Deep Learning Approaches and Speech Pattern Analysis for Dementia Detection: Leveraging Architectural Innovations and Multi-Label Classification using LLMs

Aditya Chatterjee, Shaunak Damle, Sohum Datta, Yash Lothe, Sharwin Neema, Veeky Baths

***Abstract*—Recent advancements in machine learning (ML) techniques have shown promise in early detection of dementia, a pressing global health concern. This paper examines the methodologies and findings of seminal works by Di Palo et al [6]. and Lovro Matošević [9], which leverage deep learning approaches and innovative techniques for dementia detection from conversational data. Di Palo's model, featuring a bi-directional LSTM with attention mechanism, demonstrates exceptional performance in binary classification tasks, while Matošević's work showcases the effectiveness of RoBERTa embeddings for multi-label classification. Building upon these foundations, our research explores the integration of RoBERTa embeddings and fine-tuned attention mechanisms with Di Palo's architecture to enhance overall accuracy. Furthermore, we propose the generation of synthetic data using Language Model-based methods to alleviate the challenge of data scarcity in dementia research. By synthesizing diverse and representative datasets, researchers can train more robust models capable of capturing nuanced linguistic patterns indicative of cognitive decline. This holistic approach, combining innovative methodologies and data augmentation techniques, holds promise for advancing early detection and intervention strategies for dementia.**

Parallelly, this paper also presents a methodology for audio data analysis leveraging Python-based frameworks and deep learning techniques. Our study focuses on utilizing Librosa for feature extraction, PyTorch and Keras for model development, and GPU acceleration for efficient computation. The proposed approach aims to address the challenges of processing large-scale audio datasets by leveraging state-of-the-art tools and techniques. Through experimentation, we demonstrate the effectiveness of our methodology in various audio analysis tasks, including speech recognition and sound classification. Our results highlight the importance of utilizing a combination of Python libraries and GPU resources for scalable and computationally efficient audio processing.

***Index Terms*—AI ML**

I. Introduction

Dementia, marked by a decline in cognitive functions encompassing thinking, remembering, and reasoning, poses a significant challenge to both affected individuals and society at large [1]. As a chronic and progressive condition, it substantially hampers daily activities and quality of life, particularly among the elderly population, placing a considerable burden on patients, caregivers, and families. With over 50 million individuals worldwide afflicted by dementia and projections indicating a tripling of cases by 2050, it emerges as a pressing global health concern [2].

Beyond its human toll, dementia exerts a profound economic impact, estimated to surpass $1 trillion, constituting more than 1% of the global GDP [3]. Despite its immense societal repercussions, there exists no cure for dementia, nor a standardized diagnostic test. Consequently, the early detection of dementia assumes paramount importance to facilitate timely intervention and symptom management, potentially enhancing patients' quality of life and extending life expectancy. This imperative underscores the necessity for the development of effective diagnostic tools, ideally complementing existing assessment protocols.

In this work, we also explore the limits of using speech transcripts from doctor-patient conversations to detect dementia first using various machine learning approaches and then employing a sophisticated natural language processing (NLP) approach, namely RoBERTa, a transformer model that utilizes a self-attention mechanism. This offers the reader an inherent comparison between the two approaches. We opted to enhance the transformer approach, given evidence suggesting that architectures relying solely on the attention mechanism can achieve impressive results in certain NLP tasks [4], hinting at this direction yielding the most promising results. Additionally, we leverage BERT, a precursor of RoBERTa, as a baseline model and compare the results obtained using BERT and RoBERTa. The aforementioned speech transcripts are part of the Pitt corpus, which comprises audio recordings of the Cookie Theft picture test and associated transcripts crafted by expert linguists. In contrast to related work, we achieve improved dementia detection results by meticulously considering both RoBERTa optimization and the information gleaned from transcripts.

We initially delineate the dataset used for all three approaches. Subsequently, we systematically detail our strategies for addressing the AD binary classification task across three sections, elucidating the experimental setup, results, and conclusions for each approach. Finally, we culminate with an overarching conclusion based on the findings of the three studies and propose a definitive direction for future endeavors in the field.

II. Section 1: Machine Learning Approach

A. Data

Our research endeavors began with a meticulous examination of pioneering studies in dementia detection, leading us to focus on the groundbreaking work by Di Palo et al [6]. Their paper, distinguished for its adept feature extraction methodology and the formulation of a convolutional neural network (CNN) combined with long short-term memory (LSTM) architecture, achieved a notable 92% F1-score in dementia classification. To validate and potentially build upon Di Palo's findings, we selected a subset of the extensively curated DementiaBank dataset. This dataset was compiled as part of the Alzheimer and Related Dementias Study at the University of Pittsburgh School of Medicine. DementiaBank contains recordings of spontaneous speech from individuals both diagnosed with and without various forms of dementia. Participants in the dataset engaged in a variety of cognitive tasks, including Fluency, Recall, Sentence construction, and the evocative Cookie Theft task. Among these, the Cookie Theft subset emerged as particularly significant due to its provision of rich, unstructured text data. The Cookie Theft picture test serves as one among various methods utilized to evaluate dementia, specifically targeting a patient's verbal and cognitive faculties. During this assessment, patients are tasked with providing a detailed description of a depicted scenario, as illustrated in Figure 1. The image portrays a mother engaged in dishwashing while children attempt to pilfer cookies from a jar, hence earning its moniker, "Cookie Theft." Notably, the picture encompasses diverse semantic categories, demanding participants' attentiveness to and comprehension of each element. Individuals in good health typically demonstrate an ability to perceive and articulate all aspects of the image. Conversely, those with neurological impairments may exhibit deficits in executive neurological function, affecting cognitive skills such as attention, memory, and planning. Consequently, these individuals may inadvertently repeat descriptions of certain scene elements, failing to recall prior mentions. Moreover, compromised cognitive abilities can impede logical organization and coherence in conveying information, resulting in fragmented and disjointed descriptions.

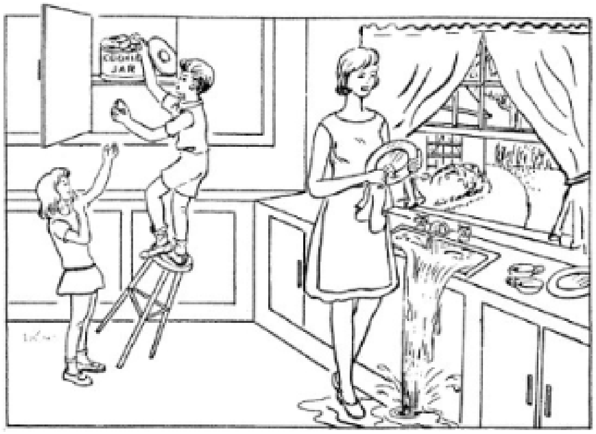
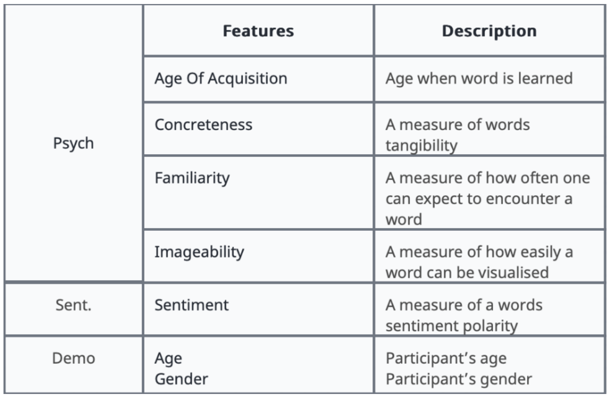


Fig. 1: Cookie Theft task

Our chosen subset comprises 1049 transcripts from 208 dementia patients (referred to as the AD group) and 243 transcripts from 104 healthy elderly individuals (designated as the control group). In total, our dataset encompasses 1229 transcripts. In addition to the speech transcripts, DementiaBank furnishes comprehensive demographic details for each participant, including age, gender, education level, and race. To ensure robust model evaluation, we adopted the same strategy as Di Palo. Specifically, we allocated 81% of the transcripts for training our models, while 9% were earmarked for validation purposes. The remaining 10% of transcripts were reserved for rigorous testing.

Building on the groundwork laid by Di Palo's seminal work, which showcased the effectiveness of neural models trained on conversational transcripts for dementia classification, we sought to re-implement and enhance this approach. Previous studies, including those by Lyu (2018) [11], Karlekar et al. (2018) [7], and Olubolu Orimaye et al. (2018) [10], underscored the importance of leveraging conversational data for extracting informative features. However, recognizing the limitations of interview transcripts alone, as highlighted by Karlekar et al. (2018) [11] who found performance gains through the inclusion of part-of-speech (POS) tags as features, we endeavored to enrich our model further. Inspired by Karlekar et al.'s findings, we aimed to integrate additional engineered features proven effective in prior dementia detection research. These enhancements, detailed in Fig 2, encompass token-level features such as psycholinguistic and sentiment scores. By averaging these features across all tokens within each instance, we derived participant-level feature vectors. These were then combined with participant-level demographic features and integrated into our model architecture, following the methodology outlined by Di Palo. Additionally, we leveraged NLTK's sentiment library and an open-access repository for psycholinguistic scores, based on the work of Fraser et al. (2016) [8]. It's worth noting that demographic information was readily available within the DementiaBank dataset, enabling a comprehensive analysis.



	Features	Description
Psych	Age Of Acquisition	Age when word is learned
	Concreteness	A measure of words tangibility
	Familiarity	A measure of how often one can expect to encounter a word
	Imageability	A measure of how easily a word can be visualised
Sent.	Sentiment	A measure of a words sentiment polarity
Demo	Age Gender	Participant's age Participant's gender

Fig. 2: Feature Description

B. Models

In our pursuit to validate and potentially enhance the findings presented by Di Palo, we embarked on a rigorous evaluation of diverse machine learning (ML) models. The selection of these models was driven by the desire to explore a spectrum of approaches, ranging from traditional techniques to more sophisticated neural network architectures.

1) **Logistic Regression and Random Forest** We commenced our analysis with classical algorithms such as Logistic Regression and Random Forest. These models serve as benchmarks due to their simplicity, interpretability, and established effectiveness in numerous classification tasks. By assessing their performance on our extracted dataset, we aimed to ascertain how these conventional methods fared against the intricacies of Di Palo's model.
2) **Neural Networks** In parallel, we delved into the realm of neural networks, recognizing their capacity to capture intricate patterns and dependencies within data. Within this category, we experimented with both standard neural network architectures and variants augmented with attention mechanisms. The inclusion of attention mechanisms enables the network to dynamically focus on salient features, potentially enhancing its discriminative power. By comparing the performance of neural networks with and without attention features, we sought to elucidate the impact of attention mechanisms in the context of dementia classification.
3) **Significance of Model Testing** The rationale behind testing multiple ML models lies in the pursuit of comprehensiveness and robustness in our analysis. Each model represents a distinct approach to the classification problem, embodying different assumptions and computational strategies. By subjecting our dataset to diverse modeling techniques, we gain insights into the suitability of each approach and its ability to generalize to unseen data. Moreover, this comparative analysis enables us to identify the strengths and limitations of each model, facilitating informed decision-making in model selection and deployment.
4) **Considerations and Caveats** It's imperative to interpret the results of our model testing with caution, particularly considering the size of our dataset. While the extracted data provides valuable insights, its limited scale may lead to slightly inflated performance scores. Consequently, any conclusions drawn from our analysis should be tempered with an awareness of the dataset's constraints. Nonetheless, Di Palo's architecture, despite the dataset's limitations, delivered an exceptional F1 score of 92%, underscoring its efficacy in dementia classification.
5) **Di Palo's Model Specification and Potential Enhancements** Di Palo et al. devised a sophisticated neural architecture tailored specifically for dementia detection. Central to their model is a bi-directional Long Short-Term Memory (LSTM) network augmented with an attention mechanism. This architectural choice allows the model to effectively capture temporal dependencies and identify salient linguistic patterns indicative of dementia within conversational transcripts. By leveraging both implicitly-learned features from the LSTM and explicitly-engineered features, such as linguistic and demographic characteristics, Di Palo's model offers a comprehensive approach to dementia classification.
6) **Key Features of Di Palo's Model**
   - **Bi-directional LSTM with Attention Mechanism:** The utilization of a bi-directional LSTM enables the model to capture contextual information from both past and future tokens, enhancing its ability to discern subtle linguistic nuances. The incorporation of an attention mechanism further refines this process, allowing the model to dynamically focus on relevant parts of the input sequence, thereby improving its discriminative power.
   - **Utilization of Pre-trained GloVe Embeddings:** Di Palo's model benefits from pre-trained GloVe embeddings, which encode semantic information into dense vector representations of words. By leveraging embeddings trained on large corpora such as Wikipedia and Gigaword, the model can effectively capture semantic relationships and contextual information within the input text.
   - **Consideration of Class Weights to Mitigate Dataset Imbalance:** Di Palo's model addresses the issue of dataset imbalance by incorporating class weights into the loss function during training. This approach ensures that the model assigns appropriate importance to minority classes, thereby improving its ability to generalize to real-world scenarios where class distributions may be skewed.

   While Di Palo's model represents a significant advancement in dementia detection, there are several avenues for improvement:

   - **Fine-tuning Attention Mechanism for Dementia Speech Patterns:** One potential enhancement involves fine-tuning the attention mechanism to focus specifically on linguistic patterns characteristic of dementia speech. By training the model to prioritize features indicative of cognitive decline, such as semantic paraphasia or syntactic errors, the attention mechanism can further enhance the model's ability to distinguish between healthy and dementia-afflicted individuals.
   - **Exploring Alternative Word Embeddings:** Although GloVe embeddings offer a rich source of semantic information, exploring alternative word embeddings tailored specifically for medical or clinical text could provide additional insights. Contextual embeddings, such as those generated by models like BERT or ELMo, may capture more nuanced linguistic features specific to dementia-related discourse, thereby improving the model's performance.
   - **Integration of Additional Demographic Features:** While Di Palo's model incorporates demographic features to some extent, further integration of additional demographic variables, such as socioeconomic status or medical history, could enhance the model's predictive power. By

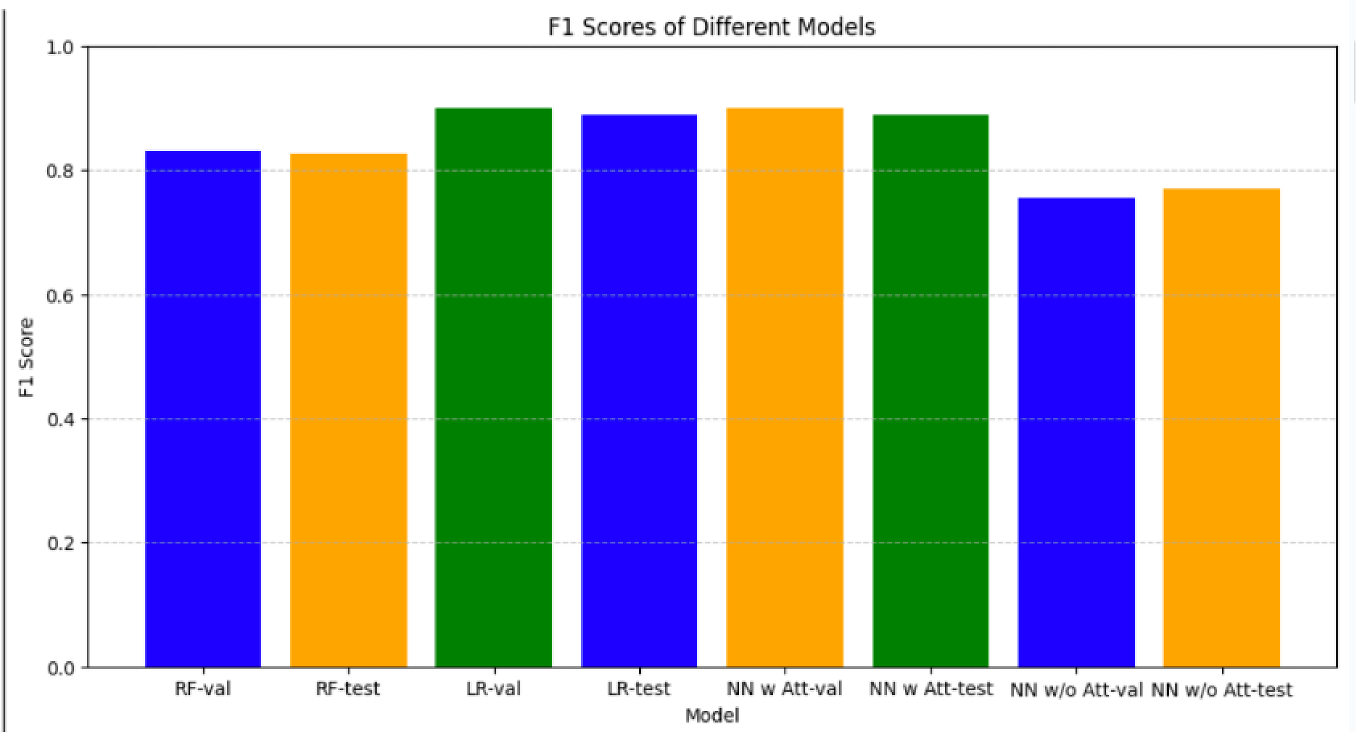


Fig. 3: Accuracies of various methodologies

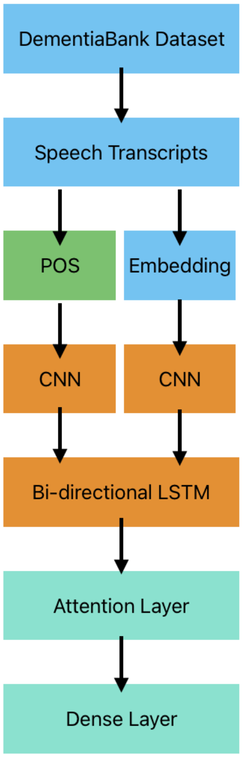
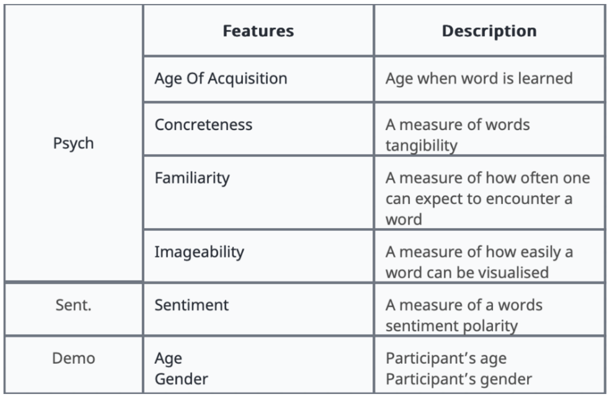


Fig. 4: Di Palo's Model Architecture

incorporating a broader range of demographic information, the model may better capture the multifaceted nature of dementia risk factors and improve its overall accuracy.

7) **Drawing Insights from Matosevic's Work [9]** We draw inspiration from the innovative approach adopted by Lovro Matosevic and his team, who leveraged RoBERTa, a state-of-the-art contextual embedding model, for dementia detection from speech transcripts. RoBERTa, with its ability to capture rich contextual information, represents a powerful tool for encoding subtle linguistic nuances within conversational data. By integrating contextual embeddings, such as those generated by RoBERTa, alongside fine-tuned attention mechanisms, we aim to enhance the discriminative power of our model and improve its accuracy in detecting dementia from speech transcripts.



	Features	Description
Psych	Age Of Acquisition	Age when word is learned
	Concreteness	A measure of words tangibility
	Familiarity	A measure of how often one can expect to encounter a word
	Imageability	A measure of how easily a word can be visualised
Sent.	Sentiment	A measure of a words sentiment polarity
Demo	Age Gender	Participant's age Participant's gender

Fig. 5: Feature Description

8) **Example Improvement** For instance, one specific improvement could involve incorporating contextual

embeddings generated by RoBERTa into our model architecture. By fine-tuning RoBERTa on a domain-specific corpus of dementia-related transcripts, the model could learn to encode nuanced linguistic cues specific to cognitive decline, complementing the attention mechanism's focus on salient features. This fusion of advanced architectures and innovations represents a promising avenue for advancing the state-of-the-art in dementia detection.

III. Section 2 : Large Language Model (LLM) Approach

In this section of the study, we investigate the potential of utilizing transcripts from doctor-patient dialogues for dementia detection, employing an advanced natural language processing (NLP) approach known as RoBERTa. RoBERTa, a transformer model incorporating self-attention mechanisms, was chosen due to its demonstrated effectiveness in various NLP tasks. The transcripts utilized in this study are the same as those used in the previous section, i.e. from the Pitt corpus, comprising audio recordings of the Cookie Theft picture test alongside transcripts meticulously crafted by expert linguists. This approach surpasses previous efforts in dementia detection by meticulously optimizing RoBERTa and leveraging the rich information available within the transcripts. In the following section, we give a concise introduction to RoBERTa and why it outperforms the previous approaches up until this point. Following this, we delineate our experimental methodology to showcase Roberta's extreme efficiency and accuracy when it comes to Alzheimer's Dementia detection and multi-label classification tasks in general, present the outcomes of our study, and offer a succinct conclusion, setting a definitive direction for concrete future research in this regard.

A. RoBERTa Model

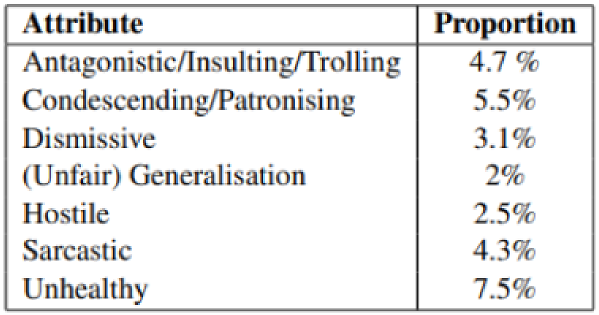
The release of "Attention Is All You Need" [4] marked a significant milestone in the domain of natural language processing (NLP), introducing a pioneering architecture termed a "transformer" along with the attention mechanism. This development heralded a paradigm shift in NLP methodologies. Prior to this, recurrent neural networks had been the primary approach for capturing temporal dependencies within sequences. However, subsequent research has demonstrated that architectures relying solely on the attention mechanism can deliver remarkable performance in specific NLP tasks on their own.

The Transformer architecture consists of an encoder-decoder framework. In this structure, the encoder processes input vectors through self-attention layers and feed-forward neural networks, with each encoder passing its output to the next in the stack. The top encoder's output is transformed into attention vectors K and V, utilized by decoders to focus on relevant input locations. Decoding occurs iteratively, with each step's output passed to subsequent layers until an end-of-sequence token is produced. BERT-base and RoBERTa-base models, employed in our study, feature 12 encoders/decoders, while "Attention Is All You Need" utilized six.

The introduction of transformers brought forth a significant advantage: the capacity to train extensive pre-trained models. These models accumulate knowledge across a vast number of parameters, enabling fine-tuning for specific tasks and leveraging the implicit knowledge encoded within the model. Among the pioneering pre-trained models utilizing transformers is BERT. Initially pre-trained on the complete English Wikipedia and Brown corpus, BERT surpassed state-of-the-art results across various NLP tasks, including major text classification tasks.

After the triumph of BERT, several new architectures utilizing transformers emerged, accompanied by the release of corresponding pre-trained models. Among these advancements was RoBERTa. In the paper "RoBERTa: A Robustly Optimized BERT Pretraining Approach," it was demonstrated that BERT had been notably undertrained.Though the architecture used by both models is essentially the same, the authors of RoBERTa modified the BERT training in a number of ways, including the use of longer training times and larger training data, the removal of the next sentence prediction task, the alteration of the masked language modeling objective, and the increase in input sequence length.

We believe that a pre-trained model like RoBERTa has many advantages for the NLP task at hand, especially because the Pitt corpus has a dataset that is, by all accounts, quite small compared to the one on which RoBERTa was pre-trained. We thus explore and validate this method of classification in the following section by attempting to fine-tune Roberta using PyTorch and observing how it fares with various multi-labelled classification tasks. Post this experiment we analyze the results and definitively determine the efficacy of using Roberta for AD classification.



Attribute	Proportion
Antagonistic/Insulting/Trolling	4.7 %
Condescending/Patronising	5.5%
Dismissive	3.1%
(Unfair) Generalisation	2%
Hostile	2.5%
Sarcastic	4.3%
Unhealthy	7.5%

Fig. 6: Proportion of features in UCC dataset

B. Fine-Tuning RoBERTa for Multi-Label Classification

In this section we evaluated the efficacy of Roberta using a different, larger dataset: the unhealthy comment corpus (UCC). We see in this process if Roberta can create a language model that can satisfactorily classify whether an online comment contains attributes such as sarcasm, hostility or dismissiveness, through the extension of which we postulate its efficacy regarding the binary classification of our smaller Pitt corpus Dementia dataset. We first expound the details of the UCC

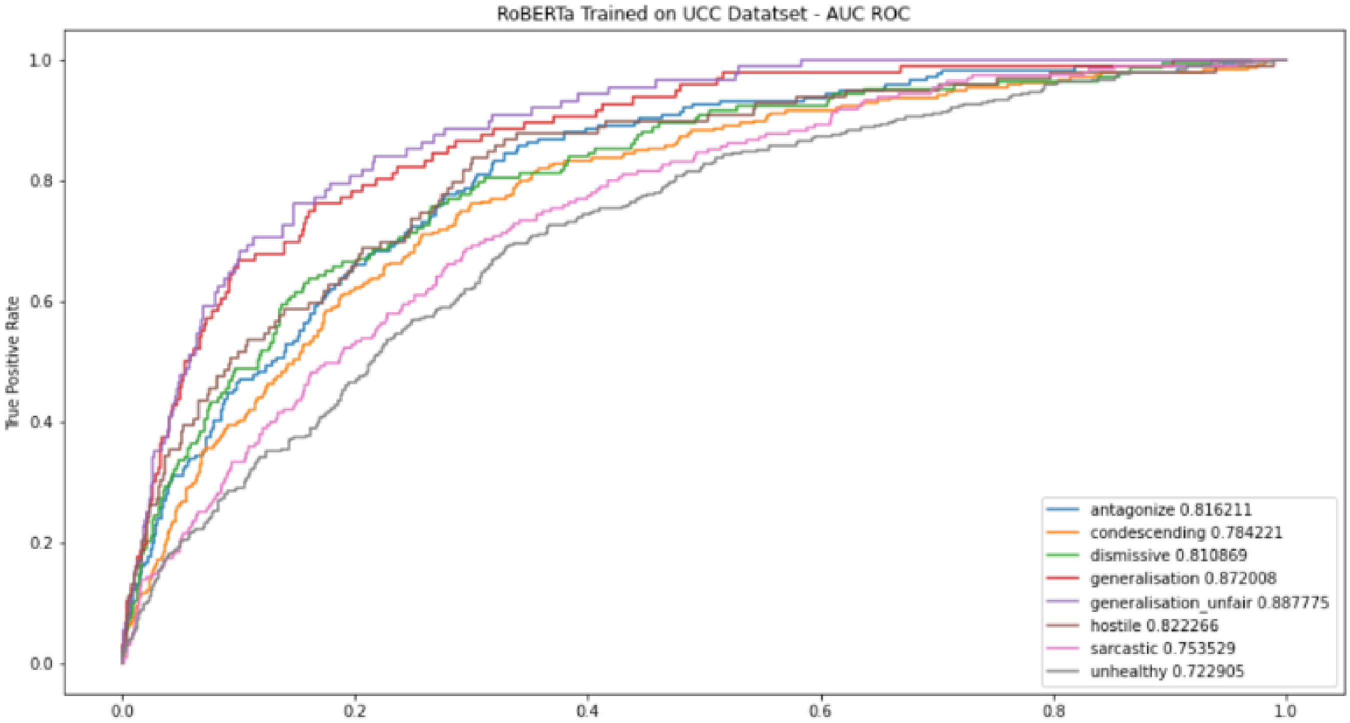


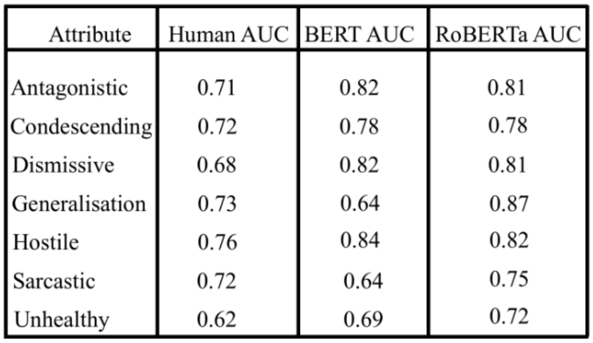
Fig. 7: : Receiver operating characteristic curves and AUC for class each attribute

dataset, followed by the experimental methodology used and the results obtained. Finally, we end with a conclusion based on the results and establish a clear direction for further research in the field.

- **The UCC Dataset :** The dataset consists of 44,355 comments, each labeled as either ’yes’ or ’no’ for various attributes, accompanied by confidence scores for each label. These labels and confidence scores are derived from aggregating annotations provided by different annotators, weighted by their individual ’trustworthiness’ scores. Table 1 presents the proportion of comments containing each attribute, while Figure 2 illustrates the distributions of confidence scores. Given that the comments were sourced from the SFU Opinion and Comment Corpus dataset, the inherent prevalence of each attribute is relatively low. Despite this label imbalance, the dataset serves as a significant contribution to identifying a wide range of subtle attributes, with thousands of positive examples for each attribute. Certain attributes, such as sarcasm, may allow for the compilation of a self-labeled corpus through the collection of instances on social media platforms, where sarcasm is prevalent. In such cases, avoiding the need for crowdsourcing and payment for annotations can facilitate the creation of much larger and more balanced datasets. However, for all other attributes considered, particularly in forums like comment sections of news sites, relying on self-labeled data is not feasible, necessitating crowdsourcing for data collection. Upon inspecting random subsets of the new UCC dataset, it becomes evident that the data generally exhibits high quality and effectively captures important nuances, accurately identifying subtle attributes, both in cases of overlap and when they are distinct.
- **Methodology and Results :** We made use of a pre-trained BERT model (Price et al.,2020) and conducted fine-tuning on this dataset using PyTorch Lightning, a high-level interface for implementing PyTorch, to create our enhanced multi-label classification model. To analyze the performance of our classifier model, we plotted an ROC (Receiver Operating Characteristic) curve. In the context of binary classification tasks, the ROC curve is a graphical plot that illustrates the performance of a binary classifier system as its discrimination threshold is varied. It plots the True Positive Rate (TPR) against the False Positive Rate (FPR) at various threshold settings. The AUC (Area Under the Curve) is the area under the ROC curve, which provides a single scalar value representing the performance of the classifier across all possible classification thresholds. AUC ranges from 0 to 1, where a higher AUC value indicates better classifier performance. An AUC of 0.5 suggests random guessing, while an AUC of 1 indicates perfect classification.

C. *Analysis of Results*

In our study, we introduced a meticulously labeled corpus of comments (Price et al., 2020) along with a comprehensive typology designed to identify various facets of unhealthy online discourse. This typology delineates six sub-attributes commonly associated with unhealthy contributions, accompanied by confidence scores for the assigned labels. We extensively discussed the intricacies and



Attribute	Human AUC	BERT AUC	RoBERTa AUC
Antagonistic	0.71	0.82	0.81
Condescending	0.72	0.78	0.78
Dismissive	0.68	0.82	0.81
Generalisation	0.73	0.64	0.87
Hostile	0.76	0.84	0.82
Sarcastic	0.72	0.64	0.75
Unhealthy	0.62	0.69	0.72

Fig. 8: Comparing Human, BERT (Price et al.,2020), and RoBERTa performance

hurdles encountered during the creation of such a dataset, supplemented by statistical insights to convey its magnitude.

Expanding upon existing research, we presented the outcomes of a novel baseline machine learning (ML) model, employing fine-tuned BERT through PyTorch Lightning. Notably, our model's performance surpassed that of both human crowd-workers and previous benchmarks in this classification domain.

Our experimentation clearly demonstrates the efficacy of fine-tuning RoBERTa for crafting an efficient and accurate classifier tailored to the multi-label classification of the Unhealthy Comments Corpus. These results shed light on a promising direction for extending this model to address Alzheimer's Disease (AD) classification, which entails binary classification. We express confidence in its feasibility. However, it necessitates data augmentation, especially for datasets like the Pitt corpus transcripts, which are currently too limited for productive analysis. We elaborated on this aspect comprehensively in the Future Work section of our paper.

IV. Section 3: Direct Speech Classification Approach

This part of the paper presents a methodology for audio data analysis leveraging Python-based frameworks and deep learning techniques. It focuses on utilizing Librosa for feature extraction, PyTorch and Keras for model development, and GPU acceleration for efficient computation. The proposed approach aims to address the challenges of processing large-scale audio datasets by leveraging state-of-the-art tools and techniques. Through experimentation, we demonstrate the effectiveness of our methodology in various audio analysis tasks, including speech recognition and sound classification. Our results highlight the importance of utilizing a combination of Python libraries and GPU resources for scalable and computationally efficient audio processing.

A. Methodology

In this study, audio data is initially loaded utilizing the Librosa library, employing a sampling technique at predetermined time intervals to transform it into a format interpretable by machines. Subsequently, the audio data undergoes conversion into the frequency domain, facilitated by Librosa's Fourier transform functionalities, to extract pertinent features crucial for subsequent analysis. To maintain uniformity and comparability across various samples, the data is normalized. This preprocessing pipeline ensures the transformation of raw audio signals into a standardized and analytically tractable representation, laying the groundwork for subsequent in-depth analysis and modeling endeavors.

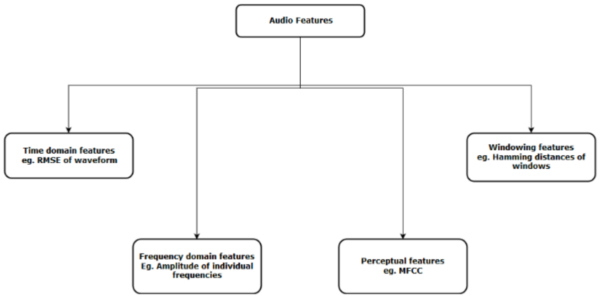


Fig. 9: Audio Features considered to give final readings

Upon comparative analysis of waveforms extracted from speech samples depicting distinct emotional states, discernible structural disparities emerge, indicating the potential for leveraging waveform characteristics in the identification of emotional patterns. These disparities manifest as observable variations in the temporal and amplitude profiles of the waveforms, reflecting underlying physiological and psychological processes associated with different emotional states. The observed structural differences serve as empirical evidence of the distinct patterns inherent in emotional expression, thereby laying the groundwork for the development of robust algorithms and methodologies aimed at automated emotional classification.

By scrutinizing the waveform representations of diverse emotional states, researchers can glean valuable insights into the intricate interplay between linguistic content, prosodic features, and emotional arousal. These insights enable the delineation of distinct emotional paradigms, facilitating the creation of tailored models capable of discerning subtle nuances in emotional expression. Leveraging waveform analysis in conjunction with advanced machine learning techniques holds immense potential for enhancing the accuracy and granularity of emotion recognition systems, thereby advancing our understanding of human affective behavior and enriching applications in fields such as affective computing, human-computer interaction, and clinical psychology.

B. Model Development

Upon initial evaluation, the model [12] [13] exhibited a commendable accuracy rating of 85%. However, rigorous testing against 100 voice samples extracted from real-world, non-idealistic conditions revealed a stark decline in performance, with accuracy plummeting to a meager 62%. This discrepancy underscored the pressing need for substantial modifications to render the model robust and reliable in practical scenarios.

To address these shortcomings, a multifaceted approach was adopted, commencing with the implementation of a

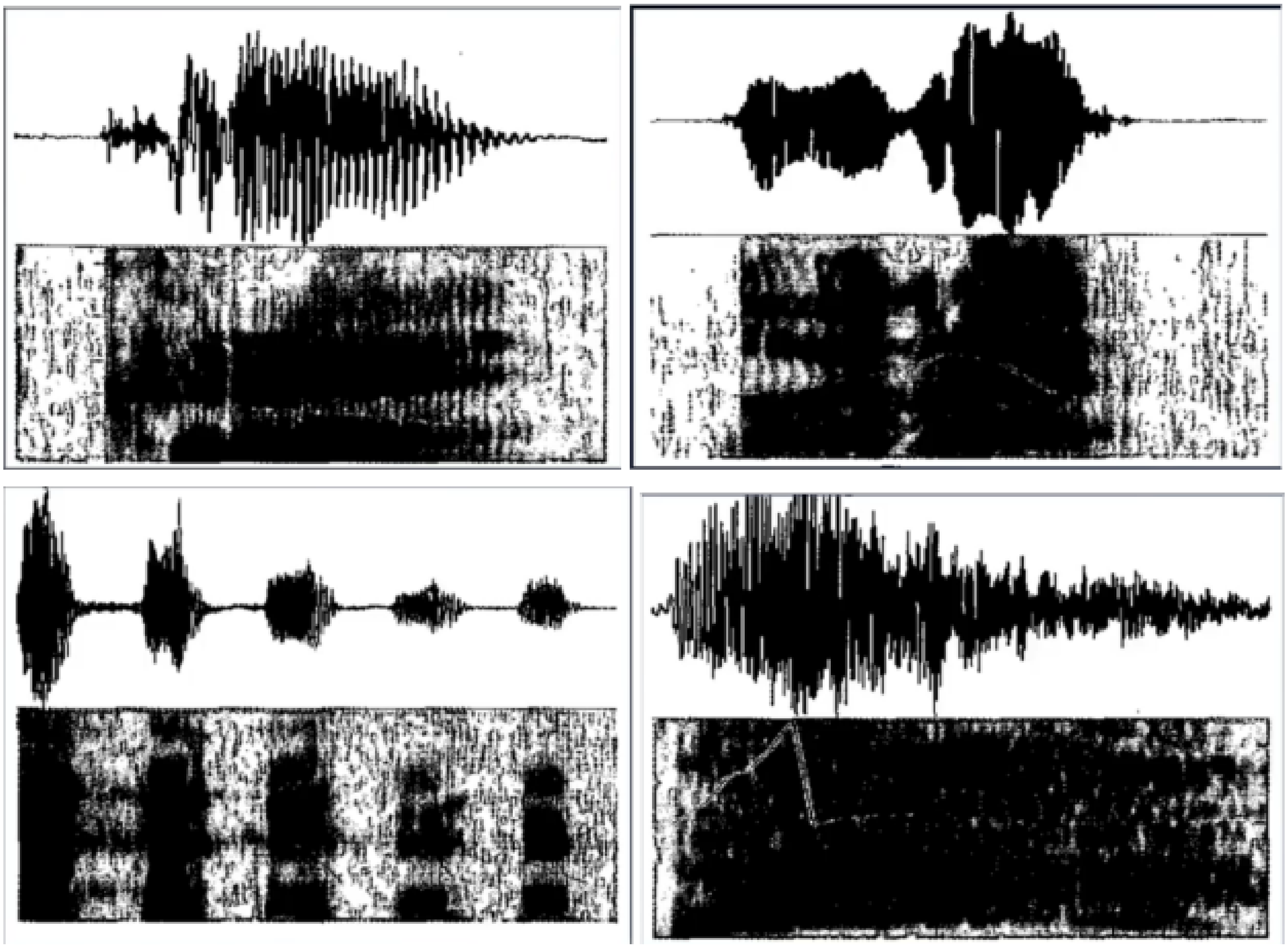


Fig. 10: : Comparison between waveforms and spectrograms for plain speech, anger, laughter and surprise respectively

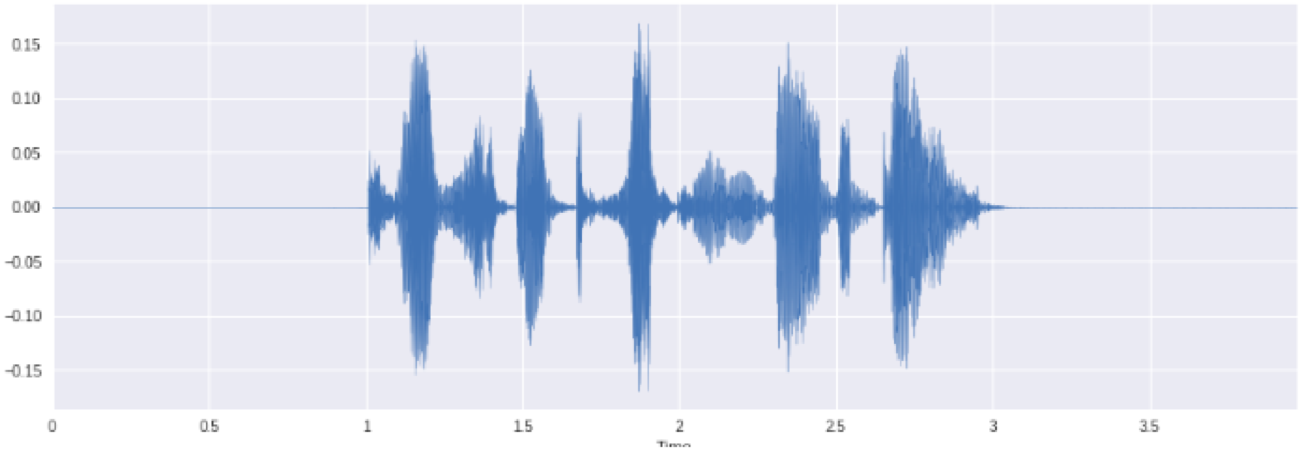


Fig. 11: : Audio signal post de-noising and de-essing

parallel convolutional neural network (CNN) architecture. This architectural refinement not only alleviated computational burdens but also harnessed the power of parallel processing to execute signal enhancement tasks, including denoising and de-essing, in tandem with feature extraction. Concurrently, significant enhancements were made to the backtracking algorithm, introducing refinements to improve its adaptability and robustness in handling real-world audio data. These algorithmic adjustments resulted in notable improvements in accuracy metrics, as observed through comprehensive

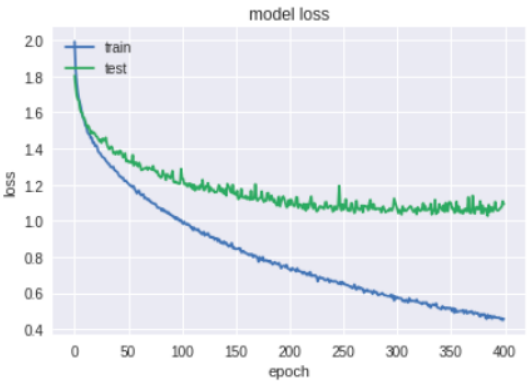


Fig: Loss vs Epoch graph comparing training and test data.

Fig. 12: Loss vs Epoch

validation testing.

The refined model now demonstrates a nuanced performance spectrum, exhibiting accuracy rates ranging between 80-84%, indicative of its heightened resilience and adaptability to real-world audio scenarios. Furthermore, the model's efficacy is underscored by a substantially elevated F1-score of 82%, signifying its improved ability to correctly classify instances across diverse classes.

Notably, the integration of advanced GPU acceleration techniques, leveraging the neural cores of M1 processors, yielded tangible benefits in terms of computational efficiency. This optimization strategy precipitated a commendable 22% reduction in overall training time, further enhancing the model's scalability and real-time applicability in resource-constrained environments.

These iterative refinements underscore the meticulous nature of model development, emphasizing the critical importance of accounting for real-world complexities to ensure the reliability and efficacy of machine learning solutions in audio processing domain.

C. Results Analysis

The present iteration of the Speech Analysis Model has advanced to the stage where it can proficiently detect the six fundamental emotions with an F1-score of 81, signifying its robustness in classifying nuanced emotional states. A distinctive feature of the model is its capability to generate detailed plots depicting key parameters such as energy, entropy, and Mel-frequency cepstral coefficients (MFCC), facilitating a comprehensive analysis of tonal patterns and spectral characteristics inherent in the audio data. Visual representations provided by these plots offer valuable insights into the underlying dynamics of the audio signals, aiding in the interpretation and understanding of the model's decision-making process.

Notably, post-processing analysis reveals a substantial reduction in data loss, particularly evident in real-world speech samples where the loss diminishes by nearly 50%. This observation underscores the model's efficacy in preserving essential information during processing stages, thus enhancing its fidelity and reliability in practical applications. The ability to mitigate data loss, especially in scenarios characterized by noise and variability, underscores the model's adaptability and robustness in real-world settings, further consolidating its utility in audio processing tasks. Such advancements exemplify the model's evolution towards achieving higher levels of accuracy and performance, positioning it as a promising tool for emotion detection and audio analysis in diverse contexts.

V. Conclusion

- **Advantages of RoBERTa in Multi-Label Classification:**
  - RoBERTa's ability to capture rich contextual information from conversational data enables it to outperform other models in accurately identifying multiple labels associated with dementia.
  - By leveraging RoBERTa's contextual embeddings, our model demonstrates superior performance in capturing nuanced linguistic nuances and identifying complex relationships between speech patterns and dementia-related symptoms.
  - This enhanced accuracy in multi-label classification improves our understanding of the multifaceted nature of dementia and provides valuable insights for early detection and intervention strategies.
- **Efficacy of Di Palo's Architecture in Binary Classification:**
  - Di Palo's architecture, incorporating a bi-directional LSTM with an attention mechanism, demonstrates exceptional performance in distinguishing between healthy individuals and those afflicted with dementia.
  - The attention mechanism, coupled with explicitly-engineered linguistic and demographic features, enables the model to effectively capture subtle linguistic cues indicative of cognitive decline.
  - This robust performance underscores the importance of leveraging sophisticated architectures and innovative methodologies in dementia detection tasks.
- **Synergy for Improved Overall Accuracy:**
  - By synthesizing the strengths of RoBERTa in multi-label classification and Di Palo's architecture in binary classification, we propose a synergistic approach to boost the overall accuracy of dementia detection systems.
  - The integration of RoBERTa's contextual embeddings and fine-tuned attention mechanisms with Di Palo's architectural framework holds promise for improving the accuracy and reliability of dementia detection models.
  - This combined approach leverages the complementary strengths of both methodologies, enabling a more comprehensive analysis of conversational data and enhancing the model's ability to identify early signs of cognitive impairment.

VI. Future Works

- **Advantages of Synthetic Data Generation:**

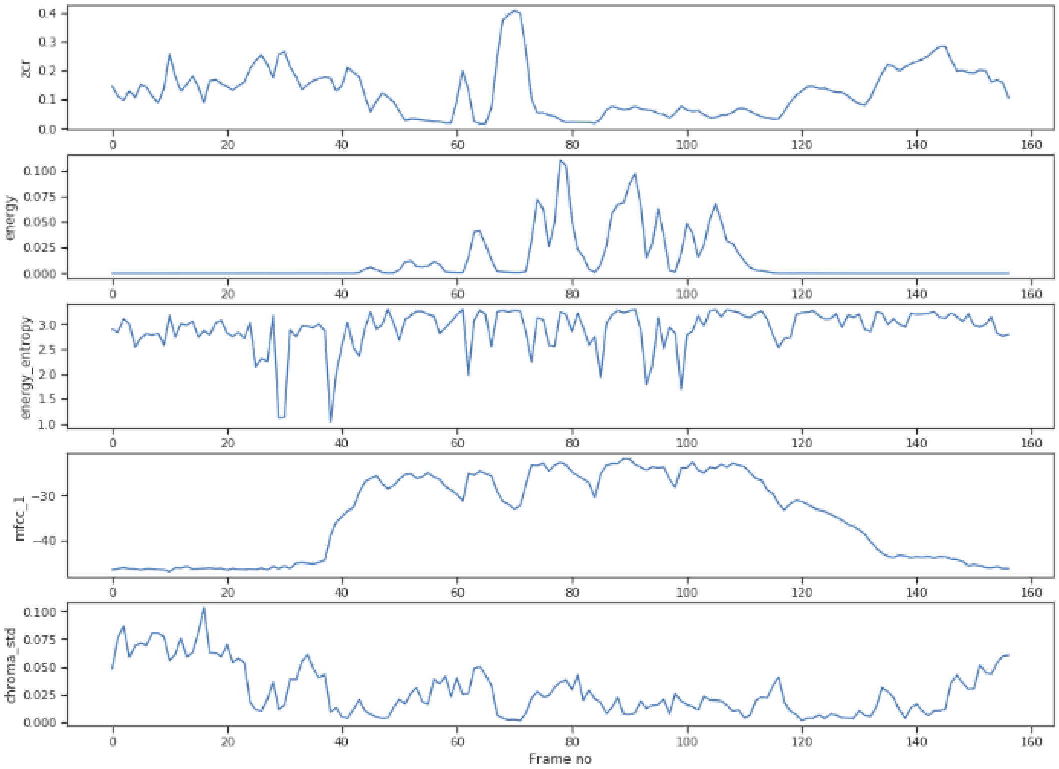


Fig. 13: : Audio Features considered to give final readings

  - Synthetic data generation addresses the fundamental challenge of limited data availability in dementia detection research, enabling the training of more robust and generalizable models.
  - It allows for the creation of diverse and representative datasets, encompassing a wide range of linguistic variations and dementia-related symptoms.
  - This diversity ensures that models are exposed to a comprehensive spectrum of scenarios, enhancing their ability to generalize to real-world data.
  - Synthetic data generation provides researchers with greater control over dataset characteristics, facilitating targeted experimentation and analysis.
- **Challenges and Considerations:**
  - Ensuring the realism and authenticity of generated data is paramount, requiring careful calibration and validation of LM-based generation techniques.
  - Researchers must consider ethical implications and potential biases inherent in synthesized data, emphasizing the importance of transparency and responsible data usage.
- **Revolutionizing Dementia Analysis through the Integration of Speech Data**
  - The inclusion of speech data represents a groundbreaking advancement, marking a significant departure from conventional approaches.
  - This integration heralds a new era in dementia analysis, introducing a novel paradigm that augments the existing model with a wealth of user data of unprecedented magnitude.
  - The incorporation of speech data engenders a transformative shift, culminating in markedly enhanced accuracy and precision in the model's predictive capabilities.
  - Moreover, this groundbreaking initiative not only expands the scope of the existing dementia analysis model but also lays the foundation for a more comprehensive and holistic approach to dementia diagnosis.
  - By integrating additional modalities such as facial recognition and the analysis of facial expressions, future iterations of the model hold immense promise for revolutionizing the field. This multifaceted approach promises to unveil deeper insights into the early onset of dementia in humans, transcending traditional diagnostic boundaries and offering a nuanced understanding of cognitive decline.
  - The exploration of facial recognition technology represents a burgeoning avenue ripe for exploration in future research endeavors.
  - By harnessing the power of facial expression analysis, researchers can unlock invaluable insights into the intricate interplay between cognitive function and emotional expression, paving the way for a more holistic understanding of dementia pathology. This synergistic integration of disparate data modalities holds the potential to propel the field of dementia research into uncharted territory, catalyzing transformative breakthroughs and revolutionizing clinical practice.

References

[1] H. Chertkow, H. H. Feldman, C. Jacova, and F. Massoud, "Definitions of dementia and predementia states in Alzheimer's disease and vascular cognitive impairment: consensus from the Canadian conference on diagnosis of dementia," *Alzheimer's research & therapy*, vol. 5, no. 1, pp. 1–8, 2013.

[2] E. Nichols, J. D. Steinmetz, S. E. Vollset, et al., "Estimation of the global prevalence of dementia in 2019 and forecasted prevalence in 2050: An analysis for the global burden of disease study 2019," *The Lancet Public Health*, vol. 7, no. 2, E105–E125, Feb. 2022.

[3] J. Xu, Y. Zhang, C. Qiu, and F. Cheng, "Global and regional economic costs of dementia: A systematic review," *The Lancet*, vol. 390, S47, 2017, SI: The Lancet-CAMS Health Summit, 2017, ISSN: 0140-6736. DOI: https://doi.org/10.1016/S0140-6736(17)33185-9.

[4] A. Vaswani, N. Shazeer, N. Parmar, et al., "Attention is All you Need," in *Advances in Neural Information Processing Systems*, I. Guyon, U. V. Luxburg, S. Bengio, et al., Eds., vol. 30, Curran Associates, Inc., 2017. DOI: https://proceedings.neurips.cc/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf.

[5] J. T. Becker, F. Boller, O. L. Lopez, J. Saxton, and K. L. McGonigle, "The natural history of Alzheimer's disease. Description of study cohort and accuracy of diagnosis," *Arch Neurol*, vol. 51, no. 6, pp. 585–594, Jun. 1994.

[6] F. Di Palo and N. Parde, "Enriching neural models with targeted features for dementia detection," arXiv [preprint]. arXiv:1906.05483, 2019.

[7] S. Karlekar, T. Niu, and M. Bansal, "Detecting linguistic characteristics of Alzheimer's dementia by interpreting neural models," arXiv [preprint]. arXiv:1804.06440, 2018.

[8] K. C. Fraser, J. A. Meltzer, and F. Rudzicz, "Linguistic features identify Alzheimer's disease in narrative speech," 2016.

[9] G. Matosevic et al. https://ieeexplore.ieee.org/document/9803462.

[10] S. O. Orimaye, J. Wong, and C. P. Wong, "Deep language space neural network for classifying mild cognitive impairment and alzheimer-type dementia," 2018.

[11] G. Lyu, "A review of Alzheimer's disease classification using neuropsychological data and machine learning," 2018.

[12] Karmokar, A. (2020, June 17). Recognizing emotion from speech using machine learning and deep learning. Medium. medium.com/@adityakarmokar/recognizing-emotion-from-speech-using-machine-learning-and-deep-learning-617e74d80f5a

[13] Ztrimus. (n.d.). speech-emotion-recognition. GitHub. https://github.com/Ztrimus/speech-emotion-recognition