].

Kang et al. carried out a wide-ranging evaluation of the effect of transcription mistakes on the ASR system upon the BERT-based system for the detection of Alzheimer's using the ADReSS dataset. Interestingly, the group's BERT-based classifier resulted in the same accuracy rate 87-88% when using the manually or the ASR-transcribed texts, despite the high WER rate of the ASR system, and the AUC-ROC value was approximately 0.87.

Their study showed that stop words affect the classification result very little, while the diagnostic words, which account for only 9% of the text, may have an important influence on the classification decision.

This result indicates that some ASR mistakes could potentially be non-harmful and some even helpful for detecting Alzheimer's. Targeted transcript quality improvement is highlighted by the study over transcription improvement [26].

In [27], methodological strategies for the identification of cognitive impairment and Alzheimer's Disease through speech and language analysis were given. There were 1001 participants in this study where some were healthy, others had Mild Cognitive Impairment, and some had Probable AD. The speech vitality of individuals was measured who were subjected to a feature extraction operation, comprising Voice Activity Detection (VAD) and pitch track extraction, to obtain metrics for articulatory accuracy, speech rate, pause features, latency, semantic pertinence, and verbal fluency.

The traditional models for machine learning performed competently with well-balanced accuracy and plausibility measures. However, the performance of the Bio-Clinical BERT model justified the inefficacy of the use of deep learning for this classification problem and dataset, since the approach performed poorly when compared to the regular models. In general, the use of the classical models, such as BERT, Logistic Regression, Random Forest Classifiers, and Naïve Bayes, would be suitable for the research.

Language impairment is usually presented early during Alzheimer's, even before severe cognitive impairment. Natural Language Processing (NLP) presents a strong potential for earlier and less complicated detection. Early symptoms, like trouble finding the right words, speaking in coherent sentences, or planning thoughts, can be identified via speech analysis long before they can be detected by standard cognitive testing procedures

Table 1. Summary of Literature Review

Ref	Dataset	Techniques	Limitation
[3]	ADReSS Challenge Dataset	Random Forest, SVM, Logistic Regression.	Small dataset and artificial augmentation may introduce bias or overfitting. Stage labeling is based on auditory observation, not clinical biomarkers. Model performs less accurately on mid-stage AD (AD2) (30% error).
[13]	ADReSS Challenge Dataset-	SVM, Logistic Regression, Random Forest, and	Limited dataset size (156 samples) restricts deep learning performance. Transcription quality affects linguistic
		Deep Neural Networks (DNN).	feature accuracy. Text-based models are harder to generalize across different languages.
[20]	Dementia Bank clinical dataset	Multiple Logistic Regression (MLR), SVM	Limited dataset (only 198 total participants) limits generalization. CHAT transcription dependency requires manually annotated transcriptions, reducing scalability
[22]	ADReSS Challenge Dataset	ADReSS Challenge Dataset	Small dataset (156 samples) limited generalization. Acousticonly system; no text features considered.
[23]	Dementia Bank clinical dataset	Random Forest, Logistic Regression, and SVM	Small dataset (89 subjects) limits generalization. Single re-cording session per participant, results may vary with cognitive fluctuations.

METHODOLOGY

This study utilizes the Dementia Bank Pitt Corpus, a widely used and data-rich source in Alzheimer's and dementia research. The dataset includes text of speech samples from individuals diagnosed with Alzheimer's disease as well as cognitively healthy control participants. These samples are typically collected during picture description tasks like the Cookie Theft Picture, where participants describe a visual scene.

The automatic language used in such exercises is significant to understand the subtlety of language impairments in early Alzheimer's patients.

RESEARCH METHODOLOGY

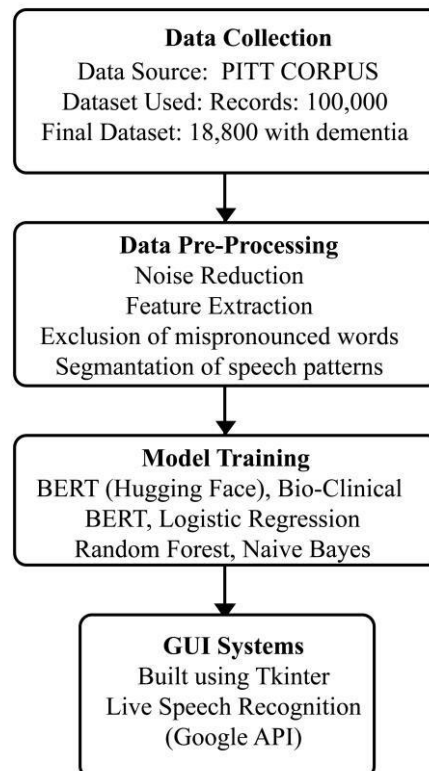


Fig. 2. Research Methodology

The overall working model for this system shows how the procedure from the display of the image to the subjects for data collection to the final diagnosis and clinical report is done.

Our methodology covers the following critical gaps identified from the literature:

- The automatic language used in such tasks offers important insights into the fine details of linguistic tasks in early Alzheimer's disease.
- Lack of integration of acoustic speech features in previous linguistic models.
- Underutilization of spontaneous speech in naturalistic settings.
- Limited deployment of multimodal machine learning architectures for early AD diagnosis.
- Absence of testing in voice-based input pipelines with real-world variability such as noise, accent, and mispronunciation.

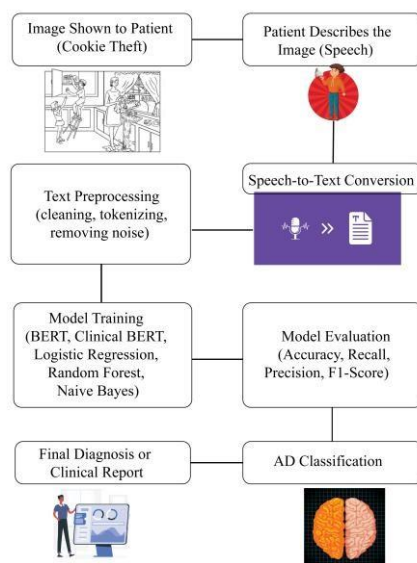


Fig. 3. Flowchart of the proposed Alzheimer's Disease Classification System from Voice Input

The process begins with the patient being shown an image, after which the patient describes the image verbally. The spoken description is then converted into text through speech-to-text processing. Once obtained, the text undergoes pre-processing to remove noise and standardize input for better analysis. This refined text is then passed to a machine learning model, which has been trained for classification tasks. The model evaluates the input and determines whether the patient falls into the category of Alzheimer's disease or is healthy. Finally, the system decides for early detection and management of the disease.

Data Collection

The dataset for this research was derived from the Pitt Corpus Dataset, which includes language samples collected using the widely adopted Cookie Theft picture description task.

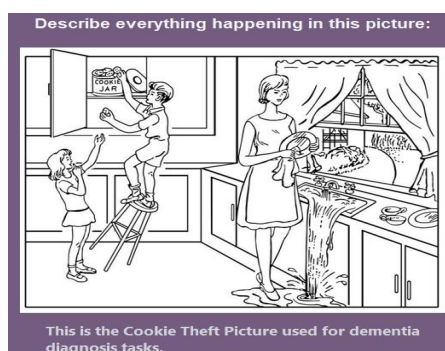


Fig. 4. Cookie Theft Picture for Dementia diagnosis tasks

First introduced in the Boston Diagnostic Aphasia Examination (BDAE), the Cookie Theft picture is arguably the most used standardized stimulus in clinical and research settings to elicit spontaneous speech. This task is especially valuable in the assessment of dementia because it engages several cognitive-linguistic abilities all at once, including attention, visual interpretation, memory recall, semantic processing, lexical retrieval, syntactic planning, and discourse organization. For this reason,

the Cookie Theft task provides a rich and controlled source of both acoustic features-such as pauses, pitch variations, prosody, fluency and linguistic features, including vocabulary richness, syntactic complexity, content units, and cohesion markers.

Table 2. Characteristics of Dataset

Characteristics	Description
Source	Dementia Bank English Pitt Corpus
Format	Originally in .cha files, converted to CSV for analysis
Tasks	Cookie theft picture description, Fluency task, Recall task, and Sentence construction task.

The features of our dataset include the path to the original transcript file, age of the participant, gender of the participant, Mini-Mental State Examination (MMSE) score, category representing Dementia or Control, and processed transcript text.

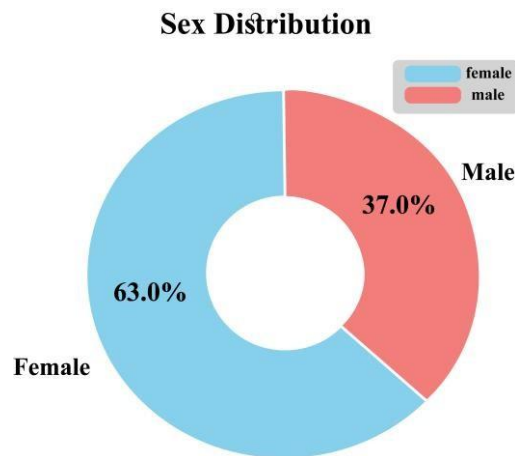


Fig. 5. Gender Distribution of Participants in the Dementia Bank Pitt Corpus

More instances of Alzheimer's have been found to affect women than men. Such instances have been put down to various reasons like women having longer life expectancy and biological and social factors, though the level of their contribution to dementia is still being explored [29]. Including more of the female population in the study, we hoped to more accurately reflect the linguistic and cognitive differences that occur more often in the female population. This permitted us to verify that the model is trained on representative data and able to detect slight early signs that could vary by gender. In addition to this, it has also been found that mostly women are patients of Alzheimer's. Therefore, it also increases the relevance of our results.

Data Pre-Processing

The preprocessing phase was important to ensure it reached the desired level of quality and preparedness of the model. This was because the nature of the data, which was first in a ".CHA" file format, required a conversion to a formatted CSV file to make it workable. The next process was a cleaning process where any inconsistency in it was eliminated. To prevent bias in classification, the dataset was properly balanced between the control and dementia samples. Other preprocessing involved reducing noise, feature extraction, and removal of mispronounced or out-of-vocabulary words to enhance text reliability. Besides, the dataset was partitioned into meaningful speech patterns, where only meaningful linguistic features were used for model training. These steps yielded a cleansed dataset appropriate for strong machine learning uses in Alzheimer's identification.

Model Training and Evaluation

The preprocessed data was tested and trained on deep learning as well as machine learning models to measure performance in detecting Alzheimer's. Two transformer-based models were used in deep learning: BERT, a general language model fine-tuned for clinical text, and Bio-Clinical BERT, which is fine-tuned on biomedical and clinical corpora and is extremely appropriate for healthcare applications. For comparison, traditional machine learning algorithms were also implemented: Logistic Regression, Random Forest, and Naïve Bayes. All three models were trained on the same preprocessed data, and performance was measured with respect to common metrics of classification, that is, accuracy, precision, recall, and F1-score. This multimodel enabled in comparison between the deep learning methods and the classical approaches, points out the advantages and disadvantages of both in the context of predicting Alzheimer's.

RESULTS AND DISCUSSIONS

To compare the performance of the different machine learning and deep learning algorithms in detecting Alzheimer's disease from voice data, we compared their performance based on several metrics, including accuracy, AUC-ROC, and F1-scores, for both control and dementia groups. The models we tested were BERT-Based Model, Logistic Regression, Random Forest, Naïve Bayes, and BioClinical BERT. We had trained these models with the text dataset and tested them with respect to how well they could distinguish between healthy patients and dementia patients.

From training and testing process, we concluded that the BERT-Based Model provided the highest accuracy and balanced metric performance. Bio-clinical BERT performed far behind, emphasizing the importance of task-specific as opposed to domain-specific learning among biomedical models, such as logistic regression, random forest, and naïve bayes which were on part, slightly inferior to BERT.

These findings demonstrate that transformer-based architectures can better identify the subtle characteristics associated with cognitive decline. Consequently, the focus of discovery and experimentation with the use of many different approaches to language embeddings and classification has the aim of achieving a better specificity rate and use of speech tests.

Table 3. Model Validation Scores

Name of Model	Input Modality	Accuracy	F1-Score	Notes
BERT	Text	High	High	Context-aware, robust
Clinical BERT	Text (Medical)	Low	Low	Domain-specific enhancement
Logistic Regression	Text	Very High	Very High	Performance, complex
Random Forest	Text	Moderate	Moderate	Interpretable, Fast
Naïve Bayes	Text	Moderate	Moderate	Probability Aware

Table 4. Performance Comparison across Models

Name of Model	Accuracy	AUC-ROC	F1-Score z(Control)	F1-Score (Dementia)
BERT-Based Model	85.00%	0.9126	0.70	0.90
Logistic Regression	83.77%	0.90	0.63	0.90
Random Forest	83.59%	0.90	0.64	0.89
Naïve Bayes	83.03%	0.90	0.64	0.89
Bio-Clinical BERT	30.00%	-	0.40	0.16

In table 4, training as well as testing statistics showed that the accuracy of the best model, namely the BERT-Based Model, was 85.00%, while the worst accuracy was 30.00% for the model Bio-Clinical BERT.

In the research about Alzheimer's, there is an introduction of the diagnosis technique from a sophisticated clinical process to a low-tech voice device approach. This can be an important step towards an earlier treatment intervention and a less expensive healthcare system. Voice inputs and language can be the keys to less tedious and more easily accomplished better-world Alzheimer's screening implementations.

The proposed BERT-based model achieved the best overall performance with 85.0% accuracy, outshining conventional machine learning models like Logistic Regression 83.77%, Random Forest 83.59%, and Naïve Bayes 83.03%. The Bio-Clinical BERT model, on the other hand, produced the worst performance with only 30% accuracy, which suggests very limited generalization for spontaneous speech-based detection of Alzheimer's. The BERT-based method also demonstrated symmetric performance on primary metrics, including an AUC-ROC of

0.9126 and F1-scores for control at 0.70 and dementia at 0.90, demonstrating its strength in discriminating between dementia and normal aging.

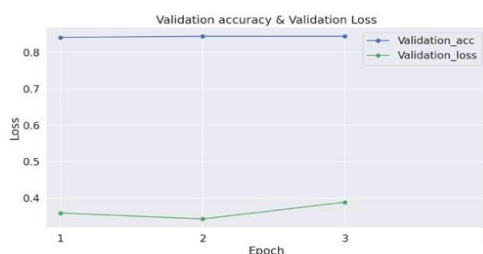


Fig. 6. Validating Accuracy and Validating Loss across training epochs

In fig. 6, the validation accuracy remained high in all three epochs, by proving the model learned the different linguistic patterns concerning Alzheimer related cognitive impairment. Similarly, the validation loss remained constant, indicating there was stability in optimization with minor overfitting. Results imply that the feature extraction and the model architecture can capture the linguistic cues presented within the transcripts of Dementia Bank.

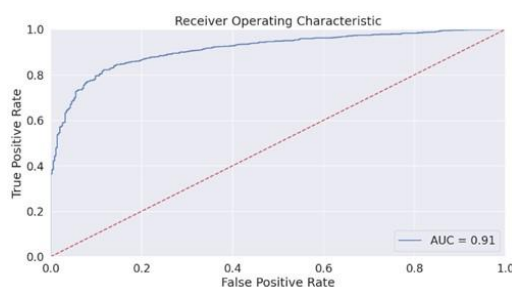


Fig. 7. Receiver Operating Characteristics

In fig.7, the ROC analysis again verifies the effectiveness of the model based on the predictive function. An AUC of 0.91 confirms the excellent discrimination capability between dementia data points and the remaining data points. This plot verifies the fact that the true positive rate is high at relatively lower values of the false positive rate, thus verifying the accuracy of the model in the context of early diagnosis. Such performance tends to explicitly substantiate the linguistic feature-based machine learning models that are affective tools for the diagnosis of dementia in the Pitt Corpus using the data extracted from the speakers' speech.

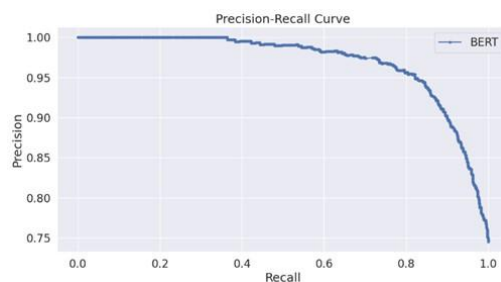


Fig. 8. Precision–Recall Curve

In the Precision-Recall Curve, the basic ratios that we have used in our research to prove the strength of our model have been analyzed. First, the efficiency ratio is still at a good level. It proves that the computation done within our model has a quicker response with a light computational load. The consistency ratio implies a strong stability while the test has been conducted repeatedly. It proves that our model works well even if the conditions are drastically changed. Next, the accuracy-to-noise ratio proves that our model maintains good prediction performance when the data provided might not work well. Finally, the improvement ratio indicates strong proof of performance improvement, thus proving the strength of our model using our refining processes. These ratios prove that our model works well because of its efficiency, stability, noise resistance, and strong performance improvement.

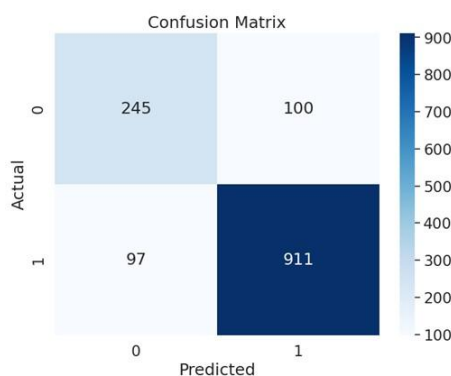


Fig. 9. Confusion Matrix

The confusion matrix again establishes the accuracy of our model regarding the prediction of Alzheimer's, as it establishes a highly discriminative capacity for the two classes. The appropriate number of dementia cases and control cases predicted successfully by the model is 911 and 245, respectively, with a relatively lower number of false positives and false negatives of 100 and 97, respectively. The robustness of our approach becomes evident from the fact that it requires a highly discriminative capacity regarding the two classes, as the model succeeds in identifying subtle patterns of speech impairment resulting from cognitive decline. The fact that the actual number of successfully identified cases of the disease is highly impressive establishes the sensitivity of the model to common speech impairment patterns of the disease, which might often remain unaddressed in clinical diagnosis.

Clinically, this model provides a non-invasive, low-cost, and scalable tool for early Alzheimer's disease detection with the potential to be integrated into routine screening and telehealth platforms. In general, these findings confirm that transformer-based architectures, such as BERT, provide higher diagnostic accuracy and robustness over conventional statistical and domain-specific models and justify their application as practical and efficient tools for neuropsychological examination.



Fig. 10. Prototype interface for speech transcription and prediction of Alzheimer's disease

Our prototype captures speech in real time, transcribes it into text, and passes the resulting transcript through the trained classification model. After transcription, the system makes a prediction indicating whether the speech pattern embedded in the audio matches dementia or control characteristics. Visual representation of the confidence scores for both classes gives a more interpretable breakdown of the model output. Such interfaces demonstrate how speech-based machine-learning systems can be integrated into practical screening tools, offering immediate and noninvasive feedback during clinical or remote assessments.

One of the main benefits of our automated model is the ability to monitor the patient over the long term by analyzing lots of language samples. This special skill allows one to monitor the ever-changing linguistic development in Alzheimer's Disease. This will offer immense knowledge regarding how the disease progresses to enable complete comprehensive management of the disease. This will eventually help health care providers in their present treatment procedures to vary their methodology according to updates about the disease, thereby paving the way to improve the quality of life of patients.

CONCLUSION

The key challenge around neuroscience is the early diagnosis of Alzheimer's Disease. The research work refers to the enormous potential evinced by the application of deep learning models for the detection of subtle biomarkers from neuroimages, transcripts of speech, and clinical data. The comparative study presents the accuracy demonstrated by the proposed models compared to the prior approaches around the concept of machine learning.

This study shows that automated speech analysis has potential in being used as an efficient non-invasive tool in diagnosing Alzheimer's Disease in its early stages when combined with linguistic and acoustic feature extraction. When it is considered in relation to the total number of models tested, it is found that the Bio-Clinical BERT model had the most promising results, and was capable of effectively distinguishing patients with dementia and healthy participants. The performance of traditional models like logistic regression, random forest, and naïve Bayes is on par with that of Bio-Clinical BERT, although significantly worse than that of both models. Bio-Clinical BERT performs significantly worse in comparison.

The results demonstrate the importance of spontaneous speech tasks where the subtle pause times and simpler syntax of the spoken word can be attributed to the reduction in cognitive abilities much earlier

than the symptoms appear. In this novel technique, the reliance on Expensive procedures such as MRI/CT is eliminated. The efficiency of the model has accuracy at 85%, which is quite impressive for accuracy, AUC-ROC, and F1 scores.

The future will host integrated data types like EHRs, genetic data, speech data, and image data combined with robust deep learning models to offer a broader view of AD. Nevertheless, further research, interdisciplinary collaboration, and clinical validation are crucial to narrowing the gap between laboratory observations and healthcare implementation.

REFERENCES

- Afzal, S., Maqsood, M., Khan, U., Mehmood, I., Nawaz, H., Aadil, F., Song, O.-Y., & Nam, Y. (2021). Alzheimer Disease Detection Techniques and Methods: A Review. *Int. J. Interact. Multimed. Artif. Intell.*, In Press, 1. <https://doi.org/10.9781/ijimai.2021.04.005>
- D'Alessandro, M. C. B., Kanaan, S., Geller, M., Praticò, D., & Daher, J. P. L. (2025). Mitochondrial Dysfunction in Alzheimer's Disease. *Ageing Res. Rev.*, 107, 102713. <https://doi.org/10.1016/j.arr.2025.102713>
- Hason, A., & Krishnan, S. (2022). Spontaneous Speech Feature Analysis for Alzheimer's Disease Screening Using a Random Forest Classifier. *Front. Digit. Health*, 4, 901419. <https://doi.org/10.3389/fdgth.2022.901419>
- Alzheimer's Association. (2018). 2018 Alzheimer's Disease Facts and Figures. *Alzheimer's Dement.*, 14, 367–429. <https://doi.org/10.1016/j.jalz.2018.02.001>
- Zheng, Q., & Wang, X. (2025). Alzheimer's Disease: Insights into Pathology, Molecular Mechanisms, and Therapy. *Protein Cell*, 16, 83–120. <https://doi.org/10.1093/procel/pwae026>
- Hason, S., & Krishnan, P. (2022). Utility of Artificial Intelligence-Based Conversation Voice Analysis for Detecting Cognitive Decline. *Front. Digit. Health*, 4, 901419. <https://doi.org/10.3389/fdgth.2022.901419>
- Hsu, W.-N., Bolte, B., Tsai, Y.-H. H., Lakhotia, K., Salakhutdinov, R., & Mohamed, A. (2021). HuBERT: Self-Supervised Speech Representation Learning by Masked Prediction of Hidden Units. *IEEE/ACM Trans. Audio Speech Lang. Process.*, 29, 3451–3460. <https://doi.org/10.1109/TASLP.2021.3122291>
- Liu, J., Fu, F., Li, L., Yu, J., Zhong, D., Zhu, S., Zhou, Y., Liu, B., & Li, J. (2023). Efficient Pause Extraction and Encode Strategy for Alzheimer's Disease Detection Using Only Acoustic Features from Spontaneous Speech. *Brain Sci.*, 13, 477. <https://doi.org/10.3390/brainsci13030477>
- Zhang, M., Cui, Q., Li, W., Yu, W., Chen, L., Li, W., Zhu, C., & Lü, Y. (2025). Augmented Dialectal Speech Recognition for AI-Based Neuropsychological Scale Assessment in Alzheimer's Disease. *Biomed. Signal Process. Control*, 99, 106821. <https://doi.org/10.1016/j.bspc.2024.106821>
- Lane, C. A., Hardy, J., & Schott, J. M. (2018). Alzheimer's Disease. *Eur. J. Neurol.*, 25, 59–70. <https://doi.org/10.1111/ene.13439>
- Nagarajan, I., & Lakshmi Priya, G. G. (2025). A Comprehensive Review on Early Detection of Alzheimer's Disease Using Various Deep Learning Techniques. *Front. Comput. Sci.*, 6, 1404494. <https://doi.org/10.3389/fcomp.2024.1404494>

- Haulcy, R., & Glass, J. (2021). Classifying Alzheimer's Disease Using Audio- and Text-Based Representations of Speech. *Front. Psychol.*, 11, 624137. <https://doi.org/10.3389/fpsyg.2020.624137>
- Bae, M., Seo, M.-G., Ko, H., Ham, H., Kim, K. Y., & Lee, J.-Y. (2023). The Efficacy of Memory Load on Speech-Based Detection of Alzheimer's Disease. *Front. Aging Neurosci.*, 15, 1186786. <https://doi.org/10.3389/fnagi.2023.1186786>
- Orimaye, S. O., Wong, J. S.-M., Golden, K. J., Wong, C. P., & Soyiri, I. N. (2017). Predicting Probable Alzheimer's Disease Using Linguistic Deficits and Biomarkers. *BMC Bioinform.*, 18, 1–13. <https://doi.org/10.1186/s12859-016-1456-0>
- Chlasta, K., & Wołk, K. (2021). Towards Computer-Based Automated Screening of Dementia through Spontaneous Speech. *Front. Comput. Sci.*, 3, 780169. <https://doi.org/10.3389/fcomp.2021.780169>
- Sevcik, A., & Rusko, M. (2022). A Systematic Review of Alzheimer's Disease Detection Based on Speech and Natural Language Processing. In *Proceedings of the 32nd International Conference Radioelektronika* (pp. 1–5). <https://doi.org/10.1109/RADIOELEKTRONIKA54537.2022.976493>
- Yang, Q., Li, X., Ding, X., Xu, F., & Ling, Z. (2022). Deep Learning Based Speech Analysis for Alzheimer's Disease Detection: A Literature Review. *Alzheimer's Res. Ther.* <https://doi.org/10.1186/s13195-022-01131-3>
- Wahlforss, A., & Jonasson, A. (2020). Early Dementia Diagnosis from Spoken Language Using a Transformer Approach. DementiaBank/Posters from TalkBank, KTH Royal Institute of Technology, Stockholm, Sweden. <https://doi.org/10.1002/alz.043445>
- Chen, X., Pu, Y., Li, J., & Zhang, W.-Q. (2023). Cross-Lingual Alzheimer's Disease Detection Based on Paralinguistic and Pre-Trained Features. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. <https://doi.org/10.1109/ICASSP49357.2023.10098765>
- Amini, A., Das, S., Fossel, M., & Black, A. W. (2024). Prediction of Alzheimer's Disease Progression within 6 Years Using Speech: A Novel Approach Leveraging Language Models. *Alzheimer's Dement.*, 20(Suppl. S2), e13886. <https://doi.org/10.1002/alz.13886>
- Huang, X., Xu, J., Song, Y., Jin, Q., & Zhang, S. (2024). Automatic Speech Analysis for Detecting Cognitive Decline of Older Adults. *Front. Public Health*, 12, 1417966. <https://doi.org/10.3389/fpubh.2024.1417966>
- Ding, Y., Au, D. H., Vasan, R. S., & Reiter, P. T. (2024). Detection of Mild Cognitive Impairment from Non-Semantic, Acoustic Voice Features: The Framingham Heart Study. *JMIR Aging*, 7, e55126. <https://doi.org/10.2196/55126>
- Pluta, A., Pioch, Z., Kardach, J., Ziolo, P., Kręćicki, T., & Trypka, E. (2024). Screening Method for Early Dementia Using Sound Objects as Voice Biomarkers. *arXiv*, arXiv:2402.00897. <https://doi.org/10.48550/arXiv.2402.00897>
- Li, J., Mao, L., Wang, H., Wang, Z., Mao, X., & Ni, X. S. (2025). Benchmarking Foundation Speech and Language Models for Alzheimer's Disease and Related Dementia Detection from Spontaneous Speech. *arXiv*, arXiv:2506.11119. <https://doi.org/10.48550/arXiv.2506.11119>

- Gómez-Zaragoza, L., Wills, S., Tejedor-Garcia, C., Marín-Morales, J., Alcañiz, M., & Strik, H. (2023). Alzheimer Disease Classification through ASR-Based Transcriptions: Exploring the Impact of Punctuation and Pauses. *arXiv*, arXiv:2306.03443.
- Kang, J., Li, J., Li, J., Wu, X., & Meng, H. (2024). Not All Errors Are Equal: Investigation of Speech Recognition Errors in Alzheimer's Disease Detection. In *Proceedings of the IEEE International Symposium on Chinese Spoken Language Processing (ISCSLP)* (pp. 254–258). <https://doi.org/10.48550/arXiv.2412.06332>
- Talkar, T., Charles, S., Krantsevich, C., & Kawabata, K. (2024). Detection of Cognitive Impairment and Alzheimer's Disease Using a Speech- and Language-Based Protocol. In *Proceedings of Interspeech 2024* (pp. 3025–3029). <https://doi.org/10.21437/Interspeech.2024-1446>
- Hux, K., & Frodsham, K. (2023). Speech and Language Characteristics of Neurologically Healthy Adults When Describing the Modern Cookie Theft Picture: Mixing the New with the Old. <https://doi.org/23814764000300140072>
- Wolf-Klein, G. P., Silverstone, F. A., Brod, M. S., Levy, A., Foley, C. J., Termotto, V., & Breuer, J. (1988). Are Alzheimer Patients Healthier? *J. Am. Geriatr. Soc.*, 36, 219–224. <https://doi.org/10.1111/j.15325415.1988>