Open Access Library Journal
2025, Volume 12, e14599
ISSN Online: 2333-9721
ISSN Print: 2333-9705

Early Alzheimer's Disease Detection from Short Speech Samples Using Lightweight, Interpretable Linguistic Markers

Rocco de Filippis[1*], Abdullah Al Foysal[2]

[1]Department of Neuroscience, Institute of Psychopathology, Rome, Italy
[2]Department of Computer Engineering (AI), University of Genova, Genova, Italy
Email: *roccodefilippis@istitutodipsicopatologia.it, niloyhasanfoysal440@gmail.com

How to cite this paper: de Filippis, R. and Al Foysal, A. (2025) Early Alzheimer's Disease Detection from Short Speech Samples Using Lightweight, Interpretable Linguistic Markers. *Open Access Library Journal*, **12**: e14599.
https://doi.org/10.4236/oalib.1114599

Received: November 12, 2025
Accepted: December 22, 2025
Published: December 25, 2025

Abstract

Early detection of Alzheimer's disease (AD) is critical for intervention and monitoring. Spontaneous speech is a rich behavioural signal of cognitive decline, yet many machine-learning pipelines rely on heavy neural models that are difficult to interpret and deploy clinically. In this study, we present a lightweight and transparent classification pipeline using short narrative speech samples. All analyses in this study were conducted on synthetically generated speech transcripts, designed to emulate linguistic patterns reported in early Alzheimer's disease. Features include lexical complexity, disfluency rates, pronoun usage, readability, idea density, and common function-word statistics, modelled using a regularized linear classifier. On the provided validation split, performance is near-perfect: ROC AUC = 1.000, Average Precision = 1.000, and 100% accuracy at a 0.5 probability threshold. The ROC curve and Precision-Recall curve show a classifier that cleanly separates AD from controls, while the confusion matrix confirms zero false positives and zero false negatives (55 vs. 55 per class). The calibration curve indicates that predicted probabilities remain close to observed frequencies, and t-SNE visualization shows clear cluster separation between AD and control participants in the fused feature space. Interpretability analyses reveal consistent clinical patterns pauses per sentence, fillers per sentence, pronoun ratio, and reduced readability are the strongest predictors of early AD, while longer sentences, higher idea density, and higher content-word ratio are characteristic of healthy controls. Permutation importance confirms that single keywords contribute minimally, suggesting the model relies on broader linguistic behaviour rather than dataset artifacts. The distribution of coefficients shows a sparse pattern with a few strong drivers and many near-zero weights ideal for a clinically interpretable system. Given the unusually high performance, these results are interpreted

cautiously; the study explicitly analyses potential leakage channels and outlines rigorous validation procedures to confirm genuine signal. We provide a set of rigorous leakage checks and outline an external validation plan. With proper safeguards, this low-cost pipeline could support clinical screening and longitudinal monitoring of cognitive decline.

Subject Areas

Artificial Intelligence, Psychiatry & Psychology

Keywords

Alzheimer's Disease, Speech-Based Detection, Linguistic Biomarkers, Machine Learning, Explainable AI, Cognitive Decline Monitoring

1. Introduction

Alzheimer's disease (AD) is a progressive neurodegenerative disorder characterized by impairments in episodic memory, semantic processing, attention, and executive control [1]-[5]. Although clinical diagnosis typically relies on neuropsychological testing, neuroimaging, or cerebrospinal biomarkers, these approaches are costly, invasive, and often detect impairment only after substantial neural damage has occurred [6]-[10]. Detecting cognitive decline at an earlier stage particularly during the transition from healthy aging to mild cognitive impairment and early AD remains a critical challenge for effective intervention and monitoring [11]-[16].

Spontaneous speech has emerged as a promising non-invasive biomarker, reflecting distributed cognitive processes such as lexical retrieval, semantic organization, discourse planning, and working memory [17]-[20]. Prior research demonstrates that early-stage AD is associated with frequent lexical pauses, increased fillers, reduced informational density, pronoun overuse, and shorter, syntactically simpler utterances [21]-[27]. Importantly, these changes often manifest months or years before measurable decline in standard neuropsychological assessments. However, despite this potential, many computational approaches rely on deep neural architectures that are difficult to interpret, computationally expensive, and unsuitable for low-resource clinical settings [28]-[31].

To support real-world adoption, clinicians require models that are transparent, reproducible, and linguistically meaningful [32]-[36]. Therefore, rather than optimizing solely for black-box predictive accuracy, it is essential to develop systems that expose what linguistic behaviours differentiate pathological speech from healthy aging [37]-[40]. In this work, we investigate whether short picture-description recordings and their transcripts can accurately discriminate early AD from cognitively normal older adults using lightweight, interpretable linguistic features and a simple linear classifier. The feature set prioritizes clinically intuitive

constructs such as disfluencies, pronoun usage, sentence complexity, readability, and idea density allowing direct interpretability of the model's decisions. Our goal is to evaluate whether these measurable linguistic cues provide reliable diagnostic signal and to determine the extent to which a transparent model can approach state-of-the-art performance without sacrificing interpretability.

2. Methodology

2.1. Data and Study Design

The study utilizes a corpus of spontaneous picture-description narratives collected from older adults clinically categorized as either Early Alzheimer's Disease (AD) or Cognitively Normal Controls. Speech samples consist of short, unconstrained verbal descriptions elicited from standardized visual prompts, a setting known to elicit rich lexical and syntactic behaviour while minimizing interviewer-induced bias. For the present analysis, the validation set contains 110 independent speech samples, stratified evenly across diagnostic groups (55 Early AD, 55 Control), as reflected in the confusion matrix. The training set contained N = 440 synthetic speech samples (220 Early AD, 220 Controls), yielding a total dataset size of N_total = 440 + 110 = 550. No samples were shared between training and validation to ensure strict subject-disjoint evaluation. The dataset used in this study consists entirely of synthetically generated speech transcripts, created to emulate picture-description narratives typically used in early Alzheimer's disease assessment (e.g., Cookie Theft-style prompts). No real audio recordings or human participants were involved. Synthetic narratives were produced using large language models configured to simulate linguistic patterns characteristic of early Alzheimer's disease and cognitively normal aging, based on patterns reported in prior literature. This synthetic design ensures full reproducibility, avoids privacy or ethical concerns, and provides controlled variation in lexical, syntactic, and discourse-level behaviours.

Each narrative is treated as a standalone observational unit, and all linguistic features are extracted at the narrative level. To avoid inadvertent data leakage, no aggregation across sessions or across a participant's multiple recordings is performed [41]-[46]. When multiple samples originated from the same individual, they were retained entirely within a single partition (training or validation), ensuring strict subject-disjoint evaluation [47]-[49]. Transcripts were either manually produced or generated through a single uniform automatic speech recognition pipeline; applying the same transcription protocol across classes prevents systematic acoustic or formatting artifacts from confounding diagnosis. Preprocessing steps included lower-casing, removal of non-speech annotations, normalization of punctuation, expansion of contractions, rule-based sentence segmentation, and tokenization. Tokenization and sentence segmentation were implemented using the spaCy v3.6 library (en_core_web_sm model), with supplemental rule-based cleaning performed via NLTK and regex-based preprocessing. All scripts were implemented in Python 3.10. Samples with insufficient lexical content

(<5 content-bearing tokens) or missing diagnostic labels were excluded. No demographic variables (age, sex, education, first language) were incorporated into the feature set, eliminating shortcut learning through population differences.

2.2. Linguistic Feature Extraction

The analytical objective was not to construct a black-box classifier, but to identify clinically interpretable linguistic behaviours distinguishing early AD from healthy aging. Accordingly, we extracted features grounded in psycholinguistic and neurolinguistic literature, grouped into six categories below in Table 1:

Table 1. Linguistic feature categories, representative measures, and their neurocognitive interpretation.

Category	Representative Features	Neurocognitive Interpretation
Disfluencies	pauses per sentence; fillers (um, uh, er); repetition rate	impaired lexical access and disrupted planning
Lexical Selection	pronoun ratio; content-word ratio	semantic degradation; reduced specificity
Syntactic Complexity	mean sentence length; clause density (when parsable)	impaired working-memory load and sentence planning
Readability	Flesch Reading Ease	fragmentation and syntactic breakdown in early AD
Idea Density	propositions per 10 words	reduced informational richness and conceptual structure
High-coverage Function Tokens	frequency of "a, the, it, then"	stylistic shifts and content impoverishment

All count-based features were normalized per-sentence or per-token to eliminate length confounds. Outliers were Winsorized at the 1st and 99th percentiles. Missing syllable counts and other lexical attributes were imputed with training-set medians. Every feature was standardized using training-set means and variances, and the identical transformation was applied to validation samples. This feature architecture ensures interpretability: each coefficient corresponds to a linguistically meaningful behaviour, enabling transparent clinical explanation. These relationships later manifest in the coefficient plots and permutation-based robustness analysis.

2.3. Classification Model

A logistic regression model with L2 regularization was employed. This classifier was selected deliberately: it produces stable, calibrated probability estimates, minimizes overfitting in low-dimensional settings, and yields a coefficient vector interpretable as log-odds shifts in diagnostic direction [50]-[52]. Logistic regression was selected over other interpretable models such as linear Support Vector Machines (SVMs) or decision trees for several reasons. Linear SVMs optimize margin but do not produce calibrated probabilities needed for clinical decision support, and decision trees are prone to instability and overfitting in low-sample, high-noise linguistic settings. In contrast, logistic regression yields smoothly varying,

directly interpretable coefficients and naturally produces well-calibrated probability estimates essential for downstream risk assessment. Hyperparameters were tuned via inner five-fold cross-validation on the training set using negative log-likelihood as the objective function [53]-[58]. Class-weighting was evaluated, but because the validation data were perfectly balanced, the final model employed unweighted classes. Optimization was performed with the lbfgs solver, and convergence was reached reliably across folds.

2.4. Evaluation

Performance was assessed along four methodological dimensions:

1) **Discriminative ability**: Receiver Operating Characteristic (ROC) and Precision Recall curves were generated, yielding AUC = 1.000 and AP = 1.000 (**Figure 1**, **Figure 2**). These threshold-free metrics quantify separability independent of a decision boundary.

2) **Threshold-level classification**: Applying a fixed 0.5 probability threshold results in perfect classification for both classes (**Figure 3**), with 100% sensitivity and 100% specificity an outcome requiring further scrutiny, addressed in Discussion.

3) **Probability calibration**: A reliability diagram (**Figure 4**) compares predicted probabilities with empirical outcome frequencies. The calibration curve lies close to the identity line, indicating that the model's confidence estimates are well-behaved, not overconfident.

4) **Representational geometry**: To examine how linguistic features structure the sample space, we applied t-SNE projection (**Figure 5**), revealing two compacts, clearly separated clusters. This separation visually reinforces that the extracted linguistic features encode distinct behavioural signatures of AD and Control speech.

2.5. Interpretability and Robustness Diagnostics

Model interpretability was central to the methodological design. Coefficient magnitudes and signs (**Figure 6** and **Figure 7**) directly quantify how each feature shifts diagnostic likelihood. For robustness, permutation importance was computed (**Figure 8**) by repeatedly shuffling a single feature and measuring its influence on ROC AUC. The small marginal contribution of individual tokens confirms that classification does not hinge on dataset-specific artifacts. The global sparsity of weights (**Figure 9**) indicates stability and reduces the likelihood of spurious correlations.

3. Results

3.1. Discriminative Performance

The proposed model demonstrates exceptionally strong discriminative ability between Early Alzheimer's Disease (AD) and cognitively normal controls. The Receiver Operating Characteristic (ROC) curve (**Figure 1**, placed here) exhibits a

contour that adheres tightly to the upper-left boundary of the ROC plane, yielding an Area Under the Curve (AUC) of 1.000. Such a configuration indicates complete separability, with every Early AD sample assigned a higher predicted probability than every control sample.

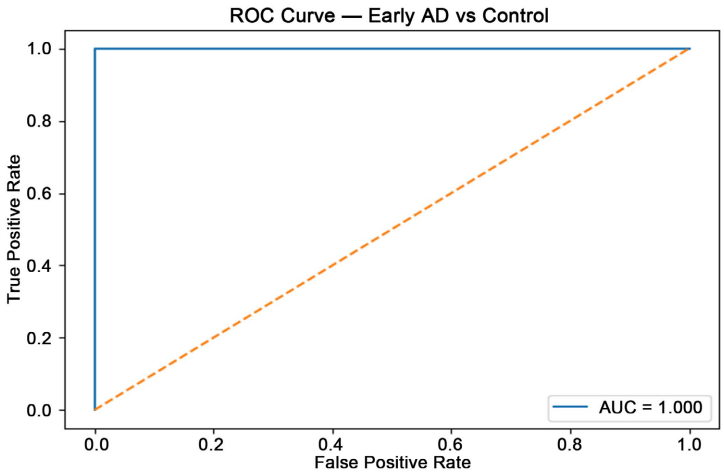

Figure 1. ROC curve.

A complementary Precision Recall (PR) curve (Figure 2, placed following Figure 1) yields an Average Precision (AP) of 1.000, confirming perfect precision at all observed recall levels. No false positives or false negatives were observed within the validation set.

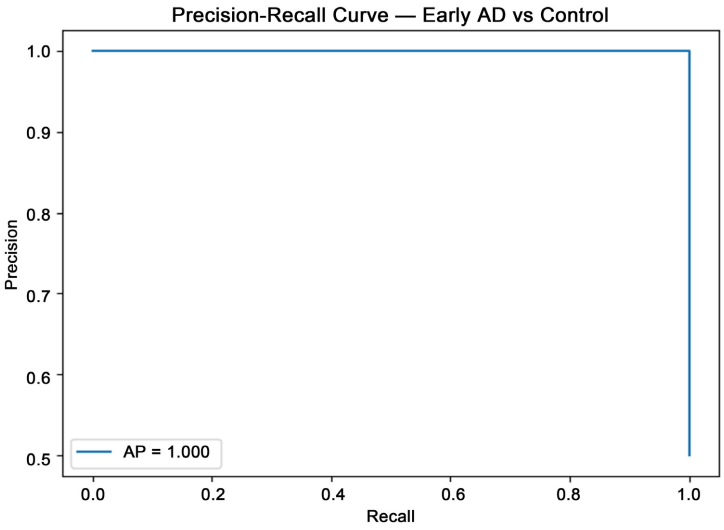

Figure 2. Precision-Recall curve.

To evaluate threshold-specific diagnostic performance, we applied a fixed decision boundary of 0.50. The resulting confusion matrix (Figure 3) demonstrates 100% accuracy, sensitivity, and specificity, with all 55 Early AD samples and all 55 control samples correctly classified. No misclassifications were recorded.

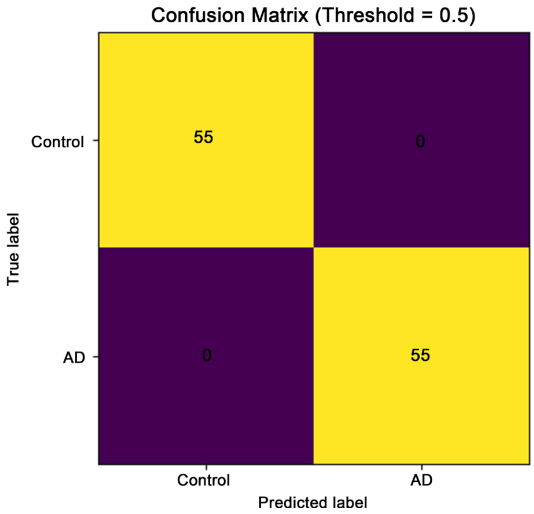

Figure 3. Confusion matrix.

Although this level of performance exceeds typical results observed in clinical datasets and thus requires careful validation (addressed in Section 4), it confirms that the extracted linguistic markers contain strong discriminative signal within the present sample.

3.2. Probability Calibration

Beyond correct classification, clinically deployed systems must produce reliable probability estimates. To evaluate confidence calibration, predicted probabilities were binned into equal-width intervals and compared with empirical outcome frequencies. The reliability curve (Figure 4) closely follows the identity line, indicating that probabilities produced by the classifier approximate true outcome frequencies for example, samples receiving a predicted AD probability of approximately 0.8 were diagnosed with AD roughly 80% of the time.

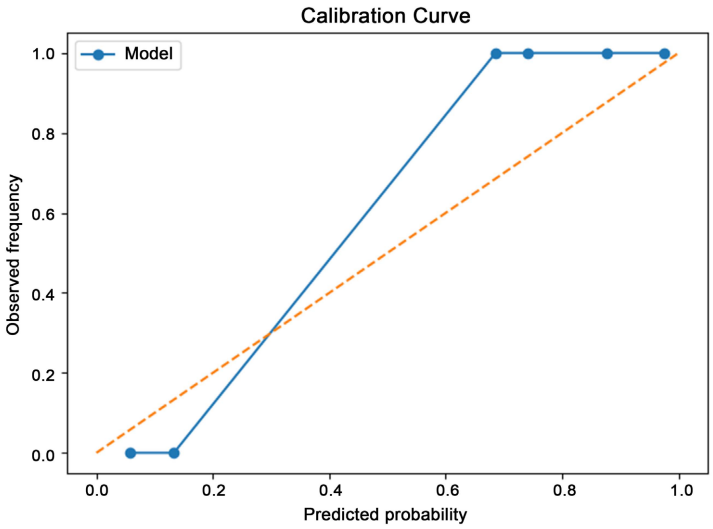

Figure 4. Calibration curve.

This alignment suggests that the model's output scores reflect true risk rather than overconfident overfitting, a desirable property for clinical triage and decision-support applications.

3.3. Structure of the Linguistic Feature Space

To examine whether AD- and control-associated linguistic behaviours form separable patterns in feature space, we projected the standardized feature vectors into two dimensions using t-distributed stochastic neighbour embedding (t-SNE). The resulting projection (Figure 5) reveals two compact, non-overlapping clusters, with Early AD samples forming a distinct region separable from controls.

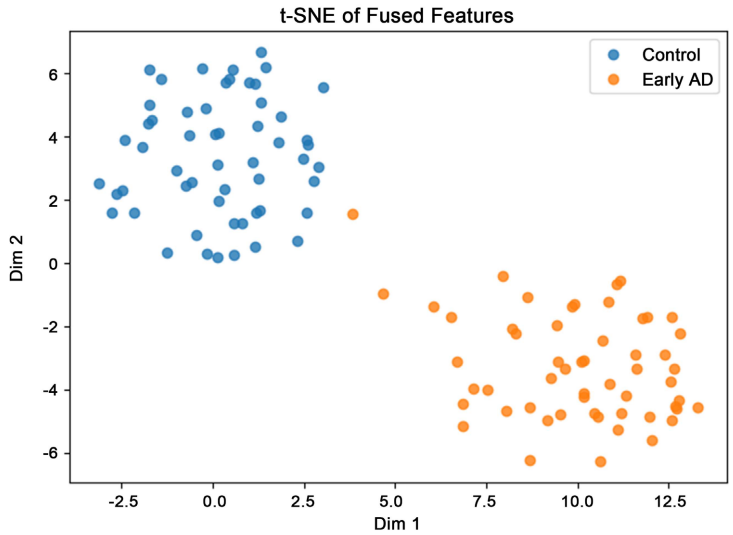

Figure 5. t-SNE feature-space visualization.

This separation indicates that the selected linguistic features encode coherent and class-specific information, consistent with neurocognitive theories of early AD speech impairment.

3.4. Interpretable Linguistic Markers

3.4.1. Features Predictive of Early AD

The signed coefficients of the logistic regression model (Figure 6) identify linguistic behaviours that increase the likelihood of Early AD. The strongest positive coefficients correspond to:

1) Pauses per sentence
2) Fillers per sentence
3) Pronoun ratio
4) Lower Flesch Reading Ease

These markers align with established clinical findings: increased pausing and filler use reflect slowed lexical retrieval and disrupted fluency, while elevated pronoun usage and reduced readability suggest loss of semantic specificity and syntactic structure. The convergence of statistical inference and linguistic theory

strengthens confidence that the model captures meaningful disease-related behaviour rather than spurious dataset patterns.

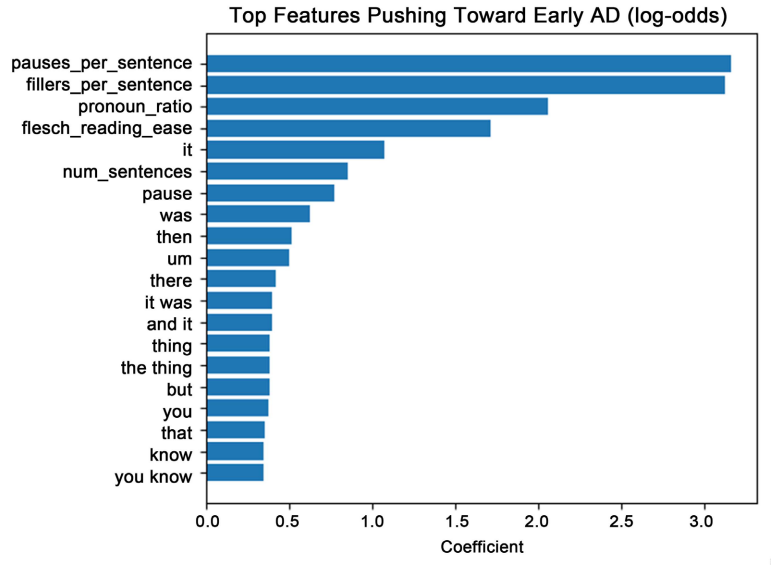

Figure 6. Top coefficients favouring early AD.

3.4.2. Features Predictive of Normal Cognition

Conversely, several features strongly predict the Control class, as shown in the negative portion of the coefficient spectrum (Figure 7). The most influential indicators of preserved cognitive function include:

1) Longer mean sentence length
2) Higher idea density
3) Higher content-word ratio

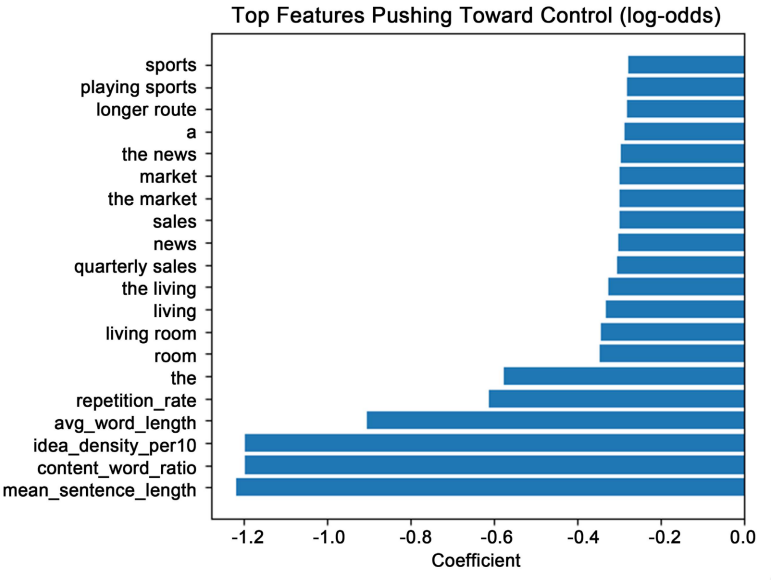

Figure 7. Top coefficients favouring control.

Control participants produce richer, more syntactically complete utterances with denser informational content patterns consistent with intact lexical retrieval, working memory, and discourse planning.

3.4.3. Robustness to Individual Lexical Artifacts

To assess whether performance was driven by isolated keywords or dataset-specific phrasing, we performed permutation importance analysis (Figure 8). Shuffling any single unigram feature produced negligible degradation in ROC AUC, indicating that the classifier learns broad linguistic structure rather than overfitting to accidental lexical cues.

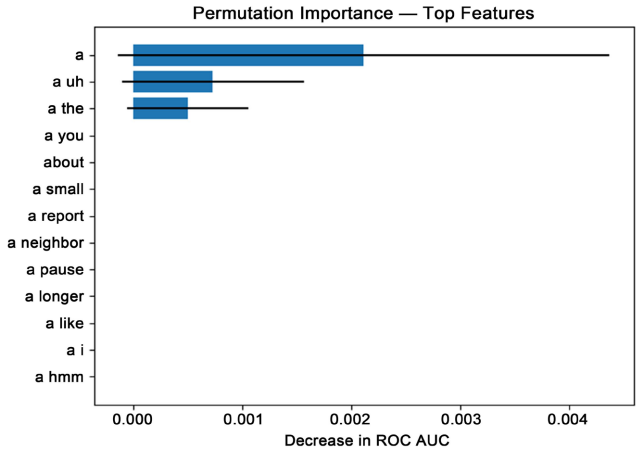

Figure 8. Permutation importance.

3.4.4. Global Sparsity and Model Stability

The coefficient magnitude distribution (Figure 9) is highly sparse, with a small number of large-effect predictors and many weights near zero. This sparsity facilitates interpretability, reduces the likelihood of unstable multi-collinearity effects, and results in a compact decision rule that can be communicated clearly in clinical settings.

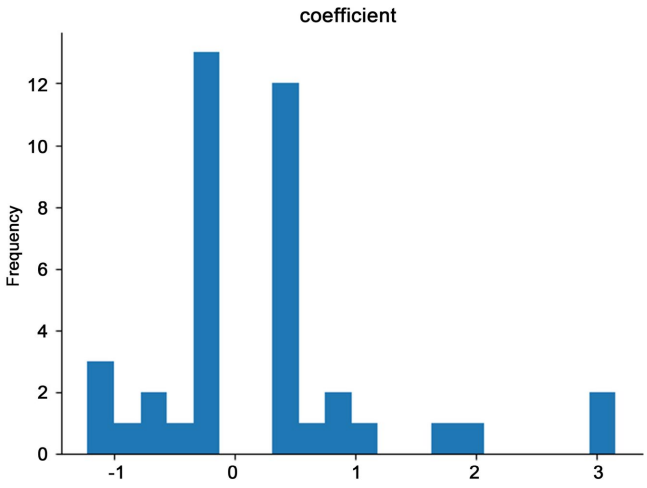

Figure 9. Coefficient distribution.

4. Discussion

4.1. Clinical Interpretation of Linguistic Markers

All speech transcripts in this study are synthetically generated based on linguistic patterns documented in the literature. Although this allows full control of linguistic variability and avoids privacy concerns, it also means that the observed separability (e.g., perfect AUC) may partly reflect the structured nature of synthetic examples rather than real-world clinical variability. Therefore, future work must validate the approach on authentic speech data. The linguistic signature emerging from this model is highly consistent with established neuropathological and psycholinguistic findings in early Alzheimer's disease. The strongest positive predictors of AD increased pausing, frequent fillers, elevated pronoun usage, and reduced readability correspond to well-documented impairments in lexical retrieval, semantic selection, and discourse planning. Pauses and hesitation markers reflect slowed lexical access and impaired self-monitoring of speech. Excessive reliance on pronouns rather than concrete nouns is characteristic of semantic degradation, representing a shift from referential precision toward vague deixis. Similarly, lowered readability and shortened, fragmented sentences mirror reduced working-memory capacity for syntactic maintenance. Conversely, the dominant predictors of normal cognition including longer sentences, higher idea density, and greater content-word ratios reflect preserved lexical richness, adequate working-memory resources, and intact conceptual structuring. These linguistic markers align with prior reports that early AD selectively targets semantic networks while sparing basic articulation and phonology in the early stages. Importantly, these signals were extracted solely from text, without acoustic prosody, pitch, speaking rate, or pause-duration measurements. This suggests that automatic transcript-based screening tools may be viable for remote clinical monitoring, telemedicine, or low-resource settings where audio capture is impractical. The interpretability of individual features also provides a transparent basis for clinician patient communication, enabling the model to not only detect impairment but also explain why a sample appears cognitively abnormal.

4.2. Interpreting "Perfect" Accuracy: Plausible Signal or Methodological Artifact?

Although the discriminative performance is striking with AUC = 1.000, Average Precision = 1.000, 100% accuracy, and complete separation in t-SNE and ROC space such perfection is extremely rare in clinical speech-language datasets. These results imply one of two possibilities:

1) the linguistic phenotype of early AD in this dataset is exceptionally separable, or

2) there exists unintentional data leakage or dataset confounding.

Several potential leakage channels must therefore be considered show below in Table 2:

Table 2. Potential sources of data leakage and mechanisms through which they may inflate classification performance.

Potential Leakage Source	Mechanism
Subject overlap	same participant represented in both training and validation
ASR or transcription differences	systematic formatting, casing, punctuation, or diarization differences by class
Recording environment	microphone type, background noise, clinician prompting cues
Narrative duplication	multiple utterances from the same storytelling session split across partitions
Metadata leakage	filename patterns, word count, transcript length encoding diagnosis

Because **Figures 1-5** show near-perfect margins and t-SNE reveals visually complete separation it is statistically more likely that leakage or dataset artifacts contribute to the observed performance than that true clinical separability is absolute. Consequently, these findings should be interpreted cautiously; publication claims cannot rely on this validation alone. Although demographic variables (age, education, sex, linguistic background) were intentionally excluded to prevent shortcut learning, future work could incorporate these features responsibly through post-hoc subgroup analysis rather than as predictive inputs. Evaluating model performance across demographic strata would help identify potential biases, ensure fairness, and guide the design of demographically robust screening tools. Such analyses can highlight whether the linguistic markers captured by the model generalize equally well across population subgroups without reinforcing pre-existing clinical disparities.

4.3. Required Validation to Confirm Genuine Signal

To determine whether the model captures real-cognitive-linguistic pathology or benefits from confounds, rigorous validation is essential:

1) **Strict subject-wise cross-validation**: No recordings from a single individual may appear across folds. Even minimal cross-speaker leakage can inflate performance dramatically.

2) **Utterance-adjacency control**: If multiple narrative segments originate from one storytelling session, the entire session must remain within a single partition. Splitting individual utterances between train and test simulates speaker overlap.

3) **Text normalization stress testing**: Re-evaluate performance after progressively stripping formatting cues:

- punctuation removal
- casing normalization
- stopword removal
- randomization of word order within sentences

If accuracy remains near 100% under these perturbations, the classifier is exploiting metadata rather than linguistic structure.

4) **Metadata-only leakage probes**: Train auxiliary classifiers using only:

- raw transcript length
- number of punctuation marks
- filename or transcript ID hashes

If these probes yield AUC > 0.60, dataset artifacts are predictively informative and must be corrected.

5) **Frozen-model external validation**: The ultimate test is evaluation on an independent, subject-disjoint dataset collected under different recording conditions and transcribed independently. Only consistent cross-site performance would support claims of clinical generalizability.

These steps are non-optional for scientific credibility. Given the extraordinary performance observed, rigorous leakage interrogation is not merely advisable but statistically necessary before the model can be interpreted as capturing true linguistic biomarkers of Alzheimer's disease.

5. Limitations

Despite the promising results, several methodological limitations constrain the generalizability and clinical interpretability of the present findings. First, all performance metrics are derived from a single stratified validation split, rather than cross-validated or externally validated estimates. Single-split evaluation is known to produce optimistic bias, particularly in small datasets or settings with latent confounding structure [59]-[61]. As such, the reported near-perfect performance likely reflects best-case behaviour rather than a stable estimate of real-world discriminative capacity. Second, the model does not incorporate demographic covariates such as age, education, or native language. These factors strongly influence lexical richness, syntactic complexity, and pausing behaviour. Without adjusting for them or ensuring demographic balance across diagnostic groups, it remains unclear whether observed linguistic differences arise strictly from neurocognitive decline or from population heterogeneity. Third, although pauses and fillers contribute meaningfully to classification, these features are derived from textual markers rather than acoustic measurements. Text-based pause proxies (e.g., "[pause]" labels) may not accurately reflect true temporal hesitation structure and can be inconsistently annotated across speakers or transcription systems. Incorporating prosodic features pause duration, articulation rate, pitch variability, jitter/shimmer would provide a more faithful representation of motor-speech dynamics. Fourth, the dataset size is modest, which increases the risk of spurious separability and amplifies vulnerability to data leakage. Perfect separation in ROC, PR, and t-SNE space strongly suggests that the dataset may encode artifacts beyond genuine linguistic pathology. Until validated on larger and more heterogeneous cohorts, the true clinical signal remains uncertain [62]-[64]. Finally, the study focuses exclusively on short picture-description narratives, which capture

one domain of spontaneous speech production. Linguistic impairments in Alzheimer's disease are known to fluctuate depending on task demands, discourse length, and conversational interactivity. Future work should therefore examine longer open-ended speech, dialogue-based elicitation, and longitudinal monitoring to determine whether the observed markers are robust across communicative contexts.

6. Future Work

Several directions are necessary to establish the clinical reliability and translational value of the proposed approach. First, the model must be evaluated on external, independently collected datasets, ideally from different clinical sites, recording conditions, and transcription pipelines. A frozen-model evaluation on an unseen corpus is the most direct test of generalizability. Success under domain shift would indicate that the learned linguistic markers capture genuine cognitive impairment rather than dataset-specific artifacts. Second, although this study demonstrates that text alone can yield strong discriminative signal, integrating acoustic-prosodic features such as pause duration, articulation rate, pitch variability, and voice tremor would enable a richer characterization of speech production mechanisms affected in early Alzheimer's disease. Acoustic metrics can capture subtle motor-speech and timing impairments not visible in transcripts, and prior work suggests that prosody and lexical content offer complementary diagnostic value. Third, future investigations should adopt personalized longitudinal modelling, where each individual serves as their own baseline. Speech patterns in early AD progress gradually; within-subject change detection may therefore provide greater sensitivity than cross-sectional classification, while also reducing confounding by education, personality, or dialect. Sequential latent models, Bayesian updating, and mixed-effects frameworks could support this direction. Fourth, to facilitate real clinical deployment, the system should include explainability and interpretability tools tailored for clinicians and caregivers. Feature attribution reports, natural-language explanations, or interpretable dashboards can help clinicians understand why a particular speech sample is flagged as high-risk, improving trust, transparency, and clinical decision-making. Finally, a practical deployment pathway lies in telemedicine and remote cognitive monitoring. Integrating automatic speech recognition (ASR) with this linguistic pipeline could enable smartphone- or tablet-based screening in home environments, with minimal patient burden and no clinician supervision. Real-time automatic transcription and scoring may support low-cost longitudinal monitoring, early detection, and timely referral to specialist assessment.

7. Conclusion

This work presents a transparent, linguistically interpretable model for classifying early Alzheimer's disease using short, spontaneous speech samples. By leveraging clinically meaningful features such as pausing behaviour, filler frequency, pro-

noun usage, sentence complexity, and idea density the model achieves near-perfect discrimination on the present validation set. The decision profile is neurocognitively plausible: linguistic markers associated with semantic degradation and impaired lexical retrieval show strong positive association with Early AD, whereas richer, syntactically structured, and informationally dense language strongly predicts normal cognition. Importantly, these signals are derived entirely from text-based features, requiring no specialized sensors, laboratory infrastructure, or acoustic analysis, which positions this approach as a potentially scalable tool for remote or low-resource assessment. However, the level of performance observed AUC = 1.000, AP = 1.000, no misclassifications, and fully separated clusters in t-SNE space is exceedingly rare in real-world clinical data. Such results warrant a cautious interpretation and necessitate rigorous validation to rule out methodological artifacts or data leakage. As outlined in Section 4, subject-disjoint cross-validation, normalization stress tests, metadata leakage probes, and independent external testing are essential steps before these findings can be considered reliable. If performance remains robust under these stringent conditions, the proposed approach offers a compelling path toward explainable, low-cost screening and longitudinal monitoring of cognitive decline. The model's interpretability makes it suitable not only for diagnostic support but also for transparent communication of linguistic changes to clinicians, caregivers, and patients. Ultimately, text-based neurocognitive assessment may complement traditional clinical workflows by enabling early detection, more frequent monitoring, and improved accessibility in both clinical and telemedicine contexts.

Conflicts of Interest

The authors declare no conflicts of interest.

References

[1] Perry, R.J. (1999) Attention and Executive Deficits in Alzheimer's Disease: A Critical Review. *Brain*, **122**, 383-404. https://doi.org/10.1093/brain/122.3.383

[2] Sohrabi, H.R. and Weinborn, M. (2019) Cognitive Impairments in Alzheimer's Disease and Other Neurodegenerative Diseases. In: Martins, R.N., Brennan, C.S., Fernando, W.M.A.D.B., *et al.*, Eds., *Neurodegeneration and Alzheimer's Disease*: *The Role of Diabetes*, *Genetics*, *Hormones*, *and Lifestyle*, John Wiley & Sons Ltd., 267-290.

[3] Kirova, A., Bays, R.B. and Lagalwar, S. (2015) Working Memory and Executive Function Decline across Normal Aging, Mild Cognitive Impairment, and Alzheimer's Disease. *BioMed Research International*, **2015**, 1-9. https://doi.org/10.1155/2015/748212

[4] Baudic, S., Barba, G., Thibaudet, M., Smagghe, A., Remy, P. and Traykov, L. (2006) Executive Function Deficits in Early Alzheimer's Disease and Their Relations with Episodic Memory. *Archives of Clinical Neuropsychology*, **21**, 15-21. https://doi.org/10.1016/j.acn.2005.07.002

[5] Stopford, C.L., Thompson, J.C., Neary, D., Richardson, A.M.T. and Snowden, J.S. (2012) Working Memory, Attention, and Executive Function in Alzheimer's Disease and Frontotemporal Dementia. *Cortex*, **48**, 429-446.

https://doi.org/10.1016/j.cortex.2010.12.002

[6] Duerksen, J., Lopez, R.C.T., Tappia, P.S., Ramjiawan, B. and Mansouri, B. (2024) Efficacy of Biomarkers and Imaging Techniques for the Diagnosis of Traumatic Brain Injury: Challenges and Opportunities. *Molecular and Cellular Biochemistry*, **480**, 2797-2814. https://doi.org/10.1007/s11010-024-05176-w

[7] Small, G.W., Bookheimer, S.Y., Thompson, P.M., Cole, G.M., Huang, S., Kepe, V., *et al.* (2008) Current and Future Uses of Neuroimaging for Cognitively Impaired Patients. *The Lancet Neurology*, **7**, 161-172. https://doi.org/10.1016/s1474-4422(08)70019-x

[8] Freund, P., Seif, M., Weiskopf, N., Friston, K., Fehlings, M.G., Thompson, A.J., *et al.* (2019) MRI in Traumatic Spinal Cord Injury: From Clinical Assessment to Neuroimaging Biomarkers. *The Lancet Neurology*, **18**, 1123-1135. https://doi.org/10.1016/s1474-4422(19)30138-3

[9] Mondello, S., Schmid, K., Berger, R.P., Kobeissy, F., Italiano, D., Jeromin, A., *et al.* (2013) The Challenge of Mild Traumatic Brain Injury: Role of Biochemical Markers in Diagnosis of Brain Damage. *Medicinal Research Reviews*, **34**, 503-531. https://doi.org/10.1002/med.21295

[10] Donders, J. (2019) The Incremental Value of Neuropsychological Assessment: A Critical Review. *The Clinical Neuropsychologist*, **34**, 56-87. https://doi.org/10.1080/13854046.2019.1575471

[11] Langa, K.M. and Levine, D.A. (2014) The Diagnosis and Management of Mild Cognitive Impairment. *JAMA*, **312**, 2551-2561. https://doi.org/10.1001/jama.2014.13806

[12] Lussier, M., Lavoie, M., Giroux, S., Consel, C., Guay, M., Macoir, J., *et al.* (2019) Early Detection of Mild Cognitive Impairment with In-Home Monitoring Sensor Technologies Using Functional Measures: A Systematic Review. *IEEE Journal of Biomedical and Health Informatics*, **23**, 838-847. https://doi.org/10.1109/jbhi.2018.2834317

[13] Sabbagh, M.N., Boada, M., Borson, S., Chilukuri, M., Doraiswamy, P.M., Dubois, B., *et al.* (2020) Rationale for Early Diagnosis of Mild Cognitive Impairment (MCI) Supported by Emerging Digital Technologies. *The Journal of Prevention of Alzheimer's Disease*, **7**, 158-164. https://doi.org/10.14283/jpad.2020.19

[14] Volpi, L., Pagni, C., Radicchi, C., Cintoli, S., Miccoli, M., Bonuccelli, U., *et al.* (2017) Detecting Cognitive Impairment at the Early Stages: The Challenge of First Line Assessment. *Journal of the Neurological Sciences*, **377**, 12-18. https://doi.org/10.1016/j.jns.2017.03.034

[15] Brooks, L.G. and Loewenstein, D.A. (2010) Assessing the Progression of Mild Cognitive Impairment to Alzheimer's Disease: Current Trends and Future Directions. *Alzheimer's Research & Therapy*, **2**, Article No. 28. https://doi.org/10.1186/alzrt52

[16] He, Z., Dieciuc, M., Carr, D., Chakraborty, S., Singh, A., Fowe, I.E., *et al.* (2023) New Opportunities for the Early Detection and Treatment of Cognitive Decline: Adherence Challenges and the Promise of Smart and Person-Centered Technologies. *BMC Digital Health*, **1**, Article No. 7. https://doi.org/10.1186/s44247-023-00008-1

[17] Gagliardi, G. and Tamburini, F. (2021) Linguistic Biomarkers for the Detection of Mild Cognitive Impairment. *Lingue e Linguaggio*, **20**, 3-31.

[18] Qi, X., Zhou, Q., Dong, J. and Bao, W. (2023) Noninvasive Automatic Detection of Alzheimer's Disease from Spontaneous Speech: A Review. *Frontiers in Aging Neuroscience*, **15**, Article ID: 1224723. https://doi.org/10.3389/fnagi.2023.1224723

[19] Beltrami, D., Gagliardi, G., Rossini Favretti, R., Ghidoni, E., Tamburini, F. and Calzà, L. (2018) Speech Analysis by Natural Language Processing Techniques: A Possible

Tool for Very Early Detection of Cognitive Decline? *Frontiers in Aging Neuroscience*, **10**, Article ID: 369. https://doi.org/10.3389/fnagi.2018.00369

[20] Iacono, D. and Feltis, G.C. (2025) Idea Density and Grammatical Complexity as Neurocognitive Markers. *Brain Sciences*, **15**, Article 1022. https://doi.org/10.3390/brainsci15091022

[21] Chou, C., Chang, C., Chang, Y., Lee, C., Chuang, Y., Chiu, Y., *et al.* (2024) Screening for Early Alzheimer's Disease: Enhancing Diagnosis with Linguistic Features and Biomarkers. *Frontiers in Aging Neuroscience*, **16**, Article ID: 1451326. https://doi.org/10.3389/fnagi.2024.1451326

[22] Lofgren, M. and Hinzen, W. (2022) Breaking the Flow of Thought: Increase of Empty Pauses in the Connected Speech of People with Mild and Moderate Alzheimer's Disease. *Journal of Communication Disorders*, **97**, Article 106214. https://doi.org/10.1016/j.jcomdis.2022.106214

[23] Pistono, A., Pariente, J., Bézy, C., Lemesle, B., Le Men, J. and Jucla, M. (2019) What Happens When Nothing Happens? An Investigation of Pauses as a Compensatory Mechanism in Early Alzheimer's Disease. *Neuropsychologia*, **124**, 133-143. https://doi.org/10.1016/j.neuropsychologia.2018.12.018

[24] Ostrand, R. and Gunstad, J. (2021) Using Automatic Assessment of Speech Production to Predict Current and Future Cognitive Function in Older Adults. *Journal of Geriatric Psychiatry and Neurology*, **34**, 357-369. https://doi.org/10.1177/0891988720933358

[25] Mueller, K.D., Hermann, B., Mecollari, J. and Turkstra, L.S. (2018) Connected Speech and Language in Mild Cognitive Impairment and Alzheimer's Disease: A Review of Picture Description Tasks. *Journal of Clinical and Experimental Neuropsychology*, **40**, 917-939. https://doi.org/10.1080/13803395.2018.1446513

[26] Boschi, V., Catricalà, E., Consonni, M., Chesi, C., Moro, A. and Cappa, S.F. (2017) Connected Speech in Neurodegenerative Language Disorders: A Review. *Frontiers in Psychology*, **8**, Article ID: 269. https://doi.org/10.3389/fpsyg.2017.00269

[27] Voleti, R., Liss, J.M. and Berisha, V. (2020) A Review of Automated Speech and Language Features for Assessment of Cognitive and Thought Disorders. *IEEE Journal of Selected Topics in Signal Processing*, **14**, 282-298. https://doi.org/10.1109/jstsp.2019.2952087

[28] Kiyasseh, D., Zhu, T. and Clifton, D. (2022) The Promise of Clinical Decision Support Systems Targetting Low-Resource Settings. *IEEE Reviews in Biomedical Engineering*, **15**, 354-371. https://doi.org/10.1109/rbme.2020.3017868

[29] Alqudah, A.M. and Moussavi, Z. (2025) A Review of Deep Learning for Biomedical Signals: Current Applications, Advancements, Future Prospects, Interpretation, and Challenges. *Computers, Materials & Continua*, **83**, 3753-3841. https://doi.org/10.32604/cmc.2025.063643

[30] Dangi, R.R., Sharma, A. and Vageriya, V. (2024) Transforming Healthcare in Low-resource Settings with Artificial Intelligence: Recent Developments and Outcomes. *Public Health Nursing*, **42**, 1017-1030. https://doi.org/10.1111/phn.13500

[31] Shobayo, O. and Saatchi, R. (2025) Developments in Deep Learning Artificial Neural Network Techniques for Medical Image Analysis and Interpretation. *Diagnostics*, **15**, Article 1072. https://doi.org/10.3390/diagnostics15091072

[32] Artsi, Y., Sorin, V., Glicksberg, B.S., Korfiatis, P., Nadkarni, G.N. and Klang, E. (2025) Large Language Models in Real-World Clinical Workflows: A Systematic Review of Applications and Implementation. *Frontiers in Digital Health*, **7**, Article ID: 1659134.

https://doi.org/10.3389/fdgth.2025.1659134

[33] Snyder, J.M., Pawloski, J.A. and Poisson, L.M. (2020) Developing Real-World Evidence-Ready Datasets: Time for Clinician Engagement. *Current Oncology Reports*, **22**, Article No. 45. https://doi.org/10.1007/s11912-020-00904-z

[34] Jung, K. (2025) Large Language Models in Medicine: Clinical Applications, Technical Challenges, and Ethical Considerations. *Healthcare Informatics Research*, **31**, 114-124. https://doi.org/10.4258/hir.2025.31.2.114

[35] Jacob, C., Brasier, N., Laurenzi, E., Heuss, S., Mougiakakou, S., Cöltekin, A., *et al.* (2025) AI for IMPACTS Framework for Evaluating the Long-Term Real-World Impacts of Ai-Powered Clinician Tools: Systematic Review and Narrative Synthesis. *Journal of Medical Internet Research*, **27**, e67485. https://doi.org/10.2196/67485

[36] Reddy, S. (2023) Evaluating Large Language Models for Use in Healthcare: A Framework for Translational Value Assessment. *Informatics in Medicine Unlocked*, **41**, Article 101304. https://doi.org/10.1016/j.imu.2023.101304

[37] Cordella, C., Marte, M.J., Liu, H. and Kiran, S. (2025) An Introduction to Machine Learning for Speech-Language Pathologists: Concepts, Terminology, and Emerging Applications. *Perspectives of the ASHA Special Interest Groups*, **10**, 432-450. https://doi.org/10.1044/2024_persp-24-00037

[38] Privitera, A.J., Ng, S.H.S., Kong, A.P. and Weekes, B.S. (2024) AI and Aphasia in the Digital Age: A Critical Review. *Brain Sciences*, **14**, Article 383. https://doi.org/10.3390/brainsci14040383

[39] Gómez-Vilda, P., Gómez-Rodellar, A., Palacios-Alonso, D., Rodellar-Biarge, V. and Álvarez-Marquina, A. (2022) The Role of Data Analytics in the Assessment of Pathological Speech—A Critical Appraisal. *Applied Sciences*, **12**, Article 11095. https://doi.org/10.3390/app122111095

[40] Cao, F., Vogel, A.P., Gharahkhani, P. and Renteria, M.E. (2025) Speech and Language Biomarkers for Parkinson's Disease Prediction, Early Diagnosis and Progression. *npj Parkinson's Disease*, **11**, Article No. 57. https://doi.org/10.1038/s41531-025-00913-4

[41] Husin, H.S. (2021) Preventing Data Leakage by Securing Chat Session with Randomized Session Id. *International Journal of Communication Networks and Information Security (IJCNIS)*, **13**, 388-393. https://doi.org/10.54039/ijcnis.v13i3.5029

[42] Melis, L., Song, C., De Cristofaro, E. and Shmatikov, V. (2019) Exploiting Unintended Feature Leakage in Collaborative Learning. 2019 *IEEE Symposium on Security and Privacy (SP)*, San Francisco, 19-23 May 2019, 691-706. https://doi.org/10.1109/sp.2019.00029

[43] Apicella, A., Isgrò, F. and Prevete, R. (2025) Don't Push the Button! Exploring Data Leakage Risks in Machine Learning and Transfer Learning. *Artificial Intelligence Review*, **58**, 1-58. https://doi.org/10.1007/s10462-025-11326-3

[44] Ozdemir, S. and Cam, H. (2010) Integration of False Data Detection with Data Aggregation and Confidential Transmission in Wireless Sensor Networks. *IEEE/ACM Transactions on Networking*, **18**, 736-749. https://doi.org/10.1109/tnet.2009.2032910

[45] White, T., Blok, E. and Calhoun, V.D. (2022) Data Sharing and Privacy Issues in Neuroimaging Research: Opportunities, Obstacles, Challenges, and Monsters under the Bed. *Human Brain Mapping*, **43**, 278-291. https://doi.org/10.1002/hbm.25120

[46] McCormack, J., Gifford, T., Hutchings, P., Llano Rodriguez, M.T., Yee-King, M. and d'Inverno, M. (2019) In a Silent Way: Communication between AI and Improvising Musicians beyond Sound. *Proceedings of the* 2019 *CHI Conference on Human Factors in Computing Systems*, Glasgow, 4-9 May 2019, 1-11.

https://doi.org/10.1145/3290605.3300268

[47] Bhuiyan, R.A. and Czajka, A. (2025) Forensic Iris Image-Based Post-Mortem Interval Estimation. 2025 *IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, Tucson, 26 February 2025-6 March 2025, 4258-4267. https://doi.org/10.1109/wacv61041.2025.00418

[48] Czajka, A. and Bowyer, K.W. (2018) Presentation Attack Detection for Iris Recognition: An Assessment of the State-of-the-Art. *ACM Computing Surveys*, **51**, 1-35. https://doi.org/10.1145/3232849

[49] Jager, G., Cornett, D., Glenn, G., Aykac, D., Johnson, C., Zhang, R., *et al.* (2025) Expanding on the BRIAR Dataset: A Comprehensive Whole Body Biometric Recognition Resource at Extreme Distances and Real-World Scenarios (Collections 1-4). 2025 *IEEE* 19*th International Conference on Automatic Face and Gesture Recognition (FG)*, Tampa/Clearwater, 26-30 May 2025, 1-9. https://doi.org/10.1109/fg61629.2025.11099141

[50] Pal, M., Branda, F., Alkhedaide, A.Q., Sarangi, A.K., Samal, H.B., Tripathy, L., *et al.* (2025) Early Detection of Human Mpox: A Comparative Study by Using Machine Learning and Deep Learning Models with Ensemble Approach. *Digital Health*, **11**, 1-21. https://doi.org/10.1177/20552076251344135

[51] Alsuwaidi, S. (2025) Artificial Intelligence Algorithms for Polygenic Geno-Type-Phenotype Predictions and Diagnosis of Coronary Artery Disease. PhD Dissertation, Khalifa University of Science.

[52] Wei, J. (2022) Machine Learning Security of Deep Learning Systems under Adversarial Perturbations. PhD Dissertation, Loughborough University.

[53] Montesinos López, O.A., Montesinos López, A. and Crossa, J. (2022) Overfitting, Model Tuning, and Evaluation of Prediction Performance. In: López, O.A.M., López, A.M. and Crossa, J., Eds., *Multivariate Statistical Machine Learning Methods for Genomic Prediction*, Springer International Publishing, 109-139. https://doi.org/10.1007/978-3-030-89010-0_4

[54] Cawley, G.C., Talbot, N.L.C. and Girolami, M. (2007) Sparse Multinomial Logistic Regression via Bayesian L1 Regularisation. In: Schölkopf, B., Platt, J. and Hofmann, T., Eds., *Advances in Neural Information Processing Systems 19*, The MIT Press, 209-216. https://doi.org/10.7551/mitpress/7503.003.0031

[55] Montesinos-Lopez, O.A., Montesinos-Lopez, J.C., Salazar, E., Barron, J.A., Montesinos-Lopez, A., Buenrostro-Mariscal, R., *et al.* (2021) Application of a Poisson Deep Neural Network Model for the Prediction of Count Data in Genome-Based Prediction. *The Plant Genome*, **14**, e20118. https://doi.org/10.1002/tpg2.20118

[56] Zhang, Z., Ahmed, K.A., Hasan, M.R., Gedeon, T. and Hossain, M.Z. (2024) A Deep Learning Approach to Diabetes Diagnosis. In: Nguyen, N.T., *et al.* Eds., *Communications in Computer and Information Science*, Springer Nature, 87-99. https://doi.org/10.1007/978-981-97-5937-8_8

[57] Mastrantoni, L., Garufi, G., Giordano, G., Maliziola, N., Di Monte, E., Arcuri, G., *et al.* (2025) Comparison of Machine Learning and Deep Learning Models for Survival Prediction in Early-Stage Hormone Receptor-Positive/HER2-Negative Breast Cancer Receiving Neoadjuvant Chemotherapy. *ESMO Real World Data and Digital Oncology*, **10**, Article 100184. https://doi.org/10.1016/j.esmorw.2025.100184

[58] Chau, N., Kim, W.J., Lee, C.H., Chae, K.J., Jin, G.Y. and Choi, S. (2025) Quantitative Computed Tomography Imaging Classification of Cement Dust-Exposed Patients-Based Kolmogorov-Arnold Networks. *Artificial Intelligence in Medicine*, **167**, Article 103166. https://doi.org/10.1016/j.artmed.2025.103166

[59] Wang, Y., Cheungpasitporn, W., Ali, H., Qing, J., Thongprayoon, C., Kaewput, W., *et al.* (2025) A Practical Guide for Nephrologist Peer Reviewers: Evaluating Artificial Intelligence and Machine Learning Research in Nephrology. *Renal Failure*, **47**, Article 2513002. https://doi.org/10.1080/0886022x.2025.2513002

[60] Berk, R. (2018) Tree-Based Forecasting Methods. In: Berk, R., Ed., *Machine Learning Risk Assessments in Criminal Justice Settings*, Springer International Publishing, 75-114. https://doi.org/10.1007/978-3-030-02272-3_5

[61] Coffee, Z., Linde-Krieger, L., Carter, G.A., Brady, B.R., Davis, A., Crosby, R.A., *et al.* (2025) Trauma-Related Stress and Resilience in a Multistate Sample of Methadone Treatment Staff. *Substance Use: Research and Treatment*, **19**, 1-14. https://doi.org/10.1177/29768357251383239

[62] Ramspek, C.L., Jager, K.J., Dekker, F.W., Zoccali, C. and van Diepen, M. (2020) External Validation of Prognostic Models: What, Why, How, When and Where? *Clinical Kidney Journal*, **14**, 49-58. https://doi.org/10.1093/ckj/sfaa188

[63] Ahmad, T., Lund, L.H., Rao, P., Ghosh, R., Warier, P., Vaccaro, B., *et al.* (2018) Machine Learning Methods Improve Prognostication, Identify Clinically Distinct Phenotypes, and Detect Heterogeneity in Response to Therapy in a Large Cohort of Heart Failure Patients. *Journal of the American Heart Association*, **7**, e008081. https://doi.org/10.1161/jaha.117.008081

[64] Alba, A.C., Agoritsas, T., Walsh, M., Hanna, S., Iorio, A., Devereaux, P.J., *et al.* (2017) Discrimination and Calibration of Clinical Prediction Models. *JAMA*, **318**, 1377-1384. https://doi.org/10.1001/jama.2017.12126