Multilingual Dementia Detection through Deep Learning

Grant Proposal

Gustavo Rodriguez

Massachusetts Academy of Math and Science

85 Prescott St, Worcester, MA 01605

Executive Summary

Alzheimer's disease, a subset of dementia, is a cognitive illness that will affect 152 million people by 2050. Despite its effects, only 50% of people with dementia have received a diagnosis (Shinkawa et al., 2018). Alternative diagnosis methods must be pursued to make diagnoses more accessible. Current methods of diagnosis are centered around machinery such as MRI or PET scans, or processes such as cerebrospinal fluid collection; both methods deter patients by being too expensive or being too intrusive. This project aims to expand on the progress made by prior research by applying speech-based dementia detection models through a multilingual lens to provide a low cost, globally accessible method of detection. Current research into such models often focuses on English, which is a gap in the field (Perez-Toro et al., 2023). The model used will combine transcription with binary classification by a large language model. Various configurations will be used including monolingual and bilingual large language models and various transcription models. Some possible implementations of a diagnostic model could be as a part of an app or as a short portion of a doctor’s appointment.

Multilingual Alzheimer's Detection through Deep Learning

Alzheimer's Disease patients face a harsh decline in their cognitive ability, decreasing their ability to perform everyday functions, access memories, and communicate. Although work on cures and methods of slowing the illness is ongoing, only 40-50% of people with Dementia in the United States are diagnosed (Shinkawa et al., 2018). This number is expected to rise significantly to 152 million people by 2050 (Shinkawa et al., 2018). Detection methods must become easier to access to provide routes to and provide earlier options to prevent Alzheimer's from stealing the late portion of a patient's life.

The current methods of Alzheimer's detection are accurate, yet often lack accessibility due to cost or invasive procedures. The most common methods are MRIs, PET scans, and cerebrospinal fluid analysis (Yang et al., 2022). MRIs and PET scans are expensive, whereas cerebrospinal fluid collection is an invasive procedure. Due to the limited availability of these machines to carry out these diagnoses, they are also difficult to scale. One recent approach to detection is the use of machine learning and deep learning models to provide diagnoses based on speech data. However, models are often trained on English-language datasets, making the implementation of early diagnosis models difficult to implement at a large scale internationally (Pérez-Toro et al., 2023). In English alone, databases containing Alzheimer's or Dementia patients, in addition to healthy controls, need to be ethically collected to train Machine Learning with sufficient data. The high bar of entry is especially detrimental to smaller communities that may share a tribal, regional, or uncommon language. 6909 languages exist currently, and training an individual model for each language is a lofty goal (Anderson, n.d.). To provide tools for diagnosis for all languages, a model developed to incorporate universal features of similar languages can provide the foundation for effectively training a machine learning model to diagnose dementia in non-English speaking patients.

One potential way of finding universality between languages is to focus on cognitive factors of language as opposed to or in combination with the syntax and word meaning. Meulemans et al. (2022) explored the pause times of different words in Dutch and found that the difference in pause times between distinct types of words was different due to the nature of the language (Meulemans et al., 2022). However, it was also observed that the cognitively impaired group's pause times per word type were consistently longer than those of the healthy control group by nearly identical amounts. Similar patterns could be further researched in other languages to see if a pattern could be developed interlingually.

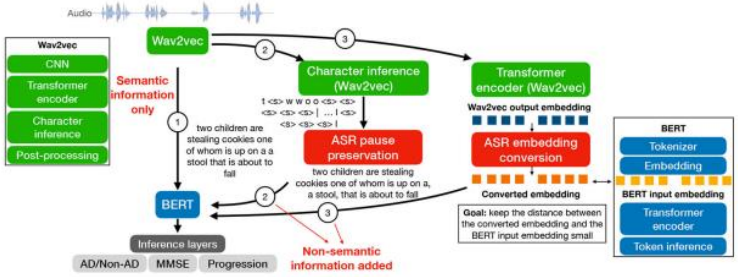


Figure 1: Flow Chart of the WavBERT model created by Zhu et al. (2022)

Another study used cognitive data to train a model titled WavBERT by implementing a novel iteration of wav2vec that could translate pauses in speech into commas in transcripts. These transcripts would be fed into a Bidirectional Encoder Representations from Transformers (BERT) classifier model to determine if a patient had Alzheimer's. The WavBERT model, shown in Figure 1, could analyze vocabulary content and cognitive factors, such as pauses, using a model that had already been pre-trained (Zhu et al., 2021). This method allowed for the simultaneous analysis of linguistic and acoustic properties, creating a model that combined multiple aspects of detection to create a complete algorithm without the need for manual transcription with 83% accuracy.

Other studies have looked at interlingual aspects of Alzheimer's disease. One study tested whether deep learning methods could link patterns in similar languages despite being only trained in one language (Pérez-Toro et al., 2023). The tests supported the idea that training a model on a language first would increase its ability to detect Alzheimer's in that language; otherwise acoustic features would

be the only left over data that could be used in an algorithm. However, there were marked similarities between English and German in acoustic embeddings. The ability of these two languages to cross diagnose to some extent could provide a potential venue for further research, as both are classified as part of the Germanic family of languages and could share other features, proving the link of languages from similar origins and a potential for diagnosis in untrained languages.

This project aims to analyze the ability of a deep learning model to detect Alzheimer's Disease across multiple languages. A model would be used to transcribe speech with quantifiable spaces included and input the transcripts into a pre-trained multilingual model to predict Alzheimer's.

Section II: Specific Aims

Specific Aim 1: The model can diagnose Alzheimer's Disease with a high level of accuracy.

Specific Aim 2: The model can retain reasonable accuracy in multiple languages.

Specific Aim 3: The model can remain interlingual without repeating pre-training in a second language.

Section III: Project Goals and Methodology

Relevance/Significance

Creating an accurate interlingual model that is necessary to provide equal opportunities for people of all communities to receive treatment for Alzheimer's Disease. With the increasing potential of speech to diagnose Alzheimer's, many communities with local languages could be left behind and not receive a proper diagnosis required for medical treatment. Most machine learning models currently are trained in major languages such as English and Spanish. However, there are not enough resources to complete the training of a model for Alzheimer's in each of the thousands of languages that are spoken currently. If a machine learning model could use specific languages that can have their qualities

extrapolated to others or use its literacy in multiple languages to apply diagnostic data, such a wide reach of diagnosis could become much more feasible.

Innovation

This project aims to innovate by experimenting on the viability of interlingual models and improving on the current models used for monolingual and multilingual diagnoses. Several configurations of pre-training on various single and multilingual transcription algorithms and BERT models will allow for this experimentation. Improvements to the creation of punctuation will yield a model stronger in one language and multiple languages.

Methodology

Specific Aim #1: The model can diagnose Alzheimer's Disease with a high level of accuracy.

	Embeddings	Model	Accuracy	Precision	Recall	F1
10-fold CV	Ada	SVC	0.788 (0.075)	0.798 (0.109)	0.819 (0.098)	0.799 (0.066)
		LR	0.796 (0.107)	0.798 (0.126)	**0.835 (0.129)**	0.808 (0.100)
		RF	0.734 (0.090)	0.738 (0.109)	0.763 (0.149)	0.743 (0.103)
	Babbage	SVC	0.802 (0.054)	0.823 (0.092)	0.804 (0.103)	0.806 (0.053)
		LR	**0.809 (0.112)**	**0.843 (0.148)**	0.811 (0.091)	**0.818 (0.091)**
		RF	0.760 (0.052)	0.780 (0.102)	0.781 (0.110)	0.770 (0.047)
Test Set	Ada	SVC	0.788	0.708	**0.971**	0.819
		LR	0.718	0.653	0.914	0.762
		RF	0.732	0.690	0.829	0.753
	Babbage	SVC	**0.803**	**0.723**	**0.971**	**0.829**
		LR	0.718	0.647	0.943	0.767
		RF	0.761	0.714	0.857	0.779

Table 2: GPT-3 embedding based model results (Williams et al., 2023)

Task	1. Classification (%)					
	Class	Precision	Recall	F1	Mean F1	Accuracy
Baseline [10]	non-AD AD	80.00 77.80	77.80 80.00	78.87 78.87	78.87	78.87
M_b	non-AD AD	71.79 75.00	77.78 68.57	74.67 71.64	73.16	73.24
M_{p1}	non-AD AD	80.00 87.10	88.89 77.14	84.21 81.82	83.02	**83.10**
M_{p2}	non-AD AD	77.50 83.87	86.11 74.29	81.58 78.79	80.19	80.28
M_e	non-AD AD	78.95 81.82	83.33 77.14	81.08 79.41	80.25	80.28
M_{e+p2}	non-AD AD	77.78 77.14	77.78 77.14	77.78 77.14	77.46	77.46

Table 1: WavBERT classification results (Zhu et al., 2021)

Note: M_{p1}, the most successful diagnostic configuration, used wav2vec transcripts with periods added.

Justification and Feasibility.

The accuracy of models, especially those that do not require professional transcription, is crucial for detection through deep learning to become mainstream. For a method to be viable, it would ideally

approach the accuracy levels of a traditional test, such as an MRI. The public opinion of these models is also important, as a method needs to appeal to professionals and the public to become widely adopted, providing a barrier to accessibility. A highly accurate model would provide significantly greater likelihood. Deep learning models are trained on thousands of parameters, allowing them to be more precise.

Improving the accuracy of a model is difficult and requires creative thinking, as many various methods have been pursued. Zhu et al. (2021) showed that using a combination of a transcription model with grammatical literacy and BERT could achieve a maximum accuracy of 83%. These findings provide a clear path to improve on this model in an attempt to make it achieve an accuracy that could be acceptable in the medical field.

The use of deep learning through BERT, as will be done in this research, is the strongest existing way to get a speech-based Alzheimer's diagnosis. When compared to traditional machine learning methods, BERT consistently had higher accuracy in studies (Balagopalan et al., 2020). Deep learning methods are also capable of simplifying complex calculations relating to characteristics of text since the model creates its parameters as opposed to preset parameters in a machine learning model. In machine learning, complex variables such as the vocabulary of a person relative to age and education level need to be crafted. Models by Williams et al. (2023) required significant expertise in data collection, normalization, as well as mathematical skills and time; however, these variables did not guarantee success. Additionally, other forms of deep learning, such as the use of a GPT-3 model, have been slightly less effective, with accuracy in the 70-80% (Agbavor & Liang, 2022) range as opposed to the 83% accuracy collected by Zhu et al. (2021), who used the WavBERT model, as shown in Tables 1 and 2 respectively.

As opposed to wav2vec, Whisper, a newer transcription model created by OpenAI, will be used. This model can utilize pauses in a more efficient way than wav2vec, which does not have punctuation built in. Whisper also can produce dashes, further increasing the variety of text and its similarity to the data that BERT was pre-trained on, the English Wikipedia.

Summary of Preliminary Data.

	precision	recall	f1-score	support
0	0.50	0.50	0.50	4
1	0.50	0.50	0.50	4
accuracy			0.50	8
macro avg	0.50	0.50	0.50	8
weighted avg	0.50	0.50	0.50	8

Figure 2: Preliminary data accuracy

	precision	recall	f1-score	support
0	0.00	0.00	0.00	4
1	0.60	1.00	0.75	6
accuracy			0.60	10
macro avg	0.30	0.50	0.37	10
weighted avg	0.36	0.60	0.45	10

Figure 3: Second preliminary test.

The model was trained on 19 dementia patients and 19 controls from the Pitt Corpus using a combination of Whisper as a transcriber and BERT as a classifier. An accuracy of 50% was achieved using an 80/20 train test split (Figure 2). Further improvements on the model will likely yield higher accuracy. Further iterations faced issues with the train test split which revealed potential problems with the

model, such as guessing a diagnosis for every input value (Figure 3). These will be addressed in future iterations.

Expected Outcome.

The expected outcome is an incremental improvement in the accuracy of the WavBERT model created by Zhu et al (2021). Ideally, this improvement could be significant, although the current experiment does not reinvent many aspects of the model and thus would not be likely to create a monumental increase in accuracy. An ideal accuracy would be 80% or above, although a higher sensitivity value would be considered stronger than a higher specificity value due to preferred false positives.

Potential Pitfalls and Alternative Strategies:

The proposed changes, including the use of a Whisper transcription model or improvements to a Wav2vec transcription model, could be particularly difficult to implement for a minor change in accuracy, which could result in failure to fully implement them. It is also possible that these improvements will have minor impact on the accuracy of the model. In that case, the use of more complex BERT models or the collection of more data could supplement the strengths of the model to prove its validity.

Specific Aim #2: The model can retain reasonable accuracy in multiple languages.

Justification and Feasibility.

Due to the nature of current detection methods, a solution is needed to detect Alzheimer's across a larger number of languages. Most research on the link between linguistics and Alzheimer's focuses on a small number of languages, primarily English. For such a form of detection to become mainstream, models for many more languages would need to be universally available for people in their

native language. Pérez-Toro et al. (2023) and Melistas et al. (2023) have researched multilingual diagnoses using machine learning. These studies have achieved accuracies in the 70% to 80% range when diagnosing across languages using various methods. They found promise in multilingual diagnoses for a model trained in multiple languages. By attempting an optimized deep learning model, these results could potentially be further improved. However, due to the black box nature of deep learning models, it is difficult to predict whether a model will be able to predict something until it has been tested. There is a strong chance that little improvements could be made on untrained languages. However, an improvement in training in multiple languages is also possible with the use of a multilingual BERT, as proposed for this study.

Expected Outcome.

It is expected that, at minimum, the multilingual diagnostic abilities of this model will be on par with those of similar studies. It is possible that the use of Whisper in combination with BERT will provide further accuracy than was seen in other studies. An ideal result would exceed 80% accuracy, although 70% to 80% accuracy with high sensitivity values would still yield positive results and potential direction for further research.

Potential Pitfalls and Alternative Strategies.

Due to the nature of deep learning, there is a possibility that this approach does not yield ideal results and that the model is not effective. In this case, further configurations could be explored, including combinations with demographic data.

Specific Aim #3: *The model can retain its accuracy in a language it has not yet encountered.*

Justification and Feasibility.

There are 6909 languages that currently exist (Anderson, n.d.). However, very few diagnostic machine learning studies have ventured beyond using English as a training set. The lack of diversity in training of models leaves a vast research gap, especially considering local languages or dialects that may be difficult to collect data for. If the origin of a language is known, it could be used for diagnosis by a language of similar origins that has more training data. Therefore, evidence that a model locating these similarities is can be used for diagnosis is crucial. The development of such a model would also save a significant amount of time, as training a large language model takes significant resources, including significant amounts of written language and computing power. Vastly opening the doors to diagnostic tools is important, especially to reach those who may not have access to or do not know of MRI scans, PET scans, or cerebrospinal fluid collection methods.

Expected Outcome.

It is expected that, at minimum, the multilingual diagnostic abilities of this model will be on par with those of similar studies. It is possible that the use of Whisper in combination with BERT will provide further accuracy than was seen in other studies. An ideal result would exceed 80% accuracy, although 70% to 80% accuracy with high sensitivity values would still yield positive results and potential direction for further research. False positives will be considered more optimal than false negatives since an abundance of diagnoses would ensure that more people seek out concrete answers and treatment. Ideally, this data could prove that a model like this one has the potential to be improved upon to form a concrete diagnostic tool.

Potential Pitfalls and Alternative Strategies.

Due to the nature of deep learning, there is the possibility that this approach does not yield ideal results and that the model is not effective. In this case, further configurations could be explored or

combinations with demographic data could be explored. More combinations of languages could also be tested to see whether a new language may provide a missing link.

Section IV: Resources/Equipment

Independent Variable:

The independent variables in this testing will be the models used to analyze Alzheimer's Disease and the language input during testing.

Dependent Variable:

The dependent variable in this experiment will be the outputs of the model (has Alzheimer's Disease or does not have Alzheimer's Disease)

Materials List:

Access to computer(s) for data processing

Dementia Bank database data (Spanish and English)

Deep Learning Software (Anaconda, TensorFlow, etc.)

Procedure:

1. Pretraining of the model (if needed) would involve ensuring that the BERT model is trained to analyze a certain language.
2. The speech recordings from the database would be inputted into an altered multilingual wav2vec or whisper model. All transcripts would be saved.
3. The BERT classifier model would be compiled and fitted to 90% of the speech transcripts generated from the whisper or wav2vec model.
4. After the BERT classifier is trained, it would be tested on the remaining 10% of the speech transcripts that were not used for training.
5. Boolean data outputs will be recorded from the classifier (Alzheimer's or no Alzheimer's).

6. Accuracy would be calculated, and data analysis would be performed on the data. Specificity and sensitivity would be recorded.

Section V: Ethical Considerations

This project will use data from databases within DementiaBank containing speech data of Alzheimer's patients. This data will be handled with care and following the guidelines set by TalkBank. All data used will be cited. The English dataset being used is the Pitt corpus (Becker et al., 1994), which is funded by grants NIA AG03705 and AG05133. No data will be stored online unless the storage is secure. All data will be processed locally or through resources provided by WPI.

Section VI: Timeline

Step 1: Get BERT to run (11/15/2023)

- Classifier model using nonsense text until database is accessible.
- Basic classifier model using transcripts.

Step 2: Figure out the link between wav2vec and BERT needed to allow commas to exist in transcribed text or set up Whisper model (12/1/23)

May require some fine tuning to figure out what the standard comma or period should be in the model.

Step 3: Train models (One can be trained on many languages and other on one language) based on either only English, only Spanish, both, or neither. (1/15/2023)

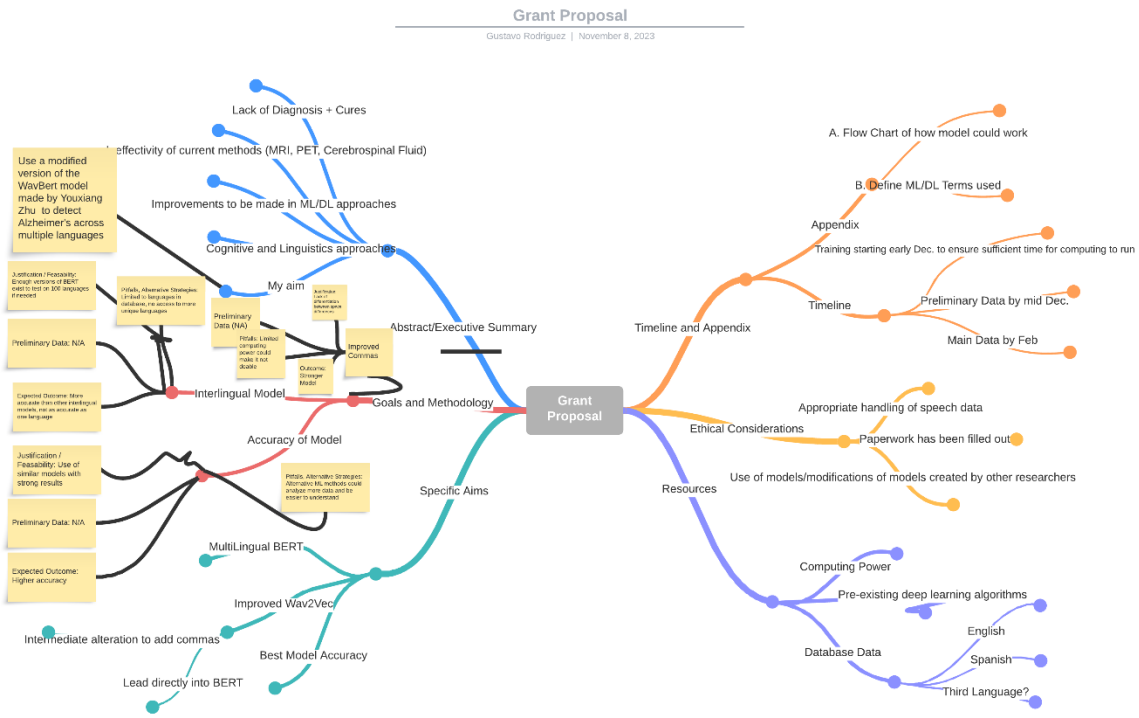
- This would involve simply inputting the transcripts from the wav2vec or Whisper model into the BERT classifier and training it. Prioritized are the multilingual BERT and training for ingenuity in the single language BERT.

Step 4: Run models (could take a long time; could gain access to added computing power depending on its necessity) (1/20/23)

Step 5: Analyze the data (2/15/23)

Section VII: Appendix

Appendix 1: Mind Map



Section VIII: References

Agbavor, F., & Liang, H. (2022). Predicting dementia from spontaneous speech using large language models. *PLOS Digital Health*, 1(12). https://doi.org/10.1371/journal.pdig.0000168

Anderson, S. R. (n.d.). How many languages are there in the world? - *Linguistic Society of America*. How many languages are there in the world? https://www.linguisticsociety.org/content/how-many-languages-are-there-world.

Balagopalan, A., Eyre, B., Rudzicz, F., & Novikova, J. (2020). To BERT or not to BERT: comparing speech and language-based approaches for Alzheimer's disease detection. *arXiv preprint* arXiv:2008.01551.

Melistas, T., Kapelonis, L., Antoniou, N., Mitseas, P., Sgouropoulos, D., Giannakopoulos, T., Katsamanis, A., Narayanan, S. (2023) Cross-Lingual Features for Alzheimer's Dementia Detection from Speech. Proc. *Interspeech 2023*, 3008-3012. 10.21437/Interspeech.2023-1934.

Meulemans, C., Leijten, M., Van Waes, L., Engelborghs, S., & De Maeyer, S. (2022). Cognitive writing process characteristics in Alzheimer's Disease. *Frontiers in Psychology*, 13, 872280. https://doi.org/10.3389/fpsyg.2022.872280

Pérez-Toro, P.A., Arias-Vergara, T., Braun, F., Hönig, F., Tobón-Quintero, C.A., Aguillón, D., Lopera, F., Hincapié-Henao, L., Schuster, M., Riedhammer, K., Maier, A., Nöth, E., Orozco-Arroyave, J.R. (2023) Automatic assessment of Alzheimer's across three languages using speech and language features. *Interspeech 2023,* 1748-1752. 10.21437/Interspeech.2023-2079.

Shinkawa, K., & Yamada, Y. (2018). Word Repetition in Separate Conversations for Detecting Dementia:A Preliminary Evaluation on Data of Regular Monitoring Service. *AMIA Joint Summits on Translational Science proceedings. AMIA Joint Summits on Translational Science,* 2017, 206–215.

Williams, E., Theys, C., & McAuliffe, M. (2023). Lexical-semantic properties of verbs and nouns used in conversation by people with Alzheimer's disease. *PLOS ONE*, *18*(8). https://doi.org/10.1371/journal.pone.0288556

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 Research & Therapy*, 14(1). https://doi.org/10.1186/s13195-022-01131-3

Zhu, Y., Obyat, A., Liang, X., Batsis, J. A., & Roth, R. M. (2021). WavBERT: exploiting semantic and mon-semantic speech using Wav2vec and BERT for Dementia detection. *Interspeech*, 2021, 3790–3794. https://doi.org/10.21437/interspeech.2021-332