Early Alzheimer's Detection Through Voice Analysis: Harnessing Locally Deployable LLMs via *ADetectoLocum*, a privacy-preserving diagnostic system

Genevieve A. Mortensen, B.S.[1] and Rui Zhu, Ph.D.[2]

[1]Indiana University, Bloomington, Indiana, USA, [2]Yale University, New Haven, Connecticut, USA

Abstract

Diagnosing Alzheimer's Disease (AD) early and cost-effectively is crucial. Recent advancements in Large Language Models (LLMs) like ChatGPT have made accurate, affordable AD detection feasible. Yet, HIPAA compliance and the challenge of integrating these models into hospital systems limit their use. Addressing these constraints, we introduce ADetectoLocum, an open-source LLM equipped model designed for AD risk detection within hospital environments. This model evaluates AD risk through spontaneous patient speech, enhancing diagnostic processes without external data exchange. Our approach secures local deployment and significantly surpasses previous models in predictive accuracy for AD detection, especially in early-stage identification. ADetectoLocum therefore offers a reliable solution for AD diagnostics in healthcare institutions.

Keywords: *Machine learning, generative AI, predictive modeling, clinical decision support, translational data science interventions, data security and privacy*

Introduction

Alzheimer's Disease (AD) is a progressive neurodegenerative disorder that significantly impairs memory, cognition, and functioning, ultimately leading to death. It is the most common cause of dementia among older adults, affecting millions worldwide[1]. Traditionally, AD is diagnosed through a combination of clinical evaluations, including neurological assessments, mental status testing, and physical examinations, often supplemented by imaging tests such as MRI or CT scans to rule out other causes of dementia[2].

The early detection of AD presents a considerable challenge due to the gradual onset of symptoms and the overlap with normal aging processes. Early symptoms, such as forgetfulness and mild confusion, are frequently dismissed as normal aging, leading to delayed diagnosis[3]. The difficulty lies not only in distinguishing early-stage AD from normal aging but also in the current diagnostic methods' reliance on observable cognitive decline, which often signifies that the disease has already progressed[3].

Despite these challenges, early detection is of paramount importance. Identifying AD at an earlier stage opens the door to potential treatments that can slow the progression of the disease, improve quality of life, and extend independence[4]. Current treatment options for early-stage AD include pharmacological interventions aimed at managing symptoms, as well as non-pharmacological approaches such as cognitive therapy and lifestyle modifications[5]. Moreover, early detection allows patients and families to plan for the future, including making care arrangements and addressing legal and financial issues, while the patient can still participate in decision-making processes[4]. Therefore, innovative approaches that enable earlier and more accurate detection of AD are crucial in the fight against this debilitating disease.

While AD poses significant diagnostic and treatment challenges due to its gradual onset and complex nature, the evolution of technology in the medical field offers promising avenues for addressing these obstacles. Among these advancements, Large Language Models (LLMs) like ChatGPT stand at the forefront, heralding a new era of medical innovation[6, 7]. These sophisticated AI tools have the potential to revolutionize the early detection and management of diseases such as AD[7, 8, 9]. By leveraging the vast capabilities of LLMs to analyze and interpret medical data, healthcare professionals can gain insights into subtle patterns and indicators of AD that may not be evident through traditional diagnostic methods. For example, LLMs can process extensive patient dialogue or written texts to identify linguistic anomalies or changes over time that may suggest early cognitive decline, providing a non-invasive and efficient tool for early AD detection[8]. Beyond diagnostics, LLMs offer support in personalized patient care and treatment planning, drawing from a wealth of medical research and data to suggest tailored interventions[10, 11, 12].

Leveraging LLMs for the early detection of AD signals a promising approach, yet it confronts numerous practical deployment challenges. Most notably, the current generation of powerful LLMs, such as ChatGPT, does not comply with the Health Insurance Portability and Accountability Act (HIPAA), a critical consideration for applications in healthcare[13, 14, 15, 16]. Furthermore, even those LLMs that are HIPAA-compliant are not open-source due to commercial considerations, limiting their accessibility and adaptability for use in hospital settings[13, 14]. Hospitals' servers are typically highly secure and closed systems, necessitating local deployment of any technological solution, a requirement that existing commercial LLMs cannot fulfill[13]. This significant gap hinders the utilization of the advanced capabilities of LLMs within healthcare facilities.

In recognition of this gap, this paper explores the development and fine-tuning of an open-source LLM specifically for the task of detecting AD through patient speech. Our model, *ADetectoLocum*, not only addresses the critical need for HIPAA-compliant, locally deployable solutions but also demonstrates the potential to achieve superior performance in AD detection compared to commercial LLMs such as GPT-3.5 and GPT-4[6, 17]. By adapting an open-source LLM to this specific medical application, we bridge the divide between cutting-edge AI technology and practical healthcare implementation, offering a viable path forward for harnessing the power of LLMs in the fight against AD.

Our contributions are threefold and underscore significant advancements in the detection of AD through *ADetectoLocum*:

1. **Enhanced Accuracy**: *ADetectoLocum* outperforms State Of The Art (SOTA) models in predicting AD, achieving an accuracy improvement of at least 2% - 5%. This marks a significant leap forward in the reliability of AD diagnostic tools, offering more precise assessments.
2. **Local Deployment Compatibility**: Unlike earlier models that rely on accessing external GPT APIs, *ADetectoLocum* is fully operational with local models. This intrinsic characteristic makes it exceptionally suited for deployment in the restrictive and privacy-sensitive environments of hospitals, ensuring patient data remains secure and within the premises.
3. **Superior Early Detection**: When it comes to identifying AD in its early stages, *ADetectoLocum* demonstrates superior performance compared to existing SOTA models. Early detection is crucial for effective AD management and treatment, and our model provides a significant advantage in this critical aspect, potentially transforming patient outcomes through earlier intervention.

Background

Challenge of AD Detection

The detection of AD presents a significant challenge due to its subtle onset and the complexity of diagnosing it accurately in its early stages. Traditional methods rely heavily on clinical assessments and neuroimaging techniques, which, while effective, can be invasive, costly, and inaccessible to many[2, 4]. Furthermore, these methods often require specialized facilities and personnel, limiting their applicability on a broad scale[4]. For example, the Boston Diagnostic Aphasia Examination (BDAE) relies on patient speech for assessment of cognitive ability, but physician evaluation of patient speech in the examination can be time consuming and subjective. Previous studies have demonstrated that OpenAI's GPT models can effectively analyze text transformed from patient speech to discern the presence of AD[18, 19]. While this approach has shown promising results, its deployment within hospital systems encounters substantial obstacles due to privacy concerns since commercial AI companies may not be subjected to follow HIPAA requirements for their own assets[13, 14, 16]. These constraints significantly challenge the practical applicability and deployment of GPT models in clinical settings, casting doubts on their feasibility for widespread clinical use.

Data Privacy and HIPAA Compliance in Hospital Systems

As aforementioned, hospital systems are characterized by their stringent data privacy protocols, ensuring that patient information does not leave the secure confines of the hospital servers. This closed environment is crucial for protecting patient confidentiality and preserving patient privacy in the healthcare system. A cornerstone of this privacy framework in the United States is the HIPAA, which sets the standard for protecting sensitive patient data[16]. HIPAA compliance requires healthcare providers to implement comprehensive security measures to safeguard patient information against unauthorized access or breaches[13, 14, 15]. These regulations cover a wide range of protections, from physical security controls, firewalls, digital encryption, de-identification, purpose limitation, and aim to preserve the integrity and confidentiality of medical records[16]. Ultimately, for AI technology solutions, including LLMs, to be integrated into hospital systems, they must protect the patient data as well as the healthcare provider. This compliance not only protects patients but also enables the responsible use of innovative technologies for enhancing patient care.

Large Language Models and Zephyr

Large Language Models (LLMs) have garnered widespread recognition, particularly following the publication of OpenAI's ChatGPT[6]. These models, trained on extensive datasets, can perform a wide range of tasks, from text generation to complex problem-solving, making them invaluable across diverse domains[18, 19, 20, 21]. Among these advancements, Zephyr stands out as a notable development[22]. Originating as a refined, chat-optimized iteration of the mistralai/Mistral-7B-v0.1 model, Zephyr-7B-β is an open-source LLM fine-tuned on a blend of public and synthetic datasets, achieving competitive performance with commercial models such as GPT-3.5[22, 23]. Previous studies have demonstrated that OpenAI's GPT models can effectively analyze text transformed from patient speech to discern the presence of AD. However, its closed-source technology does not secure HIPAA compliance, whereas open-source technology like Zephyr-7B-β does [18]. In the case of AD prediction through speech detection, a person's speech is used for diagnosis through the BDAE but is also personally identifiable information which must be de-identified prior to input to the commercial AI-platform. Internal AI utilization obviates the need for de-identification because data is curated, stored, and utilized locally by the healthcare provider, and is inherently privacy-preserving. Zephyr-7B-β's development and deployment underscore the evolving landscape of LLMs, where the quest for improved performance intersects with ethical considerations and real-world applicability.

Methods

Overview

As illustrated in Figure 1, our approach, denoted as *ADetectoLocum*, commences with the input of patient audio recordings. Initially, we employ a specific audio-to-text conversion tool (Wav2Vec2Tokenizer and Wav2Vec2ForCTC) to transcribe the audio into textual data[24]. This direct transcription method is chosen over utilizing the acoustic properties of the audio based on evidence from previous studies, which have shown that for the task of AD detection, the textual content plays a pivotal role[18, 19, 25, 26]. Incorporating acoustic features has been found not to aid, and in some instances, to detrimentally affect the performance of the text-based module[18, 25, 26].

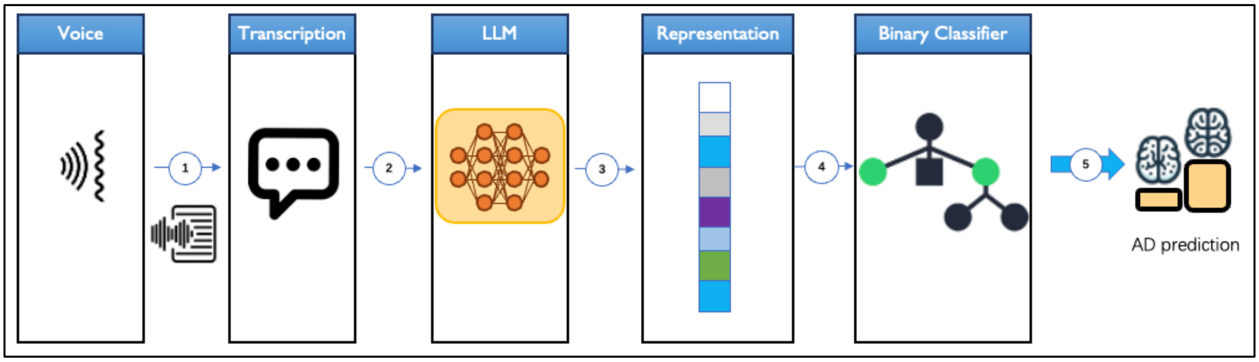

Figure 1: Overview of *ADetectoLocum*

Upon acquiring the text, this data is then inputted into the Zephyr-7B-β model for inference. We extract the feature representation from the layer preceding the output layer of the Zephyr-7B-β model to serve as the basis for our AD classification. This representation captures the nuanced linguistic patterns associated with AD, as encoded by the model's deep learning architecture.

Finally, we employ an XGBoost classifier, trained to utilize these feature representations as input[27]. The classifier is tasked with determining whether the analyzed voice, represented through the extracted features, belongs to a patient with AD. The outcome of this process is a binary classification, providing a probability score indicating the likelihood of AD presence or absence. This methodological framework not only leverages the textual analysis for AD detection but also enhances it with the advanced language understanding capabilities of Zephyr-7B-β, aiming for a more accurate and reliable diagnosis.

For a detailed explanation of our method, we proceed as follows: Steps 1 to 3, corresponding to ① to ③ in Figure 1, are thoroughly discussed in the Data and Experimental Setup section. Steps 4 and 5, associated with circles ④ and ⑤, are elaborated upon in the Binary Classification section. All code used is available at https://github.com/ginnymortensen/*ADetectoLocum*.git.

Data

We used the ADReSS20 and ADReSSo21 Challenge datasets for training, validation and testing *ADetectoLocum*, available through DementiaBank and published in 2020 and 2021, respectively[28, 29, 30]. The datasets contain spontaneous speech audio of patient descriptions of the Cookie Theft picture from the Boston Diagnostic Aphasia Examination (BDAE) and Mental Mini-State Evaluation (MMSE) scores for healthy, control patients and patients diagnosed with AD. The BDAE and MMSE are both assessments used by physicians to aid in AD diagnosis[2]. The datasets are split into training and testing cohorts, evenly stratified by age and gender. The ADReSS20 dataset differs from the ADReSSo21 dataset in that it includes audio transcriptions and normalized, sub-chunked audio in addition to full enhanced audio, whereas the latter contains only full enhanced audio. For consistency, we retain only the full enhanced audio from each dataset for the AD classification task.

In the ADReSS20 dataset, there are 54 AD and 54 cognitively normal (CN) patients in the training cohort. The testing cohort contains 24 AD and 24 CN patients. In the ADReSSo2021 dataset, there are 87 AD and 79 CN patients in the training cohort, and 35 AD and 36 CN patients in the testing cohort. We combine each dataset to form larger training and testing cohorts for a more robust model while retaining the original stratification of age and gender. We kept the testing cohort unseen by our classification models during training. The combined dataset totals 141 AD and 133 CN patients in the training cohort, and 59 AD and 60 CN patients in the testing cohort, for grand totals of 274 training samples and 119 testing samples, or 393 total samples.

Experimental Setup

In our Pipeline setup, the initial step involved securing permission to access the ADReSS20 and ADReSSo21 datasets from DementiaBank, followed by storing this data within a secure High-Performance Computing (HPC) environment[30]. Utilizing the Python package librosa with Python 3.9, we extracted waveforms from audio WAV files, marking the commencement of our analysis pipeline (as denoted by the first bin labeled "Voice" on the left side of Figure 1)[31]. At stage ① in Figure 1, the wav2vec 2.0 base model in Hugging Face, equipped with Wav2Vec2Tokenizer and Wav2Vec2ForCTC, was employed to tokenize these waveforms into textual transcriptions[28]. This speech-to-text model, fine-tuned on 960 hours of speech data, is acclaimed for its use by top performers in the ADReSSo21 Challenge[18, 19, 24]. The resulting transcriptions (stage ② in Fig.1, corresponding to the "Transcription" bin in Figure 1) were then processed through the open-source LLM, Zephyr-7B-β (stage ③ and corresponding to the "LLM" bin), to derive textual representations from the model's final layer. This LLM was accessed via Hugging Face's transformer module in Python[22]. These representations encapsulate both lexical and semantic insights, leveraging the model's extensive pre-training.

Binary Classification

Subsequently, at stage ④ in Figure 1, these textual representations were employed to train diverse binary classification models, including a Neural Network (NN), Support Vector Classifier (SVC), Extreme Gradient Boosting (XGB), and Logistic Regression (LR), (as highlighted by the "Binary Classifier" bin). We utilized the Python library Scikit-learn to import validated implementations of SVC and LR[32]. We utilized the Keras API for TensorFlow to utilize deep learning modules to build our NN. Lastly, we utilized the XGB Python package for our XGB model implementation[27]. We tune each classifier's hyperparameters using 5-fold cross validation with a grid-search over a range of values standard for each classifier. We chose these 4 classifiers as candidates because previous literature has shown success using SVC, NN, and LR for the ADReSSo21 Challenge[8, 18, 19]. However, we innovatively incorporate XGB because of its excellent performance as demonstrated in other classification challenges[33]. Indeed, this classifier proved to be the most accurate for AD classification in our approach. However, this stage also allows for flexibility in deployment, enabling users to select a binary classifier tailored to their needs, such as dataset size. Post-training, these classifiers estimate the probability of AD presence (⑤ in Figure 1). Our models were distinctively trained across both datasets and fine-tuned for specific evaluation studies, with performance metrics such as accuracy, precision, recall, and F1 score derived from their predictions. We utilized Scikit-learn library's metrics module to calculate metrics for stage ⑤[32]. To enhance processing speed and reduce inference latency, computations were accelerated using a Nvidia H100 PCIe GPU, ensuring all operations were conducted locally to avoid extra costs or data breaches.

Results

Early Detection Analysis

The most clinically important aspect of our model is its ability to accurately detect AD in early cognitive decline, visualized in Figure 2. Higher MMSE scores indicate lesser impairment, while lower MMSE scores indicates greater impairment[47]. Our model frequently correctly classifies AD patients with higher MMSE scores (and therefore less cognitive impairment), as indicated by the true positive distribution. We demonstrate our model's ability to detect AD early in the disease's progression, allowing expeditious intervention by physicians to enact treatment and improve patient outcomes for AD management.

False negatives for moderate impairment are likely a result of facilitator speech confounding the transcription quality, or bias from cognitively normal natural language used to train LLMs, including Zephyr, but segmenting audio to exclude facilitator speech was not shown to improve prediction latency in other works[29]. Additionally, false negatives are distributed lower on the MMSE impairment scale, indicating more noticeable impairment, but are less frequently incorrectly classified. Although *ADetectoLocum* is meant for AD prediction, false positives could indicate samples representing other diseases such as vascular dementia. Ultimately, our approach is accurate for cases where there is less impairment, distinguishing our work from other competitively performing models.

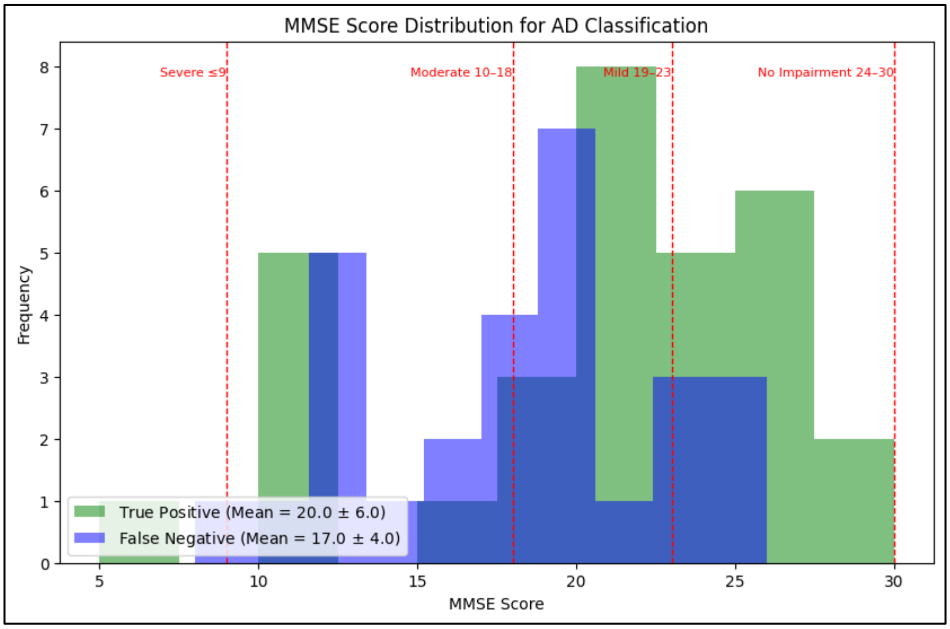

Figure 2: Histogram of correct and incorrect AD classifications with thresholds indicating MMSE cognitive impairment level. Means and standard deviations are calculated for MMSE scores per class.

Comparative Results

Our model outperforms previous methods of AD classification on linguistic features from the DementiaBank dataset using LLMs, as shown in Table 1. We compare our results to those presented in other works wherein inference was performed on language model vector representations of the linguistic features of audio transcripts for the ADReSSo21 dataset, including the baseline results for the ADReSSo21 Challenge presented by Luz et al, 2021. We chose to compare works using this dataset because the size of the data is larger and no contributors, to our knowledge, have experimented with combining ADReSS20 and ADReSSo21 data. Our model builds on the success of previous work demonstrating the feasibility of employing LLMs to achieve high accuracy in AD classification with the advantage of using locally deployable software. Our F1-score outcompetes other contributors and indicates a more effective model performance, minimizing false classifications. Specifically, when comparing *ADetectoLocum* against other models that can be deployed locally, such as those proposed by Bang et al., 2024, and Luz et al., 2021, our model demonstrates superior performance across all evaluated metrics, including accuracy, precision, recall, and F1 score. Notably, *ADetectoLocum* exceeds the benchmarks set by these studies, establishing itself as the leading solution in terms of both detection capability and efficiency.

In contrast to the GPT-based approach by Agbavor & Liang, 2022, *ADetectoLocum* showcases a remarkable improvement in key performance indicators. Our model achieves a nearly 5% increase in accuracy, a 12% improvement in precision, and a 2% uplift in F1 score (Table 1). However, it's important to acknowledge that *ADetectoLocum* exhibits a 12% lower recall rate compared to Agbavor & Liang's method. However, the high recall of Agbavor & Liang's results compared to the relatively low precision may indicate the method, while accurate, is biased and captured false positives as well. Bang's method using BERT is competitive to *ADetectoLocum*, but is not explored for early-stage detection, whereas our method has showed AD detection early in the disease's progression (Figure 2). Additionally, although GPT-3.5 + SVC is competitive, another work exploring GPT-3.5 + XGB showed an accuracy of 62% and is mentioned for completeness[36]. GPT-3.5 based comparisons are sufficient to demonstrate our approach's efficacy, as GPT-4 is not locally deployable.

Table 1: Comparative results of our model results with other contributors for AD classification.

Contributors	Model	Accuracy	Precision	Recall	F1
ADetectoLocum (ours)	Zephyr + XGB	**84.9%**	**84.7%**	84.7%	**84.7%**
Bang et al., 2024	BERT (text unimodal)	83.1%	83.1%	83.1%	83.1%
Agbavor & Liang, 2022	GPT-3.5 + SVC	80.3%	72.3%	**97.1%**	82.9%
Luz et al., 2021	SVC	78.9%	77.8%	80.0%	78.9%

LLM Comparison

Meta's open-source LLM Llama 2 has gained attention for being powerful and flexible[34]. Table 4 shows the performance of Llama 2 in our pipeline when used for stages ② and ③. Llama 2-7B has demonstrated impressive performance on medical benchmarks and there exists several iterations of Llama 2 fine-tuned for the medical domain[38]. However, Zephyr-7B-β is documented to have outperformed Llama 2-7B on several benchmarks, including the medical domain[22]. When comparing accuracy, Llama 2 shows competitive performance to current SOTA models, but does not outperform Zephyr.

Table 2. Llama 2 results on AD classification.

LLM	Classifier	Accuracy	Precision	Recall	F1
Llama 2	NN	0.731	0.865	0.542	0.667
	SVC	0.798	0.818	0.763	0.789
	XGB	0.824	0.852	**0.78**	0.814
	LR	**0.832**	**0.882**	0.763	**0.818**

Furthermore, Figure 3 shows that *ADetectoLocum*, using Zephyr for stage ②, is more capable of distinguishing between AD and CN patients than if Llama 2 is used. AUC and standard deviations are derived from 5-fold cross validation used to tune hyperparameters for each variant of our model. Not only is the variant of our proposed model, using Llama 2 for stage ②, more accurate than current SOTA models, but *ADetectoLocum* as proposed is more reliable than the variant in AD classification. For example, previous work shows that GPT-3.5's Babbage model achieves an AUC of 0.91 on AD classification, whereas we achieve an AUC of 0.94 using Zephyr for textual embedding extraction[18].

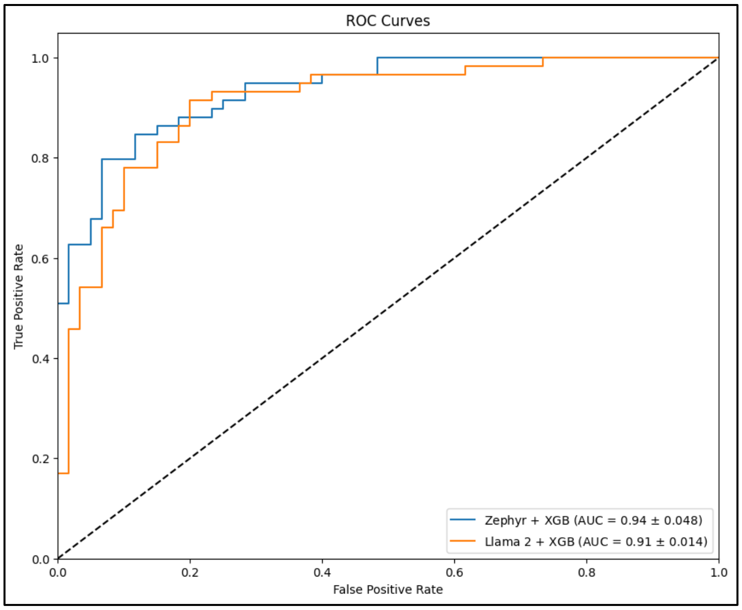

Figure 3: ROC curve comparing Zephyr and Llama 2 performance in context of *ADetectoLocum* with 5-fold cross validation derived AUC standard deviations.

Cross Validation

As shown in Table 2 we achieved the best performance on the unseen testing dataset with XGB classification. To select the best classifier to perform inference, we tuned the relevant hyperparameters for each model using 5-fold cross validation with the combined training cohort from both datasets. The best XGB model was tuned to a learning rate of 0.2, a maximum depth of 5, and a number of 50 estimators. Our results show LR performed marginally better than XGB because the fold by which testing is performed to tune the best model hyperparameters is a smaller set of data than the testing performed on the trained model following tuning. It is known that logistic regression performs well with smaller training data, while XGB performs very well on larger datasets. This insight explains XGB's superior performance on the combined testing cohort. Therefore, our method presents flexibility in utilization, where small clinical trials will favor a logistic regression classifier while large multi-institution studies would benefit from XGB classification. This gives users of *ADetectoLocum* more flexibility and control to maximize accuracy depending on the size of the data.

Table 3. Model training using cross-validation and testing results. Standard deviations are shown in parenthesis.

	Classifier	Accuracy	Precision	Recall	F1
5-fold CV	NN	0.777 (0.095)	0.557 (0.113)	0.463 (0.158)	0.502 (0.137)
	SVC	0.836 (0.042)	0.849 (0.059)	**0.836 (0.084)**	0.839 (0.045)
	XGB	0.814 (0.068)	0.834 (0.093)	0.808 (0.049)	0.819 (0.061)
	LR	**0.843 (0.057)**	**0.865 (0.071)**	0.829 (0.066)	**0.845 (0.056)**
Test	NN	0.807	0.75	**0.915**	0.824
	SVC	0.832	0.831	0.831	0.831
	XGB	**0.849**	**0.847**	0.847	**0.847**
	LR	0.832	0.842	0.814	0.828

To further examine the robustness of our selected model, we performed an out-of-time validation study as shown in Table 3. We trained our model on the training cohort from the ADReSS20 dataset using hyperparameter tuning and 5-fold cross validation, as was performed for our baseline model, and tested the resulting model with the testing cohort of the ADReSSo21 dataset. Our LR results are competitive with previous works, consistent with our cross-validation, and demonstrate the ability to perform with high accuracy given a small amount of data. Additionally, our approach offers users the flexibility to utilize different classifier options. Despite NN, SVC, and XGB suboptimal performance on smaller datasets, providing these options allows users the opportunity to experiment with these models as they acquire larger datasets. This inclusivity ensures our method caters to a broader range of user needs and data availability scenarios, facilitating the potential for improved performance with increased data volume.

Table 4. Validation time study wherein our model was trained on ADReSS20 data and tested on ADReSSo21 data (from the following year).

	Classifier	**Accuracy**	**Precision**	**Recall**	**F1**
Test	NN	0.69	0.618	**0.971**	0.756
	SVC	0.746	0.743	0.743	0.743
	XGB	0.789	0.778	0.8	0.789
	LR	**0.831**	**0.871**	0.771	**0.818**

Efficiency

Our model is advantageously lightweight with a total pipeline execution time of merely 12 minutes and 38 seconds. The extraction time of text transcriptions from audio files (stage ① in Figure 1) was 3 minutes and 46 seconds. The Zephyr-7B-β model loading and extraction of text embeddings for all samples (stage ② and ③) occurred in 47 seconds. The maximum classifier training time was 8 minutes and 5 seconds (~485 seconds) and the maximum testing time was 0.002 seconds (stage ④). Metric evaluation time (stage ⑤) was 0.003 seconds.

Discussion and Conclusions

We introduced *ADetectoLocum*, a locally deployable AD detection pipeline that outperforms existing methods and rivals GPT-based systems—which cannot be deployed in hospitals due to practical constraints. Notably, *ADetectoLocum* excels in early AD detection, facilitating timely interventions and improved patient outcomes.

Comparative Advantages

ADetectoLocum establishes a performative edge in comparison to current SOTA models. *ADetectoLocum* achieves an accuracy of 84.9%, greater than previous work utilizing similar modalities of inference for AD classification. Additionally, *ADetectoLocum* demonstrates superior ability to minimize incorrect classifications with an F1-score of 84.1%, and an AUC of 0.94. This high performance validates Zephyr's linguistic understanding of natural language. Because Zephyr is a chat-optimized model, it is most suited for understanding natural language text which include informal dialogue aligned with spontaneous speech. Impressively, Zephyr textual embeddings represent not only this natural dialogue aligned with cognitively normal spontaneous speech, but also includes necessary context for distinguishment between AD dialogue as well. This important capability poises Zephyr as the current consummate model for AD classification from linguistic features alone.

Locality, Efficiency, and Robustness

ADetectoLocum's key innovation is its local, HIPAA-compliant deployment, eliminating reliance on external models and ensuring patient privacy. Built on open-source technology, it integrates seamlessly into hospital infrastructures. Training requires slightly over 8 minutes with hundreds of examples, while inference occurs in under 1 second per test case. Its robustness is confirmed via 5-fold cross-validation (Table 3) and time-validation (Table 4), closely matching comparative testing at 84.3% and 83.1% accuracies, respectively. The ADReSS20 and ADReSSo21 datasets (393 samples total) underpin our results through even attribute stratification as provided by DementiaBank[18, 19, 29, 36].

Patient Impact

ADetectoLocum's excellent ability to classify AD, particularly in patients with higher MMSE scores (indicating a lesser degree of cognitive impairment), indicates its ability to detect AD at nascent stages. Additionally, the ability for early detection also addresses the confounding aspects of AD, particularly whether cognitive impairment is a result of natural aging or from disease progression. However, it is important to note that *ADetectoLocum*, although maximizing true positives, still detects false negatives. Results did not improve with audio segmentation of confounding BDAE facilitator speech; however, facilitator speech was minimal in the high quality DementiaBank recordings. The risk of a false negative could mean delayed treatment and a more pronounced progression of the disease, decreased quality of life, and the patient and/or family perhaps needing to make sudden decisions. False positives could mean exactly the opposite, with the family perhaps committing financial resources prematurely, or samples representing other diseases such as vascular dementia. Early detection in the disease's progression is significant because it allows physicians to preemptively implement therapeutic interventions to decelerate AD progression and improve patient quality of life.

Ethical Considerations and Limitations

Ethical Considerations and Limitations While *ADetectoLocum* represents a significant advancement in the early detection of AD through spontaneous speech and a pragmatic solution for LLM integration with hospital infrastructures, there are some limitations. *ADetectoLocum* is trained on a relatively small cohort of native English speakers from the US and is not explicitly inclusive of non-native English speakers or native English speakers with colloquial or regional accents, as no larger publicly available dataset exists. Regardless, deployment of such a system in a hospital may bias the model's cognitive normality classification towards linguistic patterns typical for the community the hospital services, which can vary by region. Further, model perplexity can be used to balance our model, as consequent works have successfully incorporated such features into their models to improve their baselines[35]. Additionally, previous work demonstrated success in increasing AD detection accuracy when combining ChatGPT opinions of AD presence and text embeddings of speech[19]. To further enhance *ADetectoLocum*, Zephyr could be fine-tuned on medical corpora, give opinions, and be re-employed in a multimodal approach.

Conclusion

ADetectoLocum represents a significant advance in early AD detection while satisfying stringent healthcare data requirements. It demonstrates that open-source AI can be HIPAA-compliant and achieve diagnostic accuracy comparable to commercial models. Future research should focus on fine-tuning and multimodal adaptations to further enhance early detection, ultimately enabling more efficient AD management and improved patient care.

Acknowledgments

We'd like to thank Professor Haixu Tang for the guidance and CCF-2207231 from National Science Foundation for the support.

References

1. 2023 Alzheimer's disease facts and figures. Alzheimer's & Dementia. 2023;19(4):1598–695.
2. Teipel S, Gustafson D, Ossenkoppele R, Hansson O, Babiloni C, Wagner M, et al. Alzheimer Disease: Standard of Diagnosis, Treatment, Care, and Prevention. Journal of Nuclear Medicine. 2022 Jul 1;63(7):981–5.
3. Britton GB, Rao KSJ. Cognitive Aging and Early Diagnosis Challenges in Alzheimer's Disease. Journal of Alzheimer's Disease. 2011 Jan 1;24(s2):153–9.
4. Prince M, Bryce DR, Ferri DC. World Alzheimer Report 2011: The benefits of early diagnosis and intervention.
5. Yiannopoulou KG, Papageorgiou SG. Current and Future Treatments in Alzheimer Disease: An Update. J Cent Nerv Syst Dis. 2020 Feb 29;12:1179573520907397.
6. Brown TB, Mann B, Ryder N, Subbiah M, Kaplan J, Dhariwal P, et al. Language Models are Few-Shot Learners [Internet]. arXiv; 2020 [cited 2024 Mar 17]. Available from: http://arxiv.org/abs/2005.14165
7. Feng Y, Wang J, Gu X, Xu X, Zhang M. Large language models improve Alzheimer's disease diagnosis using multi-modality data [Internet]. arXiv; 2023 [cited 2024 Mar 17]. Available from: http://arxiv.org/abs/2305.19280
8. Agbavor F, Liang H. Artificial Intelligence-Enabled End-To-End Detection and Assessment of Alzheimer's Disease Using Voice. Brain Sci. 2022 Dec 23;13(1):28.
9. Meskó B. The Impact of Multimodal Large Language Models on Health Care's Future. J Med Internet Res. 2023 Nov 2;25:e52865.
10. Nadarzynski T, Miles O, Cowie A, Ridge D. Acceptability of artificial intelligence (AI)-led chatbot services in healthcare: A mixed-methods study. DIGITAL HEALTH. 2019 Jan 1;5:2055207619871808.
11. Goodman RS, Patrinely JR, Stone CA Jr, Zimmerman E, Donald RR, Chang SS, et al. Accuracy and Reliability of Chatbot Responses to Physician Questions. JAMA Network Open. 2023 Oct 2;6(10):e2336483.
12. Horiuchi D, Tatekawa H, Shimono T, Walston SL, Takita H, Matsushita S, et al. Accuracy of ChatGPT generated diagnosis from patient's medical history and imaging findings in neuroradiology cases. Neuroradiology. 2024 Jan;66(1):73–9.

13. Li J. Security Implications of AI Chatbots in Health Care. Journal of Medical Internet Research. 2023 Nov 28;25(1):e47551.
14. Yadav N, Pandey S, Gupta A, Dudani P, Gupta S, Rangarajan K. Data Privacy in Healthcare: In the Era of Artificial Intelligence. Indian Dermatol Online J. 2023 Oct 27;14(6):788–92.
15. Price WN, Cohen IG. Privacy in the Age of Medical Big Data. Nat Med. 2019 Jan;25(1):37–43.
16. Moore W, Frye S. Review of HIPAA, Part 1: History, Protected Health Information, and Privacy and Security Rules. Journal of Nuclear Medicine Technology. 2019 Dec 1;47(4):269–72.
17. OpenAI, Achiam J, Adler S, Agarwal S, Ahmad L, Akkaya I, et al. GPT-4 Technical Report [Internet]. arXiv; 2024 [cited 2024 Mar 18]. Available from: http://arxiv.org/abs/2303.08774
18. Agbavor F, Liang H. Predicting dementia from spontaneous speech using large language models. PLOS Digital Health. 2022 Dec 22;1(12):e0000168.
19. Alzheimer's disease recognition from spontaneous speech using large language models - Bang - 2024 - ETRI Journal - Wiley Online Library [Internet]. [cited 2024 Mar 17]. Available from: https://onlinelibrary.wiley.com/doi/full/10.4218/etrij.2023-0356
20. Rizwan A, Sadiq T, Rizwan A, Sadiq T. The Use of AI in Diagnosing Diseases and Providing Management Plans: A Consultation on Cardiovascular Disorders With ChatGPT. Cureus [Internet]. 2023 Aug 7 [cited 2024 Mar 17];15(8). Available from: https://www.cureus.com/articles/174346-the-use-of-ai-in-diagnosing-diseases-and-providing-management-plans-a-consultation-on-cardiovascular-disorders-with-chatgpt
21. Liang H, Tsui BY, Ni H, Valentim CCS, Baxter SL, Liu G, et al. Evaluation and accurate diagnoses of pediatric diseases using artificial intelligence. Nat Med. 2019 Mar;25(3):433–8.
22. Tunstall L, Beeching E, Lambert N, Rajani N, Rasul K, Belkada Y, et al. Zephyr: Direct Distillation of LM Alignment [Internet]. arXiv; 2023 [cited 2024 Mar 17]. Available from: http://arxiv.org/abs/2310.16944
23. Jiang AQ, Sablayrolles A, Mensch A, Bamford C, Chaplot DS, Casas D de las, et al. Mistral 7B [Internet]. arXiv; 2023 [cited 2024 Mar 18]. Available from: http://arxiv.org/abs/2310.06825
24. Baevski A, Zhou H, Mohamed A, Auli M. wav2vec 2.0: A Framework for Self-Supervised Learning of Speech Representations [Internet]. arXiv; 2020 [cited 2024 Mar 17]. Available from: http://arxiv.org/abs/2006.11477
25. Balagopalan A, Eyre B, Robin J, Rudzicz F, Novikova J. Comparing Pre-trained and Feature-Based Models for Prediction of Alzheimer's Disease Based on Speech. Front Aging Neurosci [Internet]. 2021 Apr 27 [cited 2024 Mar 17];13. Available from: https://www.frontiersin.org/articles/10.3389/fnagi.2021.635945
26. Pan Y, Mirheidari B, Harris JM, Thompson JC, Jones M, Snowden JS, et al. Using the Outputs of Different Automatic Speech Recognition Paradigms for Acoustic- and BERT-Based Alzheimer's Dementia Detection Through Spontaneous Speech. In 2021 [cited 2024 Mar 17]. p. 3810–4. Available from: https://www.isca-archive.org/interspeech_2021/pan21c_interspeech.html
27. Chen T, Guestrin C. XGBoost: A Scalable Tree Boosting System. In: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining [Internet]. 2016 [cited 2024 Mar 17]. p. 785–94. Available from: http://arxiv.org/abs/1603.02754
28. Luz S, Haider F, de la Fuente S, Fromm D, MacWhinney B. Alzheimer's Dementia Recognition through Spontaneous Speech: The ADReSS Challenge [Internet]. arXiv; 2020 [cited 2024 Mar 17]. Available from: http://arxiv.org/abs/2004.06833
29. Luz S, Haider F. Detecting cognitive decline using speech only: The ADReSSO Challenge.
30. Lanzi AM, Saylor AK, Fromm D, Liu H, MacWhinney B, Cohen ML. DementiaBank: Theoretical Rationale, Protocol, and Illustrative Analyses. Am J Speech Lang Pathol. 2023 Mar 9;32(2):426–38.
31. McFee B, McVicar M, Faronbi D, Roman I, Gover M, Balke S, et al. librosa/librosa: 0.10.1 [Internet]. Zenodo; 2023 [cited 2024 Mar 18]. Available from: https://zenodo.org/records/8252662
32. Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, et al. Scikit-learn: Machine Learning in Python. Journal of Machine Learning Research. 2011;12(85):2825–30.
33. Grinsztajn L, Oyallon E, Varoquaux G. Why do tree-based models still outperform deep learning on tabular data? [Internet]. arXiv; 2022 [cited 2024 Mar 18]. Available from: http://arxiv.org/abs/2207.08815
34. Llama 2: Open Foundation and Fine-Tuned Chat Models | Research - AI at Meta [Internet]. [cited 2024 Mar 18]. Available from: https://ai.meta.com/research/publications/llama-2-open-foundation-and-fine-tuned-chat-models/
35. Guo Z, Ling Z, Li Y. Detecting Alzheimer's Disease from Continuous Speech Using Language Models. J Alzheimers Dis. 2019;70(4):1163–74.
36. Kheirkhahzadeh, Maryam. "Speech Classification using Acoustic Embedding and Large Language Models Applied on Alzheimer's Disease Prediction Task.", 2023.