Acoustic Emotion Analysis for Novel Detection of Alzheimer's Dementia

Marek Sviderski, Basel Barakat, Becky Allen
Faculty of Technology, University of Sunderland
Sunderalnd, UK
bh37sy@student.sunderland.ac.uk,{basel.barakat,becky.allen}@sunderland.ac.uk

***Abstract*—Abstract Alzheimer's Dementia (AD) presents significant diagnostic challenges, particularly in terms of early detection, where traditional methods often fall short due to their invasiveness and high costs. This study introduces a novel, non-invasive approach utilising emotional expressions captured from audio recordings to detect AD. Employing advanced digital signal processing techniques, including Facebook's Denoiser model, and deep learning methodologies through models such as Wav2Vec 2.0, this research aims to identify emotional disturbances that precede cognitive decline. Audio recordings were transformed into a tabular format, suitable for machine learning analysis. The LGBM Classifier and ensemble methods demonstrated superior performance, with the LGBM Classifier achieving the highest F1 score of 0.93 and an accuracy of 0.89 on a 3.5-second segment. These findings underscore the potential of combining emotional analysis with machine learning to enhance early AD detection, offering a simpler, more accessible diagnostic tool than currently available methods.**

***Index Terms*—Alzheimer's Dementia, acoustic emotion recognition, machine learning, audio processing, deep learning.**

I. Introduction

Alzheimer's Dementia (AD) is a progressive neurological disorder that significantly impairs cognitive function, affecting millions of individuals and their families worldwide. The condition not only leads to a decline in memory, thinking, and behavioural abilities but also places a considerable emotional and financial burden on caregivers and healthcare systems [1], [2]. Early detection of Alzheimer's is crucial as it can provide a window for timely intervention, potentially slowing the progression of the disease and improving the quality of life for patients. In the UK alone, the disease currently affects more than 944,000 people, with this number expected to rise due to the general increase in average life expectancy [3].

Despite advances in medical technology, current methods for detecting AD face significant challenges. Traditional diagnostic approaches, such as cognitive testing and neuroimaging, are often invasive, expensive, and not universally accessible [4], [5]. In response to these challenges, this study proposes the use of emotional representation extracted from audio recordings as a novel approach for AD detection. Emotional disturbances are among the early signs of Alzheimer's, manifesting before the onset of more pronounced cognitive symptoms. By analysing the acoustic properties of emotions expressed in speech, we aim to uncover subtle patterns that could indicate the presence of AD.

Depression is one of the risk factors connected with AD diagnosis. Depressive symptoms can occur in 20-30% of all AD patients [6]. Furthermore this fact is also supported by a study by Crump et al., [7] where the results indicate that an AD patient is twice as likely to have a risk of "major depression" in oppose to the healthy control (HC). With these facts in mind, we can assume that if we can employ AI to detect depression we can create models which would help with detecting the early stages of AD. If we expand the term depression to recognise the patient's full emotional state we can use the recordings traditionally used for acoustic detection and add a new dimension where we analyse the emotional state from the acoustic recordings for the pursuit of distinguishing between AD and HC patients.

The main contribution of this research is the enhancing of the field of AD detection through new acoustic emotional analysis. We employ advanced digital signal processing techniques, to first denoise the recordings, to improve the clarity and quality of the audio data. Furthermore, we utilise deep learning models to analyse the emotional content in these recordings, thereby transforming the audio data into a format that is amenable for machine learning applications. In order to use the traditional machine learning models we converted the time-series audio data into a tabular format in pursuit to predict AD presence effectively.

This paper is structured as follows: Section II contains the summary of the related work in the problem area, Section III outlines our methodology, detailing the process of data collection, denoising, emotion processing, and data transformation. Section IV presents the results and findings of the study, showcasing the performance of different machine learning models. Section V discusses the implications of our findings, limitations of the current study, and potential directions for future research. Finally, Section VI concludes the paper by summarising the key outcomes and the prospective impact of this research on the early detection of AD.

II. Related Work

In the pursuit of effective Alzheimer's Dementia (AD) diagnosis, various methods each present unique advantages and challenges. The most prevalent involves neuroimaging techniques such as MRI, fMRI, and PET scans, which are highly informative but often costly and time-consuming, pos-

ing practical challenges for healthcare systems like the NHS [8].

Alternative methods include EEG signal analysis, which offers a non-invasive option but requires complex interpretation. Cognitive function tests, such as drawing a clock face, provide straightforward assessments but may lack the sensitivity needed for early detection. Genetic testing offers insights into susceptibility to AD, but its complexity and uncertainty regarding disease manifestation limit its utility [9].

In contrast, speech and language pattern analysis presents a viable, less expensive alternative. This approach offers stability and accessibility, as it can be performed rapidly and non-invasively, requiring minimal resources compared to neuroimaging or genetic testing. Additionally, the assessment of speech can often be conducted in the patient's native environment, reducing the need for clinical visits and specialised equipment.

Importantly, our approach incorporates the analysis of depression, a condition that frequently precedes or accompanies AD. By recognising and integrating emotional states such as depression into our diagnostic model, we acknowledge the intricate relationship between mental health conditions and neurodegenerative diseases. Depression is not only a common comorbidity but can also be an early symptom of AD, further justifying its inclusion in our research.

In the category of speech analysis the detection predominantly focuses on the linguistic and acoustic features of speech. These techniques endeavour to identify the most effective feature representation be it acoustic, linguistic, or a hybrid of both. The objective is to facilitate either binary classification (healthy control (HC) vs. AD), multiclass classification (HC vs. Mild Cognitive Impairment (MCI) vs. AD), or regression analysis, with the Mini-Mental State Exam score (MMSE) often serving as the target variable.

The model architecture reported by Khan [10] currently leads in terms of accuracy score. It utilises a Stacked Deep Dense Neural Network Model derived from audio transcript data. The data processing involves an embedding layer, individual pathways through CNN, CNN+biLSTM, and biLSTM with an attention mechanism, which are then concatenated into a single vector. This feeds into a Dense NN model culminating in a Sigmoid activation function for HC or AD classification. Their results boast an accuracy score of 93.31% and an F1 score of 85.69%. While impressive, this solution does not incorporate features from the audio itself and is characterised by a complexity that poses challenges in terms of explainability and generalisability across different languages.

A more balanced approach that incorporates both acoustic and linguistic modalities is presented by Ilias and Askounis [11]. In their study, audio data is transformed into Log-Mel Spectrogram images, while textual data undergo transformation via BERT. Their model, enhanced by Co-attention, Multimodal Shifting Gate, and a Self-Attention variant, achieved an F1 score of 89.94% and an accuracy of 90%. Additionally, they provide a review of solely acoustic models, which achieve accuracy scores ranging from 54.17% to 81.25% and F1 Scores between 54.17% and 70%. This comparison underscores the underdevelopment of purely acoustic models when set against those employing both modalities and achieving state-of-the-art results.

Moreover, sentiment or emotional analysis, particularly through text transcription as shown by Liu [12], and video data analysis by Fei [13] for MCI detection, reveal the scope but also the gaps in applying emotional analysis directly to acoustic recordings.

This study seeks to address this gap by concentrating on the emotional content of speech as a novel indicator for early AD detection. By doing so, it presents the opportunity for a diagnostic tool that is not only simpler but also more accessible compared to the methodologies currently in use.

III. Methodology

This study utilises recordings from the Pitt corpus [14], comprising 1283 mp3 files from patients engaged in four cognitive tasks: the Cookie Theft picture description, recall, sentence construction, and verbal fluency tasks.

The recordings, initially in a compressed mp3 format, were converted to wav format for enhanced processing compatibility and subsequently denoised using Facebook's Denoiser model [15]. This preprocessing step was crucial for minimising background noise and enhancing the clarity of the patients' speech, particularly given the notable presence of hiss in the original Pitt corpus recordings.

A. Acoustic emotion processing

To investigate the emotional states from our audio recordings, we employed the "Speech Emotion Recognition By Fine-Tuning Wav2Vec 2.0" model developed by Calabrés from Hugging Face [16], supplemented by Facebook's "Wav2Vec2-XLSR-53" model [17]. This integration enabled us to not only identify but also quantify the intensity of emotions over time, thereby constructing a detailed emotional profile for each patient. Initially, the Calabrés model outputs a single emotion label such as angry, calm, disgust, fearful, happy, neutral, sad, or surprised, reflecting the primary detected emotion.

This ensemble approach provided a more nuanced analysis, allowing us to capture a time-series distribution of emotional probabilities for each segment, thus offering an extensive emotional profile for each recording. We started by segmenting each recording into 3-second chunks with a 0.5-second overlap. This specific duration was chosen to align with the chunk lengths used by Ilias and Askounis [11] in their methodology for spectrogram image generation, ensuring consistency with established practices while exploring its applicability to our model. The overlap helps mitigate potential edge artifacts between segments. To further validate our approach and adapt to the differing nature of our analysis compared to that of Ilias and Askounis, we also tested segments of 4 seconds and 2 seconds.

For the audio processing, each segment was transformed into a numerical waveform array. In cases of stereo recordings, we converted the audio to mono by averaging the two

channels, as our models required a single-channel input. These waveforms were then processed using a feature extractor tailored for the "Wav2Vec2-XLSR-53" architecture, which involved normalising the audio samples and adjusting the sampling rate to that of the original recordings, preparing them for effective model analysis. Figure 1 illustrates this entire process, from the initial audio waveform of the patient's speech through to the final emotional probability distribution. The diagram captures the essential steps involved, including audio segmentation, processing, and emotional probability analysis, effectively depicting how raw audio data is transformed into a structured format that can be utilized for further machine learning analysis and AD diagnosis.

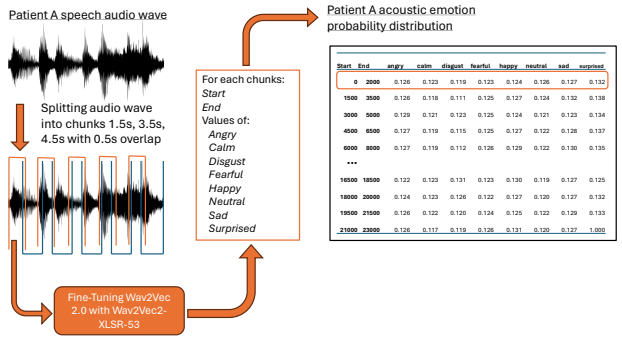

Fig. 1. Diagram illustrating the transformation process from the patient's speech audio wave to the acoustic emotion probability distribution. Each step from segmenting the audio, processing it through deep learning models, to obtaining emotional probabilities is shown.

Once prepared, the data was input into our emotion classification model. To ensure efficiency, the model was hosted on the same computational device. The outputs, or logits, from the model were converted into a probability distribution using a softmax function. This final step provided us with a detailed probabilistic breakdown of emotional states for each audio segment, enhancing our understanding of the emotional dynamics within the recordings. This comprehensive processing pipeline not only supports robust emotion recognition but also enhances the depth of our analysis, making it a pivotal component of our methodology for detecting AD through acoustic emotion analysis.

B. Transformation of the Time-Series Format into Tabular

To facilitate the use of traditional machine learning algorithms, we transformed the time-series data into a tabular format. This decision was driven by the desire to apply a novel approach to AD detection in a healthcare context, where the explainability and interpretability of machine learning models are crucial. These models allow for an examination of feature importance and provide insight into the reasoning behind decision-making processes. The first conversion strategy involved computing statistical metrics for each emotion, such as mean, standard deviation, minimum, maximum, mode, skewness, kurtosis, median, quartiles, interquartile range, range, variance, standard error of the mean, and coefficient of variation. This comprehensive feature set per emotion was complemented by demographic information and the target variable (Healthy Control or AD), which includes details such as age, sex, education level, and race of the patient.

The second strategy was designed to capture the most prominent emotion in each segment by selecting the emotion with the highest probability value from the distribution for each time slice. This effectively created a time-series dataset of the most likely emotional category over time, providing a categorical dataset that was then padded and label-encoded for analysis.

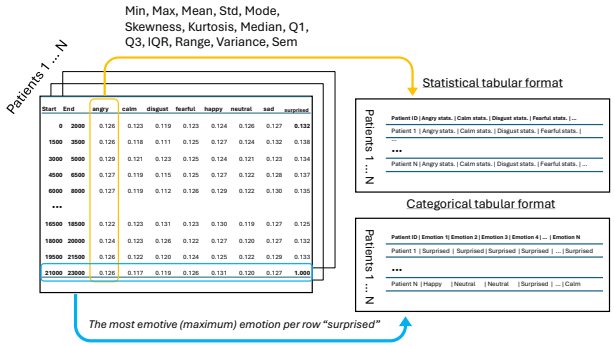

Fig. 2. Schematic representation of the transformation of time-series emotional data into tabular formats. The top part of the figure illustrates the statistical tabular format where various statistical metrics are calculated for each emotion, and the bottom part shows the categorical tabular format focusing on the most dominant emotions.

Both data transformation strategies are illustrated in Figure 2, which visually outlines the process of converting the nuanced time-series data into two distinct tabular formats: one statistical and one categorical. This figure helps convey the methodological steps involved in data preparation, from raw audio processing to the final structured data formats used for machine learning classification.

C. Proposed Experimental Design

We developed two types of tabular dataframes based on the aforementioned transformation strategies. For each type, we generated three sets of dataframes corresponding to different segment lengths, resulting in a total of six unique tabular dataframes. This array of dataframes allowed us to comprehensively test the impact of segment length on model performance.

For each dataframe, we applied several machine learning models, including Logistic Regression, K-Neighbors Classifier, SVC, Decision Tree Classifier, Random Forest Classifier, LGBM Classifier, and XGB Classifier. These models were chosen for their robustness and suitability for classification tasks. Additionally, we implemented both hard and soft ensemble models, as well as a stacking model, to enhance prediction accuracy. In the stacking ensemble, we incorporated the same sequence of models used previously but introduced

an additional Logistic Regression model as the final estimator to refine the combination of predictions.

Each model was evaluated using Stratified K-fold validation with five folds to address the imbalanced nature of the labels, with performance metrics such as F1 score and accuracy calculated for each fold. The final evaluation of model performance was determined by averaging these metrics across all folds.

This structured approach to data preparation, model application, and performance evaluation underscores our commitment to developing a robust methodology for AD detection that leverages both the depth of emotional data and the strengths of traditional machine learning techniques.

IV. Results and Findings

This section presents the performance outcomes of our experimental analysis, which aimed to assess the efficacy of various machine learning models in detecting AD using acoustic emotional data. The models were tested across three different segment lengths—1.5 s, 3.5 s, and 4.5 s—providing insights into the optimal configuration for capturing emotional nuances essential for accurate diagnosis.

We utilized two distinct types of data transformations to explore how different representations of emotional data influence model performance: categorical and statistical tabular forms. The results from these transformations are systematically detailed in two tables. Table I outlines the F1 scores and accuracy rates for models using the categorical tabular data, while Table II displays similar metrics for models employing statistical tabular data.

A. Categorical Tabular Models

In the categorical tabular model approach, emotional states captured directly from segments exhibit variance in F1 scores and accuracy across models and segment lengths. The Soft Voting Ensemble stands out, achieving the highest F1 score of 0.9231 for 4.5 s segments, as shown in Table I and visually represented in Figure 3. Similarly, the accuracy peaked at 0.8738 for the same model and segment length, indicating that longer segments may provide a more comprehensive representation of emotional states, which could be beneficial for the model's predictive performance.

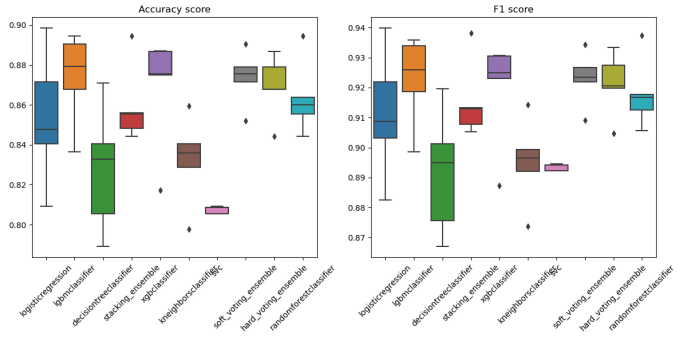

Fig. 3. Boxplot of F1 and accuracy scores for categorical tabular models with a segment length of 4.5 seconds.

TABLE I
Accuracy and F1 Score for Categorical Tabular Models

Model	1.5 s	3.5 s	4.5 s
F1 Score			
Decision Tree	0.8726	0.8728	0.8917
Random Forest	0.9066	0.9156	0.9180
k-Nearest Neighbors	0.7156	0.7261	0.8952
Logistic Regression	0.8678	0.8638	0.9112
Support Vector Classifier	0.8935	0.8935	0.8935
LGBM Classifier	0.9132	0.9100	0.9226
XGBoost Classifier	0.9144	0.9146	0.9193
Soft Voting Ensemble	0.9157	0.9138	**0.9231**
Hard Voting Ensemble	0.9150	0.9154	0.9212
Stacking Ensemble	0.9164	0.9139	0.9155
Accuracy Score			
Decision Tree	0.7966	0.7989	0.8278
Random Forest	0.8426	0.8589	0.8636
k-Nearest Neighbors	0.6126	0.6251	0.8324
Logistic Regression	0.7880	0.7748	0.8535
Support Vector Classifier	0.8075	0.8075	0.8075
LGBM Classifier	0.8574	0.8543	**0.8738**
XGBoost Classifier	0.8597	0.8613	0.8683
Soft Voting Ensemble	0.8613	0.8574	**0.8738**
Hard Voting Ensemble	0.8582	0.8597	0.8691
Stacking Ensemble	0.8613	0.8574	0.8597

B. Statistical Tabular Models

The statistical tabular model approach, which employs a detailed extraction of descriptive statistical features from the emotional data, generally delivered superior performance. The LGBM Classifier was particularly noteworthy, achieving the highest F1 score of 0.9327 for the 3.5 s segments, detailed in Table II and illustrated in Figure 4. It also attained the highest accuracy rate of 0.8893. This success emphasises the value of integrating comprehensive statistical features into the model training process, especially with segment lengths that strike a balance between capturing temporal nuances and maintaining computational efficiency.

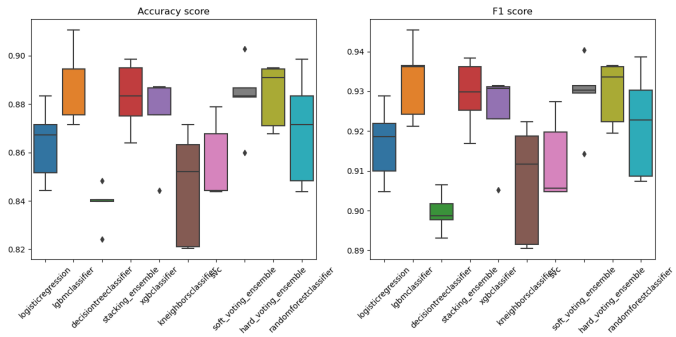

Fig. 4. Boxplot of F1 and accuracy scores for statistical tabular models with a segment length of 3.5 seconds.

Overall, both the Random Forest and various Ensemble Methods showed robust performance, confirming the effectiveness of these techniques in managing complex patterns inherent in acoustic emotion data. The findings highlight the critical balance between segment length and data repre-

TABLE II
ACCURACY AND F1 SCORE FOR STATISTICAL TABULAR MODELS

Model	1.5 s	3.5 s	4.5 s
F1 Score			
Decision Tree	0.8817	0.8996	0.8803
Random Forest	0.9115	0.9215	0.9122
k-Nearest Neighbors	0.9028	0.9070	0.9087
Logistic Regression	0.9116	0.9169	0.9174
Support Vector Classifier	0.9134	0.9125	0.9205
LGBM Classifier	0.9179	**0.9327**	0.9183
XGBoost Classifier	0.9178	0.9243	0.9220
Soft Voting Ensemble	0.9203	0.9292	0.9214
Hard Voting Ensemble	0.9158	0.9296	0.9241
Stacking Ensemble	0.9164	0.9293	0.9258
Accuracy Score			
Decision Tree	0.8090	0.8386	0.8083
Random Forest	0.8527	0.8691	0.8527
k-Nearest Neighbors	0.8355	0.8457	0.8496
Logistic Regression	0.8550	0.8636	0.8644
Support Vector Classifier	0.8566	0.8558	0.8675
LGBM Classifier	0.8652	**0.8893**	0.8660
XGBoost Classifier	0.8644	0.8761	0.8722
Soft Voting Ensemble	0.8683	0.8831	0.8699
Hard Voting Ensemble	0.8605	0.8839	0.8746
Stacking Ensemble	0.8613	0.8831	0.8769

sentation, revealing that moderate segment lengths might be most suitable for achieving high diagnostic accuracy in the context of AD detection using acoustic emotions. Our study evaluated the efficacy of multiple machine learning models across different dataset segmentations—1.5 s, 3.5 s, and 4.5 s. The results, highlighted in two primary tables, categorically present the F1 scores and accuracy rates for both categorical and statistical tabular models.

For the categorical tabular models, the F1 scores varied across different segment lengths, with the Soft Voting Ensemble achieving the highest F1 score of 0.9231 for the 4.5 s dataset. This suggests a slight advantage in performance with increased segment length for capturing more stable emotional characteristics. Similarly, the accuracy rates were the highest with the LGBM Classifier and Soft Voting Ensemble models, both reaching an accuracy of 0.8738 for the 4.5 s segments.

In the realm of statistical tabular models, the models generally performed better, reflecting the robustness of statistical features in capturing relevant patterns for AD detection. The LGBM Classifier stood out, exhibiting the highest F1 score of 0.9327 for the 3.5 s segments, and also delivering the best accuracy of 0.8893. These results indicate a strong correlation between the feature-rich statistical approach and model performance, particularly at moderate segment lengths which balance temporal resolution with emotional variability.

Overall, the Random Forest and Ensemble Methods consistently showed strong performance across both datasets and segmentations. This underscores the utility of ensemble techniques in managing the varied and complex nature of acoustic emotion data. The performance patterns observed suggest that while longer segments provide a comprehensive emotional snapshot, moderate segment lengths might offer the optimal balance for effective machine learning model training in the context of AD detection.

The analysis of feature importance within the LGBM Classifier, particularly applied to the statistical tabular model, provided pivotal insights into the factors most influential in diagnosing AD. Notably, the age of the patient and specific statistical measures of emotional expressions such as the maximum and skewness of neutral emotions, skewness of calm emotions, minimum of angry emotions, and the patient's education level were identified as having the highest importance. These features significantly contribute to the model's predictive power, underscoring the complex interplay between emotional stability, demographic factors, and cognitive health.

Conversely, certain features like the race of the patient and the variance metrics across all emotions were found to have zero importance in the model. This indicates that these features do not contribute to the differentiation between Healthy Control and AD cases within the context of our model, suggesting that they may be less relevant for the type of emotional and cognitive markers that our analysis targets.

The robust performance of the LGBM Classifier, along with the results from the ensemble methods used in the statistical tabular model configuration, highlights the substantial potential of these approaches to enhance diagnostic accuracies. The use of ensemble methods, which combine the strengths of various algorithms, has proven especially effective in our study, delivering superior results. This efficacy not only demonstrates the power of machine learning in leveraging complex datasets but also its scalability and adaptability for use in clinical settings.

V. DISCUSSION

The results of our study point to the potential of using emotion recognition from speech as a method to detect AD. Our research achieved impressive results in identifying AD through emotional cues in speech, securing the highest F1 score, although we did not surpass the accuracy score of the model developed by Khan [10]. This tells us that while our approach is on the right track, there's still scope for greater accuracy in AD detection.

Our investigation confirms that the length of speech segments we analyse plays a crucial role. In line with existing literature, we found that segments around 3.5 seconds long strike a good balance—they're not too short or too long, which is ideal for capturing the emotional nuances necessary for AD detection. This was particularly evident with the LGBM Classifier's performance in the statistical tabular models. Converting our speech data into a statistical format, consisting of a rich summary of each segment, proved to be a strong strategy. This transformation enhances the ability of machine learning models to interpret data more effectively, enabling them to detect nuanced variations in emotional expressions. This capability is essential for differentiating between healthy individuals and those affected by Alzheimer's Disease (AD).

Nevertheless, the study does have its limitations. Our reliance on a single dataset might not reflect the full variability

of speech across different populations or AD stages. Future work should, therefore, focus on testing these findings across a more diverse range of voices and settings.

The main model we used for recognising emotions in speech, while effective, showed areas for improvement in terms of loss and accuracy scores. Moving forward, combining multiple emotion recognition models could improve the reliability of our emotion analysis. This future work might also include merging these emotion features with established speech features like eGeMAPS or measures of silent pauses. Such combinations could enhance the accuracy of AD detection methods.

VI. Conclusion

This study introduces a novel methodology for AD detection using the emotional content extracted from audio recordings. By analysing the acoustic emotions expressed by patients, we propose a non-invasive approach that could potentially streamline the preliminary screening process for AD. Our research demonstrates the feasibility of using machine learning models to interpret and classify emotional nuances, which are indicative of cognitive changes associated with AD.

The most compelling findings from our investigation revealed that ensemble methods and the LGBM Classifier consistently outperformed other models, with the highest F1 score of 0.9327 and an accuracy score of 0.8893 obtained from the LGBM Classifier on the 3.5-second segments in the statistical tabular models. These results underscore the effectiveness of our proposed method, particularly when employing a moderate segment length that optimally balances the granularity of emotional detection with computational efficiency.

Our study not only highlights the potential of acoustic emotion recognition as a diagnostic tool but also sets the groundwork for future research to explore the integration of multiple emotion recognition models. By employing a combination of diverse datasets and advanced modelling techniques, subsequent studies can aim to enhance the reliability and accuracy of emotion-based AD detection.

In conclusion, the integration of acoustic emotion analysis into the diagnostic landscape for AD offers a promising new avenue for early detection and monitoring. While the current models provide a strong foundation, continuous advancements in model accuracy and robustness are necessary to fully realise the potential of this innovative approach. The pursuit of refining these techniques and validating them in broader clinical settings remains a crucial next step in the evolution of AD diagnostics.

References

[1] Z. Breijyeh and R. Karaman, "Comprehensive review on alzheimer's disease: Causes and treatment," *Molecules*, vol. 25, no. 24, p. 5789, Dec 2020.

[2] A. Nandi, N. Counts, J. Bröker, and et al., "Cost of care for alzheimer's disease and related dementias in the united states: 2016 to 2060," *npj Aging*, vol. 10, p. 13, 2024. [Online]. Available: https://doi.org/10.1038/s41514-024-00136-6

[3] N. Choices, "What is dementia," 2024. [Online]. Available: https://www.nhs.uk/conditions/dementia/about-dementia/what-is-dementia/

[4] S. H. Omar and J. Preddy, "Advantages and pitfalls in fluid biomarkers for diagnosis of alzheimer's disease," *Journal of personalized medicine*, vol. 10, no. 3, p. 63–63, Jul 2020. [Online]. Available: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7563364/

[5] Edwin, C. Kuhl, Y. Anzai, P. Desmond, R. L. Ehman, Q. Gong, G. Gold, V. Gulani, M. Hall-Craggs, T. Leiner, T. Lim, J. G. Pipe, S. Reeder, C. Reinhold, M. Smits, D. K. Sodickson, C. Tempany, H. A. Vargas, and M. Wang, "Value of mri in medicine: More than just another test?" *Journal of magnetic resonance imaging*, vol. 49, no. 7, Aug 2018. [Online]. Available: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7036752/

[6] O. Sáiz-Vázquez, P. Gracia-García, S. Ubillos-Landa, A. Puente-Martínez, S. Casado-Yusta, B. Olaya, and J. Santabárbara, "Depression as a risk factor for alzheimer's disease: A systematic review of longitudinal meta-analyses," *Journal of clinical medicine*, vol. 10, no. 9, p. 1809–1809, Apr 2021. [Online]. Available: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC8122638/

[7] C. Crump, W. Sieh, B. G. Vickrey, A. C. Edwards, J. Sundquist, and K. Sundquist, "Risk of depression in persons with alzheimer's disease: A national cohort study," *Alzheimer's & dementia. Diagnosis, assessment & disease monitoring*, vol. 16, no. 2, Apr 2024. [Online]. Available: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC11016814/

[8] S. Aramadaka, R. Mannam, R. S. Narayanan, A. Bansal, V. R. Yanamaladoddi, S. S. Sarvepalli, and S. L. Vemula, "Neuroimaging in alzheimer's disease for early diagnosis: A comprehensive review," *Curēus*, May 2023. [Online]. Available: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC10239271/

[9] A. G. Vrahatis, K. Skolariki, M. G. Krokidis, K. Lazaros, T. P. Exarchos, and P. Vlamos, "Revolutionizing the early detection of alzheimer's disease through non-invasive biomarkers: The role of artificial intelligence and deep learning," *Sensors*, vol. 23, no. 9, p. 4184–4184, Apr 2023. [Online]. Available: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC10180573/

[10] Y. Khan, B. Kaushik, M. K. I. Rahmani, and M. Ahmed, "Stacked deep dense neural network model to predict alzheimer's dementia using audio transcript data," *IEEE Access*, vol. 10, pp. 1–1, 01 2022.

[11] L. Ilias and D. Askounis, "Multimodal deep learning models for detecting dementia from speech and transcripts," *Frontiers in aging neuroscience*, vol. 14, Mar 2022. [Online]. Available: https://www.frontiersin.org/articles/10.3389/fnagi.2022.830943/full

[12] M. Liu, X. Xie, J. Xie, S. Tian, X. Du, H. Feng, and H. Zhang, "Early-onset alzheimer's disease with depression as the first symptom: a case report with literature review," *Frontiers in psychiatry*, vol. 14, Apr 2023. [Online]. Available: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC10174310/

[13] Z. Fei, E. Yang, L. Yu, X. Li, H. Zhou, and W. Zhou, "A novel deep neural network-based emotion analysis system for automatic detection of mild cognitive impairment in the elderly," *Neurocomputing*, vol. 468, p. 306–316, Jan 2022. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0925231221015186

[14] J. T. Becker, F. Boiler, O. L. Lopez, J. Saxton, and K. L. McGonigle, "The Natural History of Alzheimer's Disease: Description of Study Cohort and Accuracy of Diagnosis," *Archives of Neurology*, vol. 51, no. 6, pp. 585–594, Jun. 1994. [Online]. Available: https://doi.org/10.1001/archneur.1994.00540180063015

[15] A. Defossez, G. Synnaeve, and Y. Adi, "Real time speech enhancement in the waveform domain," in *Interspeech*, 2020.

[16] E. H. Calabrés, "wav2vec2-lg-xlsr-en-speech-emotion-recognition (revision 17cf17c)," 2024. [Online]. Available: https://huggingface.co/ehcalabres/wav2vec2-lg-xlsr-en-speech-emotion-recognition

[17] A. Conneau, A. Baevski, R. Collobert, A. Mohamed, and M. Auli, "Unsupervised cross-lingual representation learning for speech recognition," 2020.