Interspeech 2025
17-21 August 2025, Rotterdam, The Netherlands

Understanding Dementia Speech Alignment with Diffusion-Based Image Generation

Mansi [*], *Anastasios Lepipas* [*], *Dominika Woszczyk, Yiying Guan, Soteris Demetriou*

Imperial College London, UK

m.-24@imperial.ac.uk, a.lepipas20@imperial.ac.uk, d.woszczyk19@imperial.ac.uk, yiying.guan23@imperial.ac.uk, s.demetriou@imperial.ac.uk

Abstract

Text-to-image models generate highly realistic images based on natural language descriptions and millions of users use them to create and share images online. While it is expected that such models can align input text and generated image in the same latent space little has been done to understand whether this alignment is possible between pathological speech and generated images. In this work, we examine the ability of such models to align dementia-related speech information with the generated images and develop methods to explain this alignment. Surprisingly, we found that dementia detection is possible from generated images alone achieving 75% accuracy on the ADReSS dataset. We then leverage explainability methods to show which parts of the language contribute to the detection.
Index Terms: dementia, text–to–image models, privacy

1. Introduction

Problem Statement. The rapid advancement of text-to-image (T2I) diffusion models, such as Stable Diffusion [1] has significantly expanded their applications, enabling highly realistic and contextually rich image generation from textual descriptions. By simply describing the desired image in natural language, users can leverage T2I models to produce visually accurate representations. However, as these models become increasingly integrated into speech and assistive technologies, critical privacy concerns emerge—-particularly in the context of vulnerable populations, such as individuals suffering from dementia.

This study investigates an unexplored privacy risk: the potential leakage of dementia-related characteristics through the output space of T2I diffusion models. Specifically, we explore whether images generated from speech-derived textual descriptions(from the ADReSS dataset) can inadvertently encode and reveal cognitive decline markers. Since dementia affects speech patterns like lexical choice, syntactic complexity, and fluency, diffusion models trained on text representations of speech may implicitly learn and propagate these features into their outputs. In our work, we focus our analysis on *Information Units* (IU) which we define as the set of nouns and verbs which are present in the image, similar to [2]; and *discourse tokens* (pauses and filler words, e.g.: "um","uh", etc). Alarmingly, we show that diffusion models can inadvertently infer and externalize cognitive health indicators that can result in unauthorized profiling, discrimination, or stigmatization with generated images potentially carrying sensitive neurocognitive information and posing ethical, security, and privacy concerns.

Prior works. The field of dementia classification from speech [3] has grown significantly, particularly since the 2020 *ADReSS* challenge. These models can be categorized into three main types: those using exclusively acoustic data [4, 5], those only using speech transcriptions [5, 6, 7, 8], and hybrid models combining both [9, 10, 11, 12]. More related to our dementia inference classifiers are models trained on transcriptions. Among those, BERT–based models have been shown to be highly effective yielding 81%–83% accuracy [7]. More similar to our approach are the models by Zhu et al. [13] which combine information from the original "Cookie Theft Picture" with transcriptions to achieve between 80.63% and 89.6% accuracy on the *ADReSS* dataset which matches the SOTA result (89.6%) from Yuan et al's ERNIE-based text–audio approach [9]. The main scope of our work is not to improve the dementia inference accuracy but to understand whether pathological speech information leakage is possible in image generation.

In addition, previous works have focused on analysing the input space features which lead to dementia classification. Model agnostic explainability techniques like SHAP [14] and LIME [15] have been used to identify linguistic patterns [16] as well as key attributes that influence classification outcomes [17, 18]. Other works [19, 20] leverage techniques like GradCAM and SHAP on multimodal data to identify the key aspects responsible for distinguishing dementia. To the best of our knowledge, there has been no work towards analysing the relationship of these characteristics across modalities in generative models.

Our Approach. To investigate the potential leakage of dementia-related information through T2I diffusion models, we use a three-stage analysis framework which consists of speech-to-text conversion, text-to-image generation, and image-based inference analysis. This approach allows us to study the relationship between sensitive information from speech and generated images, and evaluate the extent to which diffusion models encode and propagate cognitive decline markers. We leverage the transcription samples from the ADReSS dataset which resemble natural image descriptions as input prompts to a diffusion model to instruct it to generate images.

Below we summarize the **main contributions** of this work:

• **Novel Application Domain.** To the best of our knowledge, this is the first study to explore dementia-related speech leakage in T2I diffusion models. By bridging pathological speech analysis with generative AI security, we highlight a previously overlooked privacy dimension for a vulnerable population.

• **New Dementia Inference Models.** We demonstrate that images generated from dementia-affected speech descriptions can encode implicit markers of neurocognitive decline posing a risk to individuals' privacy.

• **New Understanding.** We demonstrate that dementia can be

[*]The first two authors contributed equally to this work.

10.21437/Interspeech.2025-871

leaked through T2I model outputs, with background details being the most discriminating features. Additionally, discourse tokens impact detection but are not the sole cause of leakage.

2. Threat Model

Our threat model focuses on T2I diffusion models. Such models leverage natural language descriptions to guide the generation process. Therefore, we hypothesize that they might inadvertently encode linguistic markers of cognitive decline into their visual outputs. This raises significant privacy concerns, as generated images may serve as unintended carriers of sensitive neurocognitive information, allowing adversaries to infer a speaker's cognitive status without direct access to their speech or text. The ability of T2I models to propagate such information underscores the urgent need for privacy-preserving mechanisms in generative AI applications for healthcare. To examine our hypothesis we focus on dementia-related speech transcriptions. Note that real systems already exist which take speech as input, then convert it to text before passing it to the T2I model [21].

We consider an adversary $\mathcal{A}$ with black–box access to T2I models, as it reflects a more realistic attack scenario compared to a white–box setting, which would require access to the model's internal components. We define two adversary models with different levels of access to resources, text–based ($\mathcal{A}_i$) and image–based ($\mathcal{A}_o$) only. Considering the related work, we do not aim to improve the classification performance of the $\mathcal{A}_i$ model which uses the natural descriptions for inference. Instead, we focus on a) exploring the feasibility of the $\mathcal{A}_o$ model which leverages a different modality to detect Dementia, and b) better understanding which information units from the input space are leaked in the output space contributing to dementia classification. Using these insights, we hope that our work inspires further work on protecting affected populations.

3. Analysis Methods

3.1. Overview

Our analysis comprises three stages: extracting speech-to-text transcriptions; generating images based on those transcriptions; analyzing the relationship between the ability to detect dementia from T2I language inputs and generated images.

Transcriptions. We use ADReSS [3], a subset of the DementiaBank dataset [22]. 156 participants are provided with the same 'Cookie Theft Picture' and are asked to describe it orally. The descriptions are manually transcribed eliminating the effect of ASR errors. The ADReSS dataset contains healthy-control (CC) and dementia–labelled (AD) descriptions, with 54 train and 24 test samples for each class, for a total of 156. In our evaluations, 20% of the train set is set aside as the validation set. All the evaluations were done on the provided test set.

Image Generation. We leverage Stable Diffusion v2.1 an open source T2I model to generate images. Specifically, each sample (description) in the ADReSS dataset is used as an input to the model. Stable Diffusion v2.1 clips the input to 77 tokens which means parts of the descriptions are not considered in the generation. However, this does not affect our analysis which focuses on the difference between AD and CC using the same number of available tokens as the frequency distribution and total count of nouns within the first 77 tokens are nearly identical for both groups. We therefore conclude that restricting the study to the first 77 tokens does not introduce any significant bias.

Analysis. The last part is the core of the framework. For the analysis we (a) develop binary dementia classification models based on text, and based on images; (b) use explainable AI techniques to study what the classifiers learn; (c) identify and introduce new metrics to evaluate the relationship between the input descriptions and the generated images. We describe this in more detail below.

3.2. Dementia Classification.

For our classification task, we use all the classic machine learning algorithms integrated in the scikit–learn library [23]. Due to space limitations, we report only our best classifier results in the Evaluation section. Notably, our classifiers are trained based on embeddings which encode joint representations of text and images in a latent space. This is paramount for our analysis to better identify what characteristics of text are present in the generated images. CLIP [24] and ViT [25] embeddings are suitable for this purpose. For our classifiers, we use CLIP embeddings.

3.3. XAI Methods

Explainable AI (XAI) methods provide insights into the decision-making processes of deep learning models by highlighting the most influential features that drive predictions. For instance, a well–established approach is GradCAM [26], which generates heatmaps to visualize the regions of an input image that contribute most to a model's prediction. Thus, we leverage XAI techniques to understand the contribution of relevant features in the input and output spaces. In this work, we focus our analysis on *Information Units (IU)*. The set of IUs we used is constructed by performing POS tagging on the entire dataset and then collecting the unique and meaningful nouns and verbs. More specifically, for the input space, we compute the SHAP values for the IUs. For the output space, we use GradCAM on top of SVM to identify the attention score by the classifier over the image. Then, we use Grounded-SAM [27] to get the mask for the IUs. We compute the contribution score for an IU as the sum of the attention for the mask of the IU divided by the area of the mask. We divide the sum of attention score by the area of the mask so as to normalise any bias towards the contribution score due to the area occupied by the IU.

3.4. Metrics

We aim to take a step towards understanding dementia-related information leakage from speech transcriptions to generated images. To this end we evaluate our text–image embedding dementia classifiers in terms of *privacy* and *utility*

Privacy. We measure dementia leakage via dementia classification accuracy. In other words we assume that if dementia classification is possible in the input or the output space then dementia-related linguistic markers are learned by the classifiers and can be used to infer one's condition.

Utility. CLIPScore [28] and TIFA [29] are used for measuring the semantic similarity between the input prompts and the output images. For comparing the quality of the images between the two groups, we leverage the Inception Score (IS), FID, and LPIPS. We consider the original cookie theft image as the ground truth image. Furthermore, we introduce two new metrics, namely Information Units Propagation Score (IPS) [30] (in Equation 1) and Extraneous Content Score (ECS) (in Equation 2) which specifically evaluate the extent to which the desired content in terms of IUs is propagated into the output space and how much of the unintended content appears in the output

space respectively. More formally:

$$\text{IPS}_{\text{avg}} = \frac{1}{N}\sum_{i=1}^{N}\frac{\sum_{j=1}^{M}\mathbb{1}(\text{IU}_j \in \text{prompt}_i)\cdot\mathbb{1}(\text{C}(\text{IU}_j)\neq 0)}{\sum_{j=1}^{M}\mathbb{1}(\text{IU}_j \in \text{prompt}_i)} \tag{1}$$

$$\text{ECS}_{\text{avg}} = \frac{1}{N}\sum_{i=1}^{N}\frac{\sum_{j=1}^{M}\mathbb{1}(\text{IU}_j \notin \text{prompt}_i)\cdot\mathbb{1}(\text{C}(\text{IU}_j)\neq 0)}{M-\sum_{j=1}^{M}\mathbb{1}(\text{IU}_j \in \text{prompt}_i)} \tag{2}$$

where M is total number of IUs, N is the size of the test set and C is the function which computes the contribution score. Lastly, we note that the contribution scores computed for the IUs are used for evaluation.

4. Evaluation

4.1. Research Questions.

Our evaluation aims to answer the following overall research questions: **RQ1:** Can we detect Dementia from generated images? **RQ2:** What regions of an image are responsible for the leakage? **RQ3:** What features are leaked during the transformation of a prompt to an image?

4.2. Dementia Leakage

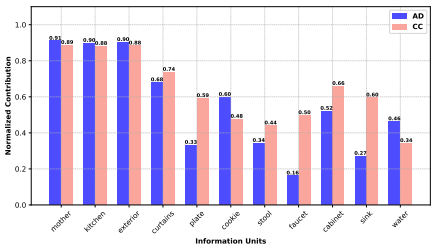
To answer RQ1, for the Input model ($\mathcal{A}_i$) we extract the CLIP embedding from text and we train for 100 epochs, using the Binary Cross Entropy Loss, a Neural Network with 3 hidden layers and batch size 16. We apply a 70 : 30 training/test split on the dataset and use 20% of the training set for validation to avoid overfitting to the test set. Our $\mathcal{A}_i$ model achieves 83.33% accuracy. For the Output models, we use the original descriptions to generate images with Stable Diffusion v2.1. As mentioned in Section 3, we leverage clip embeddings and we use the same dataset splits as in $\mathcal{A}_i$. Our $\mathcal{A}_o$ model achieves 75% using SVM (based on the *One–Vs–One* approach).

Observations. The results suggest that T2I models can inadvertently reveal sensitive medical information through their outputs. More importantly and in contrast with prior work [13], our $\mathcal{A}_o$ approach does not depend on audio information. It merely requires the image generated by the T2I model which is known to the model providers by default. Such images are also publicly exposed when a user shares them online.

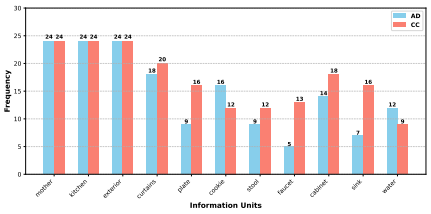
4.3. Salient Regions

To understand how generated images contribute to dementia classification, we analyze the specific image regions or patches that are most influential in the model's decision–making (RQ2). Our analysis is conducted in two settings: (1) within each group separately (AD vs. CC), and (2) across both groups collectively to identify shared and distinct classification patterns. This dual approach allows us to highlight common artifacts or regions that are consistently important for classification, as well as group-specific differences in visual features. Inspired by Zhu et al. [31] we also try to understand the contribution of different information units in the image towards classification. However, we use explainability techniques and fine segmentation models, like Grounded SAM, to extract the fine contribution of all the units, rather than relying on selective search, which identifies regions based on another model's assumptions.

Figure 1 summarizes our results. For space limitation, we include only a few information units which can highlight the trend (in Figure 1a). In detail, Figure 1a shows that the contribution of IUs like 'mother', 'kitchen', 'exterior' are very high for both classes. More importantly. we see that finer-grained details like 'faucet', 'stool', 'plate', and 'sink' are not very important features for the AD group as in CC. This same trend is followed by both the groups. Furthermore, Figure 1b indicates that the contributions of the features across the two groups directly correspond to the frequency distribution of these information units in the images generated for each group. This suggest that the contribution scores are mostly driven by the representation of specific information units in the output space for both groups.



(a) *Comparison of AD and CC groups in the output space.*



(b) *Frequency of the information units in the output space for both groups.*

Figure 1: *Analysis of features responsible for classification in the output space.*

4.4. Feature Propagation

To answer RQ3, we focus on analysing how the contribution of IUs to classification changes from the input space to the output space. Strikingly, Figure 2a reveals that the most important IUs in the output space were among the least significant in the input space. For instance, the IU 'kitchen' which had minimal importance for classification in the input space, emerges as the most critical in the output space. Similarly, units like 'faucet' and 'cabinet' which were entirely absent in the input space, play a key role in classification within the output space. Beyond that, there is a general increase in the contribution score for all the information units.

Figure 2b shows that, similar to the output space, there is a general correspondence between the frequency distribution of information units and their contribution scores in the input space. However, the relationship between the frequency of information tokens in the input space and their actual representation in the output space is highly inconsistent. For example, tokens such as 'kitchen', 'curtains', 'faucet', 'cabinet' are scarcely present in the input space but appear prominently in the output space, likely due to the T2I model's tendency to enhance prompts with additional details. In contrast, information units like 'stool', 'sink', 'water'-—despite being prominent in the input space—-are largely absent in the output space. This

suggests that T2I models lose information and introduce noise during the generation process, disrupting the accurate propagation of features into the output space.

We hypothesize that the deviations from the intended generation goal may be influenced by the presence of discourse tokens in the input space. To study their impact, we remove discourse tokens and compute the contribution scores and classification accuracy for the two groups. While the input space accuracy experiences only a slight decline, from 83.33% to 81.25%, the output space accuracy drops significantly, from 75% to 62.13%. This suggests that *discourse tokens played a crucial role in differentiating the two classes in the output space*. Additionally, we observe that removing the discourse tokens improves the classifier's attention towards minor details in the output bringing the contribution scores for AD and CC closer for most information units (see Figure 2c). This reduction in differentiation results in lower classification accuracy in the output space. For example, the contribution score gap of 'faucet' between the groups decreased from 0.34 to 0.02.

To better understand how the discourse tokens influence the output space, we analyse the fidelity of T2I model outputs across the two groups. Specifically, we examine which elements of the prompts are consistently carried over into the generated images using CLIP, TIFA, and IPS score. The results in Table 1 suggest that removing the discourse tokens improves semantic similarity for both groups. However, interestingly, despite the CC group's prompt being more clear and concise, the AD group exhibits higher semantic similarity in the generated images. This discrepancy can be attributed to significant neglect in T2I diffusion models, where the presence of multiple entities in a prompt increases the likelihood of omitting key elements during generation. In our case, CC prompts contain nearly twice as many entities as AD prompts on average, leading to a greater omission of topics and, consequently, lower semantic similarity.

Group	IU Present	CLIP		TIFA		IPS	
		Org	ND	Org	ND	Org	ND
AD	7.20	0.15	0.16	0.28	0.28	0.75	0.77
CC	12.62	0.15	0.15	0.23	0.20	0.76	0.72

Table 1: *Comparison of IU present, CLIP, TIFA, and IU Propagation Score(IPS) scores for AD and CC groups.* ND *refers to 'No Discourse', and* Org *refers to 'Original'.*

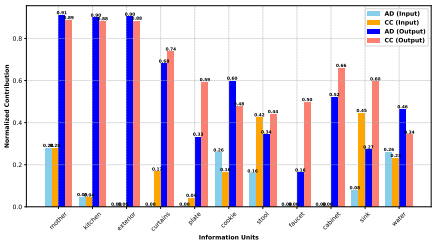
To further analyze the extent of creative freedom exhibited by the T2I model in generating additional content, we compute the ECS score. Table 2 shows that ECS score decreases from 0.67 to 0.66 for the AD group and from 0.75 to 0.74 for the CC group, indicating that the removal of discourse tokens from the prompt limited the T2I model's ability to generate extraneous content[1].

Additionally, we evaluate the impact of discourse tokens on generation quality by computing IS, FID, and LPIPS scores for both groups, as summarized in Table 2. An increase in these scores indicate that removing discourse tokens enhances generation quality. In conclusion, while discourse tokens play a significant role in classification within the output space, they are not the sole contributing factor. As shown in Figure 2c, even though their removal reduces the gap between the contribution scores of information units across the two groups, a noticeable difference remains. This highlights the presence of additional distinguishing features influencing classification.

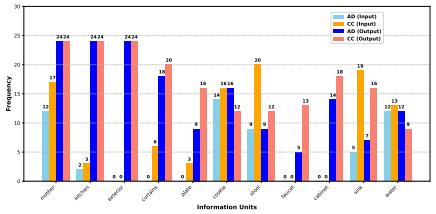
[1]Note that ADReSS dataset has a limited number of data samples, and might not be indicative for large ones.

Group	IS		FID		LPIPS		ECS	
	Org	ND	Org	ND	Org	ND	Org	ND
AD	3.16	3.62	425.29	433.06	0.71	0.72	0.67	0.66
CC	3.08	2.80	435.96	411.14	0.72	0.73	0.75	0.74

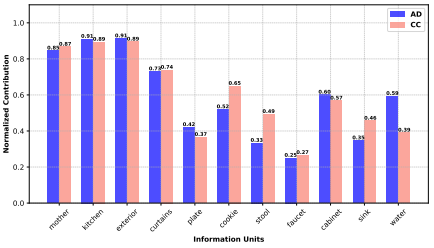
Table 2: *Comparison of IS, FID, LPIPS, and ECS scores for AD and CC groups.* Org *refers to 'Original',* ND *refers to 'No Discourse'.*



(a) *Comparison of overall contribution of both AD and CC groups in the input v/s output space.*



(b) *Comparison of frequency of information tokens present in the test set for AD and CC groups in the input v/s output space.*



(c) *Contribution score of IUs in the output space after removal of discourse tokens.*

Figure 2: *Comparison of frequency and contribution of IUs across AD and CC in the input as well as output space along with evaluation after removing the discourse tokens.*

5. Conclusion

Our work highlights a critical and previously overlooked privacy risk associated with T2I models: the potential for generated images to inadvertently reveal sensitive neurocognitive information. Our findings demonstrate that visual artifacts within these images can serve as unintended indicators of dementia, aligning with linguistic patterns present in speech descriptions. Through the use of explainability techniques, we identified specific regions within the generated images that contribute to classification, reinforcing concerns about implicit leakage of medical conditions through generative AI outputs. These results underscore the ethical and privacy challenges posed by diffusion models. The ability to infer cognitive health status from generated images raises the risk of unauthorized profiling or discrimination, highlighting the urgent need for mitigation strategies. While adversaries could exploit this vulnerability, various techniques exist to obfuscate sensitive linguistic cues before they are transformed into images [32]. Future research should focus on developing robust defenses that ensure the responsible deployment of T2I models, preserving both privacy and inclusivity.

6. References

[1] R. Rombach, A. Blattmann, D. Lorenz, P. Esser, and B. Ommer, "High-resolution image synthesis with latent diffusion models," 06 2022, pp. 10 674–10 685.

[2] Y. Zhu, N. Lin, X. Liang, J. A. Batsis, R. M. Roth, and B. MacWhinney, "Evaluating picture description speech for dementia detection using image-text alignment," *arXiv preprint arXiv:2308.07933*, 2023.

[3] S. Luz, F. Haider, S. de la Fuente, D. Fromm, and B. MacWhinney, "Alzheimer's dementia recognition through spontaneous speech: The adress challenge," *arXiv preprint arXiv:2004.06833*, 2020.

[4] K. Chlasta and K. Wołk, "Towards computer-based automated screening of dementia through spontaneous speech," *Frontiers in Psychology*, vol. 11, p. 623237, 2021.

[5] P. Mahajan and V. Baths, "Acoustic and language based deep learning approaches for alzheimer's dementia detection from spontaneous speech. front aging neurosci. 2021; 13," 2021.

[6] T. Millington and S. Luz, "Analysis and classification of word co-occurrence networks from alzheimer's patients and controls," *Frontiers in Computer Science*, vol. 3, p. 649508, 2021.

[7] A. Balagopalan, B. Eyre, F. Rudzicz, and J. Novikova, "To bert or not to bert: comparing speech and language-based approaches for alzheimer's disease detection," *arXiv preprint arXiv:2008.01551*, 2020.

[8] Y. Guo, C. Li, C. Roan, S. Pakhomov, and T. Cohen, "Crossing the "cookie theft" corpus chasm: applying what bert learns from outside data to the adress challenge dementia detection task," *Frontiers in Computer Science*, vol. 3, p. 642517, 2021.

[9] J. Yuan, Y. Bian, X. Cai, J. Huang, Z. Ye, and K. Church, "Disfluencies and fine-tuning pre-trained language models for detection of alzheimer's disease." in *Interspeech*, vol. 2020, 2020, pp. 2162–6.

[10] Y. Zhu, X. Liang, J. A. Batsis, and R. M. Roth, "Domain-aware intermediate pretraining for dementia detection with limited data," in *Interspeech*, vol. 2022. NIH Public Access, 2022, p. 2183.

[11] R. Haulcy and J. Glass, "Classifying alzheimer's disease using audio and text-based representations of speech," *Frontiers in Psychology*, vol. 11, p. 624137, 2021.

[12] M. Martinc, F. Haider, S. Pollak, and S. Luz, "Temporal integration of text transcripts and acoustic features for alzheimer's diagnosis based on spontaneous speech," *Frontiers in Aging Neuroscience*, vol. 13, p. 642647, 2021.

[13] Y. Zhu, X. Liang, J. A. Batsis, and R. M. Roth, "Exploring deep transfer learning techniques for alzheimer's dementia detection," *Frontiers in computer science*, vol. 3, p. 624683, 2021.

[14] S. Lundberg, "A unified approach to interpreting model predictions," *arXiv preprint arXiv:1705.07874*, 2017.

[15] M. T. Ribeiro, S. Singh, and C. Guestrin, "" why should i trust you?" explaining the predictions of any classifier," in *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*, 2016, pp. 1135–1144.

[16] L. Ilias and D. Askounis, "Explainable identification of dementia from transcripts using transformer networks," *IEEE Journal of Biomedical and Health Informatics*, vol. 26, no. 8, pp. 4153–4164, 2022.

[17] V. Viswan, N. Shaffi, M. Mahmud, K. Subramanian, and F. Hajamohideen, "Explainable artificial intelligence in alzheimer's disease classification: A systematic review," *Cognitive Computation*, vol. 16, no. 1, pp. 1–44, 2024. [Online]. Available: https://doi.org/10.1007/s12559-023-10192-x

[18] A. S. Alatrany, W. Khan, A. Hussain, H. Kolivand, and D. Al-Jumeily, "An explainable machine learning approach for alzheimer's disease classification," *Scientific Reports*, vol. 14, no. 1, p. 2637, 2024. [Online]. Available: https://doi.org/10.1038/s41598-024-51985-w

[19] S. Jahan, K. A. Taher, M. S. Kaiser, M. Mahmud, M. S. Rahman, A. S. M. S. Hosen, and I.-H. Ra, "Explainable ai-based alzheimer's prediction and management using multimodal data," *PloS One*, vol. 18, no. 11, p. e0294253, 2023. [Online]. Available: https://doi.org/10.1371/journal.pone.0294253

[20] K. J. Junior, K. S. Carole, T. P. Theodore Armand, H.-C. Kim, and T. A. D. N. Initiative, "Alzheimer's multiclassification using explainable ai techniques," *Applied Sciences*, vol. 14, no. 18, 2024. [Online]. Available: https://www.mdpi.com/2076-3417/14/18/8287

[21] Ashutosh-AIBOT, "Voice and text to image," 2025, accessed: 2025-02-20. [Online]. Available: https://github.com/Ashutosh-AIBOT/Voice-and-Text-to-Image-

[22] 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, 1994.

[23] P. Fabian, "Scikit-learn: Machine learning in python," *Journal of machine learning research 12*, p. 2825, 2011.

[24] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark *et al.*, "Learning transferable visual models from natural language supervision," in *International conference on machine learning*. PMLR, 2021, pp. 8748–8763.

[25] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly *et al.*, "An image is worth 16x16 words: Transformers for image recognition at scale," *arXiv preprint arXiv:2010.11929*, 2020.

[26] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, "Grad-cam: Visual explanations from deep networks via gradient-based localization," in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 618–626.

[27] T. Ren, S. Liu, A. Zeng, J. Lin, K. Li, H. Cao, J. Chen, X. Huang, Y. Chen, F. Yan *et al.*, "Grounded sam: Assembling open-world models for diverse visual tasks," *arXiv preprint arXiv:2401.14159*, 2024.

[28] A. Ramesh, M. Pavlov, G. Goh, S. Gray, C. Voss, A. Radford, M. Chen, and I. Sutskever, "Zero-shot text-to-image generation," in *Proceedings of the 38th International Conference on Machine Learning*, ser. Proceedings of Machine Learning Research, M. Meila and T. Zhang, Eds., vol. 139. PMLR, 18–24 Jul 2021, pp. 8821–8831. [Online]. Available: https://proceedings.mlr.press/v139/ramesh21a.html

[29] Y. Hu, B. Liu, J. Kasai, Y. Wang, M. Ostendorf, R. Krishna, and N. A. Smith, " TIFA: Accurate and Interpretable Text-to-Image Faithfulness Evaluation with Question Answering ," in *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*. Los Alamitos, CA, USA: IEEE Computer Society, Oct. 2023, pp. 20 349–20 360. [Online]. Available: https://doi.ieeecomputersociety.org/10.1109/ICCV51070.2023.01866

[30] L. Kumar, R. Rajaan, N. Choudhary, and A. Sharma, "A comprehensive review of image super-resolution metrics: Classical and advanced approaches," *Acta IMEKO*, vol. 12, no. 2, pp. 1–9, 2023. [Online]. Available: https://acta.imeko.org/index.php/acta-imeko/article/view/1679

[31] Y. Zhu, N. Lin, X. Liang, J. A. Batsis, R. M. Roth, and B. MacWhinney, "Evaluating picture description speech for dementia detection using image-text alignment," *arXiv preprint arXiv:2308.07933*, 2023.

[32] D. Woszczyk and S. Demetriou, "Didots: Knowledge distillation from large-language-models for dementia obfuscation in transcribed speech," *arXiv preprint arXiv:2410.04188*, 2024.