Entropy-Guided Graph Neural Networks for Cognitive Brain Disorder Detection from Speech Transcripts

JANNATUL FERDUSH, Federation University, Australia
AZADEH NOORI HOSHYAR, Federation University, Australia
JAISON MULERIKKAL, Department of Information Technology, Rajagiri School of Engineering and Technology, India
BINU A, Department of Artificial Intelligence and Data Science, Rajagiri School of Engineering and Technology, India
ADEL AL-JUMAILY, School of IT and Engineering, Melbourne Institute of Technology, Australia, Department of Computer Science and Software Engineering, University of Western Australia, Australia, UNITAR International University, Malaysia, and School of Science, Edith Cowan University, Australia

Early and accurate diagnosis of brain diseases such as dementia, autism, and right hemisphere damage is crucial for effective intervention and treatment. Traditional diagnostic approaches often rely on clinical assessments, which can be time-consuming and subjective. In this study, we propose a methodology leveraging graph neural networks (GNNs) for brain disease classification based on speech transcripts or textual data. Our approach incorporates a subject similarity graph with an entropy-guided reinforcement regularisation mechanism to model complex linguistic patterns and capture inter-subject relationships, improving discrimination between neurological conditions. Experimental results demonstrate consistent performance gains over strong baseline models. The proposed framework provides a cost-effective and computationally efficient diagnostic tool with potential applicability to early-stage clinical decision support.

Additional Key Words and Phrases: Brain disease, GNN, Autism, Dementia, RHD, Entropy-guided regularisation

1 Introduction

Ageing is an inescapable aspect of human life. Despite endeavours to maintain a healthy and meaningful lifestyle, individuals frequently face a range of health-related challenges arising from genetic factors, environmental conditions, or wider socio-economic influences. Among these challenges, neurological disorders have become a significant concern, particularly those that affect cognitive functions.

Cognitive brain disorders such as dementia, autism spectrum disorder (ASD), and right hemisphere damage (RHD) significantly affect an individual's ability to process language, engage in effective communication, and interact with their surroundings. These disorders are characterised by disruptions in memory, attention, emotional

Authors' Contact Information: Jannatul Ferdush, Federation University, Berwick, VIC, Australia, j.ferdush@federation.edu.au; Azadeh Noori Hoshyar, Federation University, Berwick, VIC, Australia, a.noorihoshyar@federation.edu.au; Jaison Mulerikkal, Department of Information Technology, Rajagiri School of Engineering and Technology, Kochi, India, jaisonmpaul@rajagiritech.edu.in; Binu A, Department of Artificial Intelligence and Data Science, Rajagiri School of Engineering and Technology, Kochi, India, binua.cusat@gmail.com; Adel Al-Jumaily, School of IT and Engineering, Melbourne Institute of Technology, Sydney, NSW, Australia and Department of Computer Science and Software Engineering, University of Western Australia, Perth, WA, Australia and UNITAR International University, Petaling Jaya, Selangor Darul Ehsan, Malaysia and School of Science, Edith Cowan University, Joondalup, WA, Australia, adel.al-jumaily@ieee.org.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than the author(s) must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.
© 2026 Copyright held by the owner/author(s). Publication rights licensed to ACM.
ACM 2637-8051/2026/6-ART
https://doi.org/10.1145/3827607

Just Accepted

regulation, and social behaviour, often leading to a decline in quality of life [29, 37, 39, 41, 54]. With the global population continuing to age and awareness of neurodevelopmental and neurodegenerative conditions[1] on the rise, gaining understanding of the mechanisms, impacts, and interventions associated with cognitive brain disorders [12] has become a vital area of research, shown in Figure 1.

Brain disease statistics

Table 1. Statistics about brain diseases

Disease	Prevalence	Main Cognitive Effects	Age Group Affected
Dementia	55M cases worldwide	Memory loss, cognitive decline	Mostly 65+ years old[a]
Autism spectrum disorder (ASD)	1 in 100	Impaired social communication, repetitive behaviours	Diagnosed in childhood[b]
Right hemisphere damage (RHD)	30–50% of stroke patients	Impaired attention, spatial neglect	Varies, often older adults[c]

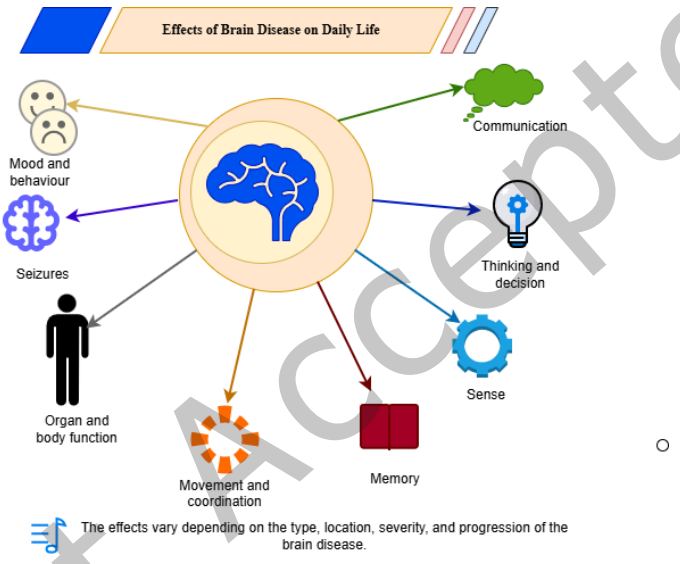

Fig. 1. Impact of Brain Disease on Daily Functioning[d]

[a]https://www.who.int/news-room/fact-sheets/detail/dementia; accessed 5th April, 2025
[b]https://www.who.int/news-room/fact-sheets/detail/autism-spectrum-disorders; accessed 5th April, 2025
[c]Sabbouh, Toni, and Michel T. Torbey. "Malnutrition in stroke patients..." Neurocritical care, 2018; accessed 5th April, 2025
[d]source of inspiration: https://brainfoundation.org.au/what-is-brain-disease/; accessed 26 March 2025

Traditional diagnostic methodologies, including clinical assessments, neuroimaging, and neuropsychological evaluations, although foundational, are often constrained by high operational costs, lengthy procedures, and dependence on specialist expertise [13, 27, 52]. Recent advancements in natural language processing (NLP) and transcript-based computational modelling have revealed that linguistic patterns encode critical cognitive markers of neurological dysfunction. Whilst several studies have utilised speech and text data for early-stage detection

[1]Neurodevelopmental conditions originate during early brain development and often affect cognition, behavior, and social functioning (e.g., autism spectrum disorder, ADHD). Neurodegenerative conditions involve the gradual loss of neuron structure or function over time, commonly seen in disorders like Alzheimer's and Parkinson's disease.

Just Accepted

[36, 38, 58], the prevailing body of work remains predominantly disorder-specific, frequently necessitating independent models for each condition. Existing approaches are further limited by sensitivity to inter-subject linguistic variability, restricted scalability, static feature representation, and insufficient dynamic modelling of semantic relationships within natural language. These constraints highlight the necessity for unified and adaptable frameworks capable of capturing evolving cognitive-linguistic impairments across broader neurological contexts. In particular, developing a single model that can robustly and simultaneously detect dementia, autism spectrum disorder (ASD), and right hemisphere damage (RHD) remains unresolved. The overlapping yet distinct linguistic manifestations associated with each disorder- ranging from pragmatic deficits to discourse-level incoherence- pose significant challenges for constructing generalisable, clinically interpretable models. Each of these disorders exhibits distinct linguistic and cognitive impairments, yet there is considerable overlap in their symptoms, shown in Table 1. Dementia is characterised by progressive cognitive decline, affecting memory, comprehension, and speech coherence [24]. Autism often presents with atypical language use, pragmatic difficulties, and repetitive speech patterns [42]. Right hemisphere damage (RHD) can lead to deficits in discourse coherence, prosody, and the interpretation of figurative language[5].

The complexity of these language impairments, combined with individual variability in expression, makes automated diagnosis using textual data particularly difficult. Current methods often struggle to accurately differentiate between these disorders due to limitations in feature extraction, interpretability, and the generalisability of algorithms (described above). Moreover, developing separate models for each disorder may lead to increased computational overhead and time consumption.

So, the motivation for this research stems from the potential of computational approaches to revolutionise neurological diagnostics by harnessing speech(transcript) and text data for early disease detection with a generalisation technique. By addressing the existing challenges of generalisation, this research contributes to the development of advanced diagnostic tools that can aid clinicians in early detection and intervention. The key contributions of this study are:

- Graph-based representation: Propose a graph-based learning approach with entropy-guided reinforcement regulariation to model complex relationships in language patterns, improving diagnostic precision.
- Demonstrate the capability of a unified modelling approach to accurately detect multiple brain disorders, thereby improving the adaptability, scalability, and clinical applicability of the proposed framework.
- Provide an extensive evaluation of conventional machine learning models and transformer-based baselines against graph-based techniques, highlighting their strengths and limitations.

2 Related works

Natural language processing (NLP) has become a powerful tool for analysing speech and text data in brain disease detection, particularly for dementia, autism spectrum disorder (ASD), and right hemisphere damage (RHD) [24, 36, 38]. Early studies typically relied on conventional, feature-engineering approaches such as n-grams, part-of-speech (PoS) statistics, lexical richness measures, and handcrafted syntactic features [4, 20, 50]. These traditional NLP pipelines were frequently combined with classical machine-learning classifiers (e.g., SVM, logistic regression, random forests) to detect linguistic abnormalities. More recent work has expanded to tasks such as sentiment analysis, topic modelling, and semantic coherence modelling, where researchers have applied both classical methods and modern deep-learning or transformer-based approaches [2, 46]. This evolution highlights a shift from surface-level linguistic features toward context-aware representations capable of capturing deeper cognitive and semantic patterns.

Just Accepted

However, these approaches often struggle to capture the deeper semantic relationships and contextual dependencies present in speech patterns. To address these limitations, advanced deep learning models, including transformers, recurrent neural networks (RNNs), and graph-based approaches, have been employed, significantly improving the accuracy and robustness of disease classification [6, 34].

Given the growing interest in brain disorders, this section reviews key contributions to brain disease detection. To provide a structured discussion, we divide this section into three main parts: (i) Dementia detection (DD), (ii) Autism detection (AD), and (iii) RHD detection, each focusing on the distinct linguistic markers and computational techniques applied to these disorders.

2.1 Dementia detection

Dementia is a progressive neurodegenerative disorder affecting cognition, language, and communication [45]. It also impacts memory, problem-solving abilities, emotional regulation, and daily functioning, gradually diminishing a person's independence.

In order to systematically compare approaches in dementia detection (DD), we organized the research literature into three modes: audio-only, transcript-only, and multimodal approaches. This organization maintains the nature of each approach while also providing an easy comparison of each approaches' strengths, weaknesses, and functions across multiple studies, highlighting the progression from single mode models to multi-modal approaches that are capable of leveraging both acoustic and linguistic information.

In 2022, Liu et al. [33] proposed a transformer-based feature purification network (GP-Net) to improve speech-based Alzheimer's disease (AD) detection by removing non-discriminative linguistic features from learned representations. By explicitly separating common and class-specific features, the model achieved state-of-the-art performance on the Pitt corpus, demonstrating the importance of representation refinement in neurodevelopmental and neurodegenerative disorder detection. However, the approach focuses exclusively on speech transcripts and requires complex adversarial training, which may limit scalability and adaptation to multimodal or real-time clinical settings.

Recent studies further highlight the effectiveness and interpretability of transformer-based language models for cognitive disorder detection. In 2021, Matosević et al. [34] employed RoBERTa on speech transcripts from the Pitt corpus and demonstrated that carefully preprocessed textual data alone can achieve over 90% accuracy in dementia detection, outperforming baseline BERT models. Despite strong results, the method depends heavily on preprocessing choices and transcript availability. In 2023, Wen et al. [57] investigated the role of part-of-speech (PoS) features using a transformer combined with causal explainable AI, achieving 92.2% accuracy while identifying linguistically meaningful biomarkers for AD. Although highly interpretable, this approach is limited to structured PoS features and does not leverage richer semantic embeddings or multimodal behavioural cues, indicating scope for broader fusion-based frameworks applicable to autism and related cognitive disorder detection.

The Table 2 summaries significant studies within each mode showcasing the methodologies, performance statistics, and other relevant comments. It provides a comprehensive overview of the field's development, from the early rule-based or shallow learning approaches, to the recent development of deep learning-based, end-to-end models, while utilising transfer learning, cross-mode embeddings, and contextual reasoning.

As shown across the studies summarised in Table 2, unimodal systems (audio-only or text-only) often achieve strong results on benchmark datasets. However, prior research has noted that their performance may vary across different recording conditions, speaker populations, or transcript sources, especially when dealing with ASR errors, background noise, or differences in elicitation protocols [19, 56]. Our analysis does not imply that unimodal systems are ineffective. In fact, our proposed method is also unimodal but rather that their robustness can depend heavily on data quality and context, which motivates continued exploration of improved text-based modelling

Just Accepted

Table 2. Summary of dementia Studies

Study	Approach	Performance	Notes	Modality
[55], 2018	Unsupervised diarisation and acoustic features	UAR = 0.645	Limited detection of mild cognitive decline due to class imbalance	Audio-Only
[22], 2019	Low-level acoustic features + ADR	Accuracy = 78.7%	No transcription needed; focus on voice quality, prosody	Audio-Only
[8], 2020	Acoustic biomarkers with classifiers	Accuracy = 82%	Good at distinguishing MCI and AD, limited by class overlap	Audio-Only
[3], 2022	Spectrograms + Autoencoders	Accuracy = 93.3%	Signal-level features only, no linguistic analysis	Audio-Only
[31], 2012	ASR-based story retelling (WLM) + SVM	Comparable to manual scoring	Automatic pipeline for scalable screening	Audio-Only
[1], 2023	Multifeature (acoustic + linguistic + phonetic)	AUC=92.6%	needs external validation	Audio-Only
[58], 2022	Stop words + PoS + hybrid models	Accuracy = 81.54%	Misses deeper semantic analysis	Transcript-Only
[6], 2022	semantic and syntactic language features extracted via transfer learning	Accuracy =81.54%	Focus on semantic and syntactic utility	Transcript-Only
[57], 2023	Causal explanation with PoS	Accuracy = 92.2%	Reinforces syntactic structures	Transcript-Only
[33], 2022	Feature purification network	Accuracy=93.5% on pitt dataset	Removes noisy/irrelevant features	Transcript-Only
[47], 2021	BERTLarge + logistic regression	Accuracy = 88.08%	Depends on transcript quality	Transcript-Only
[49], 2022	SBERT for semantic coherence	Accuracy = 77%	Requires extensive engineering	Transcript-Only
[28], 2022	Evaluation of BERT, XLNet	Accuracy = 93%	Needs pre-cleaned transcripts	Transcript-Only
[34], 2022	Roberta on Pitt Corpus	Accuracy=90.16%	Depends on long-range dependencies	Transcript-Only
[51], 2024	Pause-encoded transcripts + BERT	Accuracy = 86%	Symbolic pauses may limit robustness	Transcript-Only
[17], 2024	OpenAI embeddings + MLP	Accuracy = 86.93%	Lightweight but misses nonverbal cues	Transcript-Only
[24], 2024	Syntactic + semantic + SVM/KNN/RF/ANN	Accuracy>99	Improved early detection and lightweight	Transcript-Only
[56], 2017	ASR vs manual transcription	UAR = 0.623	High ASR errors reduce performance	Multimodal
[25], 2014	Lexical + acoustic features	Accuracy = 88%	Small dataset, shallow features	Multimodal
[23], 2022	BERT + Vision Transformers + Co-attention	Accuracy = 90%	Complex; overfit risk on small datasets	Multimodal
[9], 2023	Neural architecture search + multimodal fusion	Accuracy=92.08%	Sensitive to ASR errors	Multimodal
[10], 2025	Whisper ASR + Roberta	RMSE = 2.69	Fully multimodal; transcript + signal	Multimodal
[26], 2024	Whisper transcripts + Roberta	RMSE = 2.6911	Text-only from high-fidelity ASR	Multimodal
[35], 2025	GPT + BERT + CLAP	F1=83.33%	contextual embedding	Multimodal
[43], 2024	Cross-transformers	AUC=85.3%	Deeper intermodal interaction	Multimodal

Just Accepted

approaches.

Furthermore, the multimodal studies reviewed in Table 2 demonstrate that while combining acoustic and linguistic cues can improve performance, these methods often introduce higher computational cost, dependence on audio quality, and sensitivity to recording conditions. These observations motivated us to design a unified, text-based framework that aims to retain the semantic strengths of linguistic features while avoiding the practical limitations associated with multimodal audio–text pipelines.

2.2 Autism detection

Table 3 provides a systematic summary of recent research activity on autism detection using a variety of computational and data-driven automated technologies. The types of studies reviewed in this work relates to conventional machine learning, rule-based systems, federated learning, natural-language-based systems (NLP), and deep learning-based systems as they are used to improve scalability, accuracy, interpretability, and privacy for autism screening on diverse data types including clinical assessment data, questionnaires, social media, biomedical literature, and physiological signals such as eye movements. In 2020, Elbattah et al. [16] proposed an NLP-based machine learning framework for autism spectrum disorder (ASD) detection by modelling eye-tracking scanpaths as textual sequences composed of fixations and saccades. By transforming gaze behaviour into sequence representations, the study successfully connected visual attention patterns with sequence learning techniques and achieved a ROC-AUC of up to 0.84, demonstrating the feasibility of non-invasive, data-driven ASD screening. However, the approach is constrained by a limited sample size and controlled experimental conditions, which may reduce its robustness and applicability in real-world clinical environments.

More recently, autism detection research has expanded toward scalable and text-driven approaches. In 2023, Farooq et al. [18] introduced a federated learning–based ASD detection model using behavioural screening features from children and adults, emphasising data privacy and distributed learning while reporting high accuracy, particularly for pediatric cases. Despite its strong performance, the method relies primarily on questionnaire-based features and lacks multimodal behavioural signals. In 2024, Pal et al. [40] further advanced text-based autism research by applying BERT to predict ASD-related behavioral traits from textual descriptions of toddler behaviours, achieving 88% accuracy. While this work highlights the strength of deep language models for early ASD risk assessment, limited works are present in the literature.

The above mentioned reasearches (In table 3) demonstrate the variation of work in computational autism detection, showing the development of algorithms and the methods selected, use of example sources of all data collection, and data processing. Data collection methods vary from structured clinical assessments to one of full decentralization and community use to collect data and provide opportunities for scalable, inclusive, and non-intrusive ways for detecting and screening autism. While there are optimistic results and methods, it is clear there are challenges still to overcome, including data privacy, harmonization of features across multiple domains, Another commonly reported challenge in ASD detection is the limited interpretability of deep learning models. These models often operate as black boxes, making it difficult to understand which linguistic, behavioural, or semantic cues they rely on and why a particular prediction (e.g., Autism vs. Non-Autism) is made. This lack of transparency limits clinical trust and prevents researchers and clinicians from verifying whether the model is basing its decisions on meaningful indicators or dataset-specific biases. While these broader challenges remain open research problems, our work specifically addresses the need for scalable and robust text-based modelling by proposing a unified methodology that avoids dependency on audio quality. Although our method does not solve all limitations in ASD detection, it contributes toward more practical and adaptable solutions that can support future clinical screening efforts.

2.3 RHD detection

Research on the detection of right hemisphere damage (RHD) using machine or deep learning methods is still limited. While computational approaches have gained momentum in brain abnormality detection, most efforts focus on broader categories such as tumors or infarcts, rather than hemisphere-specific functional impairments.

Gauriau et al. [21] developed a deep learning model trained on MRI data to distinguish between likely normal and likely abnormal brain scans. Although effective in improving radiological workflows with an F1 score of 0.72, the model does not address the functional consequences of RHD, particularly those affecting language and communication. On the other hand, Kim et al. [29] highlighted the importance of these functional effects

Just Accepted

Table 3. Summary of autism studies

Study	Approach	Performance	Notes	Modality	Key Techniques
[30], 2015	Feature selection on ADOS behavioral scores	accuracy ≈ 98%	Clinical data, structured environment	Behavioral	Machine Learning, Feature Selection
[53], 2020)	Rule-based ML on app-based questionnaires	0.95-0.96(based on error rate)	Scalable, interpretability screening via self-reports	Questionnaire or Text	Rule-Based ML, Mobile App
[18], 2023	Federated Learning on 4 decentralized datasets	98% (children), 81% (adults)	Preserved data privacy, issues with feature harmonization	Questionnaire or Text	Federated Learning, SVM, Logistic Regression
[40], 2024	BERT classifier for behavioral trait prediction	Accuracy: 0.88	Effective for traits, not binary diagnosis	Text	BERT, Trait Classification
[16], 2020	NLP on eye-tracking data encoded as symbolic sequences	ROC-AUC: 0.84	ConvNets	Physiological (Eye-Tracking)	Eye Tracking, NLP, CNN
[15], 2018	NLP for suicidality detection in EHRs	F1 > 0.85	EHR-based passive risk monitoring for adolescents with ASD	Text (EHR)	Clinical NLP, Suicidality Detection
[38], 2023	BERT + ChatGPT for sentiment analysis of social media narratives	Accuracy = 83%	analysed parental narratives, informal text sources	Text (Social Media)	Social Media, Sentiment Analysis, BERT, ChatGPT
[48], 2024)	DL on 400,000 tweets using Bi-LSTM, BERT, BERTweet	Accuracy ≈ 88%	Self-reported ASD detection via Twitter	Text (Social Media)	Social Media, Deep Learning, BERTweet

Just Accepted

by demonstrating that individuals with RHD, despite intact basic language skills, often struggle with topic maintenance, conversational turn-taking, and pragmatic language use. These are key indicators of communicative disruption, but they are often undetectable in imaging alone, underscoring the need for language-based diagnostic tools.

In this context, natural language processing (NLP) emerges as a crucial avenue for RHD detection. RHD-related impairments often manifest in subtle pragmatic and discourse-level featuressuch as conversational breakdowns, reduced topic coherence, and emotional prosody that are better captured through speech and text analysis than through structural imaging. Cui et al. [11] proposed an NLP-driven vulnerability detection system that leverages both acoustic and textual features from short speech segments. Their model classifies speakers by age group, accent, and health status, including the presence of neurological conditions. By using publicly available datasets such as RHDBank which includes speech from RHD patients: they achieved detection accuracies of up to 93.91% in the patient vs. non-patient classification task. This demonstrates the strong potential of NLP-based models in identifying communication disorders that may otherwise go unnoticed in standard clinical assessments. In addition, Li et al. [32] demonstrated the usefulness of combining functional near-infrared spectroscopy (fNIRS) with continuous performance tests (CPT) to detect post-stroke cognitive impairment in RHD patients. Their findings indicated that specific oxygenated hemoglobin patterns in the prefrontal cortex can differentiate RHD patients from controls, suggesting that multimodal neuroimaging combined with task-based paradigms offers complementary insights beyond linguistic data alone. Similarly, Di Gregorio et al. [14] explored EEG-based markers and machine learning to predict functional recovery in RHD stroke patients, finding that fronto-parietal connectivity and low-frequency EEG power dynamics strongly correlate with post-rehabilitation outcomes. This line of research highlights that multimodal and electrophysiological data, when integrated with speech and text analytics, could significantly enhance RHD detection and prognostication frameworks. Finally, Rava et al. [44] assessed AI-driven intracranial hemorrhage (ICH) detection from CT scans, demonstrating high sensitivity and specificity, but also underscoring that structural abnormalities alone cannot fully capture the nuanced communicative deficits resulting from RHD, reinforcing the necessity for language and cognitive assessments.

Given the unique communication deficits associated with RHD, expanding research into NLP-based detection is essential for developing early, accessible, and functionally meaningful diagnostic tools that go beyond what traditional imaging can provide.

3 Motivation and Rationale for a Unified Text-Based Framework

While multimodal approaches that integrate both audio and transcript features have demonstrated strong diagnostic performance, particularly in the context of dementia, they present notable practical limitations. Audio-based systems depend on clean, high-quality recordings and robust signal processing pipelines to extract meaningful paralinguistic features. However, in real-world scenarios such as telehealth, home monitoring, or clinical environments with variable recording conditions, such audio quality cannot always be assured. Additionally, acoustic features are often sensitive to background noise, microphone variability, and speech irregularities, all of which are frequently observed in individuals affected by cognitive and neurological disorders.

By contrast, working with transcripts removes the need to process raw audio features directly, but it does not eliminate the challenges that come from audio. Transcripts still rely on ASR systems, and their quality is shaped by the recording environment, background noise, speaker characteristics, and speech impairments - which are very common in dementia, autism, and RHD. In other words, transcript-based models shift the difficulty to the ASR stage rather than removing it. While modern ASR tools have improved, especially for general speech, they can still struggle in clinical conversations, so this dependency needs to be acknowledged when discussing practicality and real-world use.

Moreover, text-only approaches reduce the computational overhead associated with audio processing and are less vulnerable to environmental variability. They also allow for greater flexibility in deployment across low-resource settings or remote assessments. Recent research has shown that well-designed text-based models can perform comparably to multimodal systems, especially when enhanced with context-aware embeddings and domain-specific features.

Just Accepted

However, despite these advantages, existing methods still face several limitations: including the scarcity of text focused approaches for certain disorders (e.g., RHD), the heavy reliance on audio quality in multimodal pipelines, and the lack of lightweight, clinically deployable models. To address these specific limitations, our work focuses on transcript-based analysis to develop accurate, interpretable, and clinically relevant models for detecting cognitive and communicative impairments without reliance on audio input.

4 System architecture

The proposed Graph Neural Network–based Brain Disease (GNN-BD) framework comprises two integrated components: (i) subject similarity graph construction using task-adapted language representations and (ii) graph neural network learning with entropy-guided reinforcement regulariation.

In the first component, a pretrained transformer language model is fine-tuned for subject-level brain disease classification. Fixed-length subject representations are then obtained from the final-layer [CLS] embeddings, capturing clinically relevant linguistic and semantic characteristics of each subject's speech. These embeddings are used as node features to construct a subject similarity graph, where each subject is connected to its k most similar peers based on cosine similarity, enabling the modelling of latent inter-subject relationships.

In the second component, a Graph Attention Network (GAT) is trained on the resulting similarity graph. To encourage reliable and confidence-aware predictions, the GNN is augmented with an entropy-guided reinforcement regulariation mechanism that operates during optimisation without altering the graph topology. This mechanism jointly learns a scalar value function and reweights the supervised classification loss based on an entropy-weighted reward, promoting confident correct predictions while penalizing overconfident errors. The overall workflow of the proposed framework is illustrated in Figure 2.

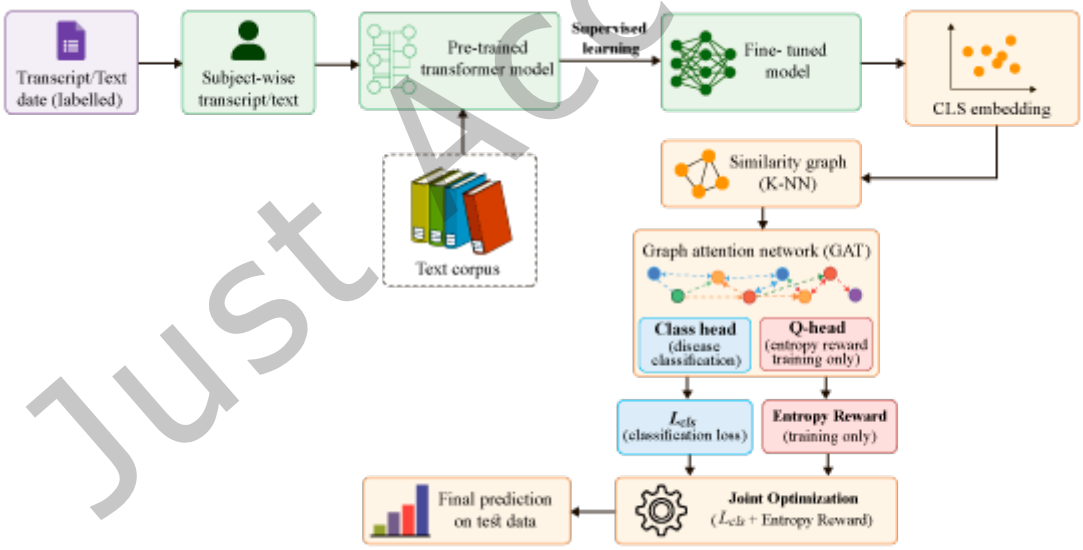

Fig. 2. Overview of the proposed GNN-BD framework for transcript-based brain disease classification.

Just Accepted

4.1 Part I: Subject Similarity Graph Construction

4.1.1 Subject-Level Transcript Aggregation. For each subject i, all utterances from the Cookie Theft task are concatenated into a single transcript:

$$C_i = \text{concat}\big(\text{utterances of subject } i\big). \tag{1}$$

This aggregation reflects the clinical nature of the task, where each participant produces a single spontaneous description rather than a multi-turn dialogue.

4.1.2 Task-Adapted Transformer Embedding. A pretrained transformer language model is fine-tuned for subject-level brain disease classification. After fine-tuning, a fixed-length representation is obtained by extracting the final-layer `[CLS]` embedding:

$$\mathbf{X}_i = \mathbf{E}_i^{\text{CLS}} \in \mathbb{R}^d, \tag{2}$$

where d denotes the hidden dimension of the transformer. Since the encoder is optimized for the diagnostic task, the resulting embeddings are task-aware and encode clinically relevant lexical, syntactic, and discourse-level information.

Each subject is represented as a graph node:

$$v_i = (\mathbf{X}_i, y_i), \quad y_i \in \{0, 1\}, \tag{3}$$

where y_i denotes the diagnostic label.

4.1.3 k-Nearest Neighbor Graph Construction. To model latent inter-subject relationships, a similarity graph is constructed using cosine similarity between subject embeddings:

$$\text{Sim}(\mathbf{X}_i, \mathbf{X}_j) = \frac{\mathbf{X}_i \cdot \mathbf{X}_j}{\|\mathbf{X}_i\| \, \|\mathbf{X}_j\|}. \tag{4}$$

Prior to graph construction, the embeddings are standardized using statistics computed on the training set and linearly projected to a lower-dimensional space as a preprocessing step to improve numerical stability. For each training node v_i, edges are added to its k most similar training neighbors:

$$E_{\text{train}} = \{(v_i, v_j) \mid v_j \in \text{kNN}(v_i)\}. \tag{5}$$

This yields a sparse, undirected subject-level graph:

$$G = (V, E_{\text{train}}, X), \tag{6}$$

where V denotes the set of subjects and X the matrix of node features. Edges are computed exclusively among training subjects to prevent information leakage.

4.1.4 Leakage-Free Inference Graph. During validation and testing, each test subject is added to the training graph by connecting it only to its k nearest training neighbors. No edges are created between test subjects, ensuring strict subject-level isolation during evaluation.

4.2 Part II: Entropy-Guided Graph Neural Network Learning

4.2.1 Graph Attention Network Encoder. Given the subject similarity graph G, node features are updated using a Graph Attention Network (GAT). For each node v_i, the GAT produces a latent representation:

$$\mathbf{h}_i = \text{GAT}(\mathbf{X}_i, \mathcal{N}(i)), \tag{7}$$

where $\mathcal{N}(i)$ denotes the neighborhood of node i. Residual connections are employed to preserve the original linguistic representations and improve training stability.

Just Accepted

4.2.2 Dual-Head Prediction. Each node representation $\mathbf{h}_i$ is passed to two output heads:

$$\hat{y}_i = f_{\text{cls}}(\mathbf{h}_i), \quad Q_i = f_Q(\mathbf{h}_i), \tag{8}$$

where f_{cls} denotes the softmax classification head and f_Q denotes a scalar value estimator.

4.2.3 Entropy-Guided Reward Formulation. To encourage confident and reliable predictions, an entropy-based reward is defined at the node level. Let

$$\mathbf{p}_i = \text{softmax}(\hat{y}_i) \tag{9}$$

denote the predicted class probability distribution, and

$$H(\mathbf{p}_i) = -\sum_c p_{i,c} \log p_{i,c} \tag{10}$$

denote its entropy. The reward is computed as:

$$r_i = (2 \cdot \mathbb{I}[\hat{y}_i = y_i] - 1)\left(1 - \frac{H(\mathbf{p}_i)}{\log C}\right), \tag{11}$$

where C is the number of classes.

4.2.4 Value Learning and Joint optimisation. The proposed framework does not employ temporal-difference learning or Bellman updates. Instead, the value head (Q-head) is trained to directly regress an entropy-guided reward computed from the model's current predictions. The reward reflects both classification correctness and prediction confidence and is defined at each training iteration without state transitions.

The value loss is formulated as a regression objective:

$$\mathcal{L}_Q = \|Q_i - r_i\|_1, \tag{12}$$

where r_i denotes the entropy-guided reward and $\|\cdot\|_1$ represents the smooth L1 loss.

The supervised classification loss is defined as:

$$\mathcal{L}_{\text{cls}} = \text{CE}(\hat{y}_i, y_i). \tag{13}$$

The final training objective jointly optimizes both components:

$$\mathcal{L} = \mathcal{L}_{\text{cls}} + \alpha \mathcal{L}_Q, \tag{14}$$

where α controls the influence of the reinforcement-style regulariation. This joint optimisation shapes the node embedding space while keeping the graph topology fixed.

5 Dataset Description

In this study, we have selected three distinct datasets representing different brain diseases, as shown in Table 4. These datasets include textual data and are chosen to analyse language patterns and cognitive impairments associated with various neurological conditions.

- **Dementia (English Pitt Corpus) - TalkBank**:This dataset is sourced from the TalkBank repository and consists of transcribed speech data from individuals diagnosed with dementia. The English Pitt Corpus contains recorded conversations that are transcribed into text, enabling the analysis of linguistic features associated with dementia. We extract the text transcripts from this dataset. It is labelled into two categories: *Dementia* and *Healthy*.
  When we downloaded the dataset, each diagnostic category contained four tasks: cookie, recall, frequency, and sentence. We selected only the cookie task and discarded the others. The justification for selecting only the cookie task is that if we included all task and then there is very less data for other task in control patient than dementia. When data are aggregated per subject, the total transcript length for

Just Accepted

dementia patients is often longer than that of control subjects, which can artificially inflate classification accuracy without truly capturing underlying linguistic patterns. For this reason, we restrict the dementia experiments to the Cookie Theft task only. The cookie task subset contains 99 subjects, with a total of 243 transcripts in the control group and 192 subjects (306 transcripts) in the non-control group. Some subjects have multiple transcripts. After excluding mild cognitive impairment (MCI) cases from non-control group, the final dementia cohort contained 175 unique patients.

- **Toddler Autism Dataset**: This dataset contains textual data along with behavioral indicators used for diagnosing autism in toddlers. It includes features related to social interactions, communication skills, and repetitive behaviours, which are crucial for identifying autism spectrum disorder (ASD). We focus on the text-based attributes for analysis. It is not transcript. The dataset has two labels: *Autism* and *Non-Autism*.
- **Right hemorrhagic stroke dataset - TalkBank**: This dataset, also from TalkBank, includes transcribed speech samples from individuals who have experienced a right-hemisphere stroke. The dataset provides insights into the linguistic impairments resulting from brain damage in the right hemisphere. We extract the text transcripts for analysis. It is categorized into two labels: *Stroke* and *Healthy*. There are also two tasks for each stroke and healthy participant. These tasks are balanced in quantity; therefore, we selected both tasks for our dataset.

Each dataset was carefully selected to ensure a diverse representation of neurological disorders affecting speech and language. The primary focus of this study is to analyse these text-based datasets to identify distinct linguistic patterns associated with each condition.

Table 4. Summary of datasets used in this study

Dataset Name	unique patient/subject	Dataset Link
Dementia (English Pitt Corpus) - TalkBank	Dementia:175, Healthy:99	talkbank.org/.../Pitt.html
Toddler Autism Dataset	Autism:174,Non-Autism:158	archive.ics.uci.edu/.../Autism+Screening
Right Hemorrhagic Stroke Dataset - TalkBank	Stroke:60, Healthy:30	talkbank.org/rhd

6 Experimental Setup

In our experiment with the dementia dataset, we initially split the data using an 80:20 ratio for training and testing. The training set was then further divided into a 90:10 split for training and validation. The validation set was used to identify the optimal parameters and to select the appropriate number of epochs.

Preprocessing

We restrict our analysis to the *Cookie Theft* picture description task in order to maintain a consistent elicitation context across subjects and to reduce variability introduced by heterogeneous speech tasks. Diagnoses related to mild cognitive impairment (MCI) are excluded to avoid label ambiguity, and only the selected diagnostic groups are retained. Only participant speech (`*PAR:` lines) is included in the analysis to prevent contamination from examiner language. The extracted text is then cleaned using a dementia-aware normalization procedure that removes CHAT-specific artifacts, including timestamps, non-speech event markers (e.g., coughing or laughter), and bracketed or parenthetical annotations. At the same time, filled pauses are preserved by normalizing forms such as `&-uh` and `&-um` to their lexical equivalents (`uh`, `um`), as disfluencies are known to carry clinically relevant

Just Accepted

information in neurodegenerative disease assessment. After text preprocessing, each transcript is encoded using a fine-tuned Roberta model, and the resulting utterance-level representations are aggregated into a single subject-level embedding. These embeddings are standardized using z-score normalization (mean zero, unit variance), where the scaling parameters are estimated on the training set only and subsequently applied to validation and test sets to prevent information leakage. For models requiring dimensionality reduction or structured inductive bias (e.g., graph neural networks), a trainable linear projection is applied to map the high-dimensional Roberta embeddings into a lower-dimensional latent space prior to downstream classification.[2]

Finally, all cleaned utterances belonging to the same subject are concatenated into a single subject-level transcript, yielding one sample per subject. This subject-level aggregation supports a leakage-free experimental design and enables reliable downstream transformer-based representation learning and graph neural network modelling.

Handling Class Imbalance

To address class imbalance, we adopt a class-weighted loss strategy during the transformer fine-tuning stage, where class weights are computed exclusively from the training subjects based on inverse class frequencies and incorporated into the cross-entropy loss. This approach increases the influence of underrepresented classes without modifying the data distribution. In addition, the dataset is split using stratified subject-level sampling, ensuring that the training, validation, and test sets maintain similar class distributions. The validation set is drawn only from the training subjects and preserves the same class ratio, while the held-out test set follows the same stratified distribution as the training data. In the graph neural network stage, class imbalance is further mitigated implicitly through entropy-guided reinforcement regularization, which emphasizes confident and correct predictions while reducing the impact of uncertain samples. All imbalance-handling mechanisms rely solely on training data, preserving a leakage-free and statistically consistent evaluation protocol.
Depending on the validation data we find the hyper-parameter. The following combinations were explored:

- `lr` $\in \{1 \times 10^{-5}, 2 \times 10^{-5}, 3 \times 10^{-5}\}$
- `gnn_hidden_dim` $\in \{128, 256\}$
- `gnn_k_neighbors` $\in \{5, 10, 15\}$
- `gnn_num_layers` $\in \{2, 3\}$

The best-performing hyperparameters were selected based on the with accuracy. The final configuration used for training on the dementia dataset is presented in Table 5.

Table 5. Optimal hyper-parameters for dementia dataset

Parameter	Value
Learning Rate (`lr`)	1×10^{-5}
GNN Hidden Dimension (`gnn_hidden_dim`)	128
k-Neighbors (`gnn_k_neighbors`)	5
Number of GNN Layers (`gnn_num_layers`)	2

To identify the optimal configuration for the proposed model, we conducted a series of hyper-parameter tuning experiments using five fold cross-validation. Each trial explored a different combination of key parameters including the Roberta fine-tuning learning rate, the GNN hidden dimension size, the number of neighbors

[2]The preprocessing strategy was adapted (not fully) from prior study [34], and their successful fine-tuning of RoBERTa further encouraged our methodological design.

Just Accepted

used for graph construction (k-NN), and the number of GNN layers. Table 6 summarises the performance of selected trials on validation. We observed that while multiple configurations produced competitive results, the best overall performance was achieved with `lr`=1×10^{-5}, `gnn_hidden_dim`=128, `gnn_k_neighbors`=5, and `gnn_num_layers`=2, yielding an out-of-fold accuracy of 0.8942, precision of 0.9451, recall of 0.8857, F1 score of 0.9145, and ROC-AUC of 0.9091.

Table 6. Performance metrics for different hyperparameter combinations

Hyperparameter combination	Accuracy	Precision	Recall	F1 Score	ROC-AUC
LR=1×10^{-5}, Hidden=128, k=5, Layers=2	0.8942	0.9451	0.8857	0.9145	0.9091
LR=2×10^{-5}, Hidden=256, k=10, Layers=2	0.8796	0.8944	0.9200	0.9070	0.9113
LR=3×10^{-5}, Hidden= 256, k= 10, Layers=3	0.883212	0.903955	0.914286	0.909091	0.894228

6.1 Experimental result

The dataset was divided into an 80:20 ratio for all three datasets: the dementia dataset, the autism dataset (AD), and the RHD dataset. The same algorithm and the same parameter settings, as identified in the previous section, were used for all three conditions. For each disease, 10% of the training data was further reserved for epoch selection, which was performed exclusively on the training set. We evaluated several pre-trained transformer-based embedding models to examine their impact on classification performance. The compared models include Roberta-large, Roberta-base, BERT-large, and BERT-base. Table 7 summarises the performance across folds in terms of Accuracy, Precision, Recall, and F1 score, while Figure 3 provides a visual comparison of the mean metric values.

Overall, Roberta-large demonstrates the strongest performance across most evaluation metrics. It achieves the highest accuracy (0.9273), precision (0.9697), and F1 score (0.9412), indicating superior overall discrimination capability and balanced performance. Its recall (0.9143) is also competitive, remaining close to the best-performing models. These results suggest that the higher representational capacity of Roberta-large is beneficial for capturing complex linguistic patterns in the dataset.

Roberta-base emerges as a close second, particularly excelling in precision (0.9677) while maintaining comparable accuracy (0.8909) and F1 score (0.9091). This indicates that RoBERTa-base is especially effective at making reliable positive-case predictions, while offering substantially lower computational cost compared with large-scale models.

BERT-large shows stable but slightly weaker performance, with an F1 score of 0.8857 and lower precision than both Roberta-large and BERT-base. Roberta-base, although achieving reasonable results, consistently underperforms its larger counterpart, particularly in recall and F1 score, suggesting limited benefit from the base configuration for this task.

Table 7. Performance metrics for dementia dataset

Model	Accuracy	Precision	Recall	F1 Score	ROC-AUC
Roberta-large	0.9273	0.9697	0.9143	0.9412	0.9343
Roberta-base	0.8909	0.9677	0.8571	0.9091	0.9443
BERT-large	0.8545	0.8857	0.8857	0.8857	0.8829
BERT-base	0.8727	0.9118	0.8857	0.8986	0.9457

Just Accepted

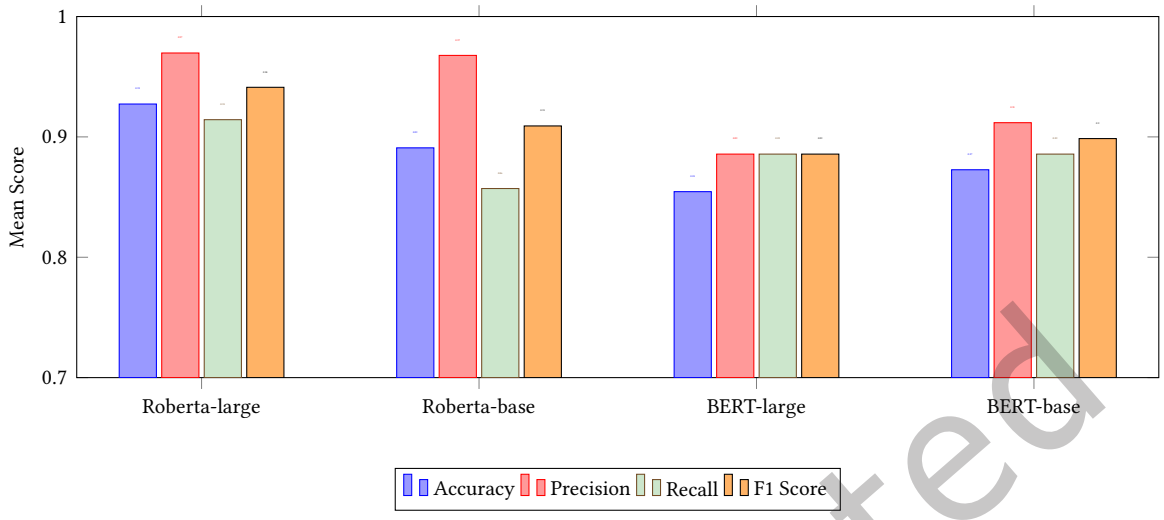

Fig. 3. Metric values across 80:20 distribution for dementia dataset

Table 8. Ablation study on the Cookie Theft task (subject-level 80/20 split). Roberta-large is trained end-to-end using full-dimensional representations. All other methods operate on projected Roberta CLS embeddings (1024→500). The proposed method (GNN+Q+Roberta) is highlighted.

Model	Acc	Prec	Recall	F1	ROC-AUC	Advantage
Roberta-large (full dim., end-to-end)	0.9091	0.9412	0.9143	0.9275	0.9357	Strong text-only baseline
Roberta + LR (projected CLS)	0.8727	0.8889	0.9143	0.9014	0.8971	Simple linear decision boundary
Roberta + MLP (projected CLS)	0.8545	0.8857	0.8857	0.8857	0.9186	Nonlinear feature mixing
Roberta + RF (projected CLS)	0.8727	0.8889	0.9143	0.9014	0.9329	Temporal aggregation of chunked embeddings
Roberta + GNN (projected CLS)	0.8364	0.8611	0.8857	0.8732	0.9257	Captures subject similarity graph
Roberta + Q-head (fine-tuning stage)	0.8727	0.9667	0.8286	0.8923	0.9543	Entropy-guided fine-tuning without graph structure
Roberta + GNN + Q (Proposed)†	**0.9273**	**0.9697**	0.9143	**0.9412**	0.9343	**Confidence-aware graph learning with entropy-guided Q-updates**

†Q denotes entropy-based value regression with reward-driven loss reweighting.

Ablation Analysis: Table 8 presents an ablation study comparing a full end-to-end Roberta-large classifier with several projection-based downstream models operating on Roberta CLS embeddings. The end-to-end Roberta-large baseline leverages the full 1024-dimensional representation and achieves strong performance, highlighting the effectiveness of large pretrained language models for dementia detection from spontaneous speech.

When the CLS embeddings are projected to a lower-dimensional space and used with standard classifiers ((LInear regression (LR), Multi layer perceptron (MLP), Random forest(RF)), performance remains competitive but generally decreases in accuracy and F1 score, indicating that dimensionality reduction alone does not fully preserve all discriminative information. Introducing graph-based learning on projected embeddings (Roberta+GNN) captures inter-subject similarity structure but results in reduced accuracy, likely due to noisy neighborhood aggregation and over-smoothing effects in the similarity graph. This observation indicates that naive graph propagation is insufficient when subject relationships are imperfect or heterogeneous. We also tested applying the Q-head directly on top of RoBERTa, without using any graph (Roberta+Q-head). The results show that the entropy reward still helps at the transformer stage, achieving a precision of 0.9667 and ROC-AUC of 0.9543. However, its recall (0.8286) and F1 score (0.8923) remain notably lower than the proposed model (0.9143 and 0.9412, respectively),

Just Accepted

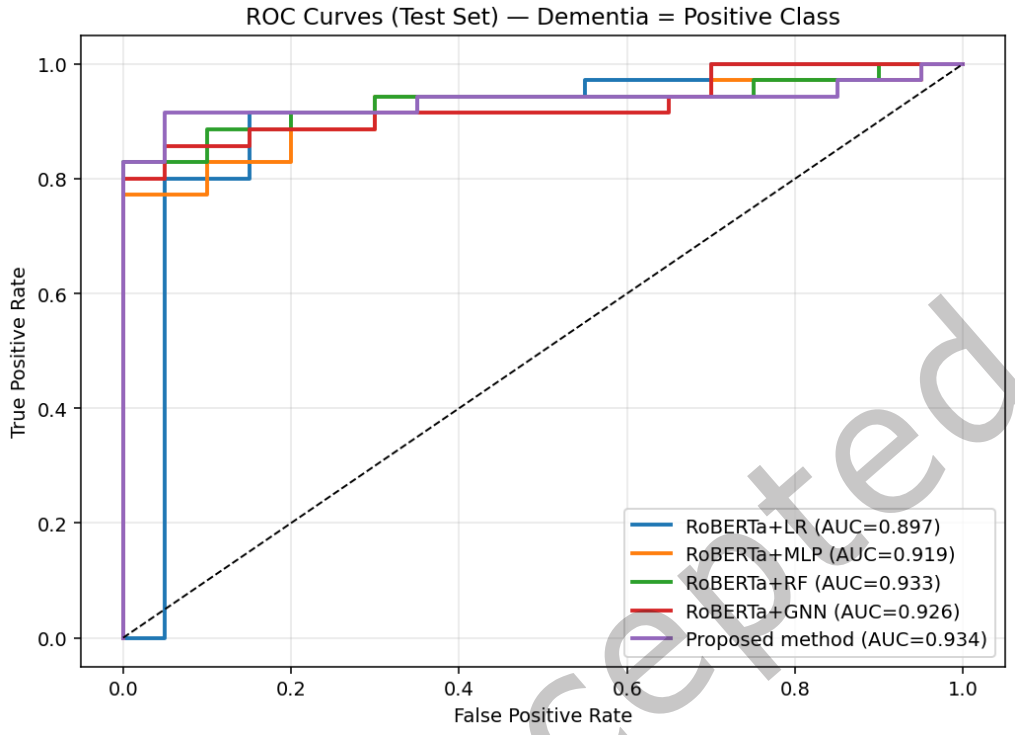

Fig. 4. Receiver operating characteristic (ROC) curves on the test set for dementia detection.

Just Accepted

suggesting that the graph structure captures complementary inter-subject relationships that the entropy reward alone cannot replicate.

The proposed Roberta+GNN+Q model substantially improves upon the plain GNN by incorporating an entropy-based Q-learning mechanism that jointly reweights the classification loss and learns a confidence-aware value function. By assigning higher importance to confident and correctly predicted nodes while down-weighting uncertain or misleading ones, the model selectively exploits reliable graph information and mitigates the impact of noisy neighbors. As a result, the proposed method achieves the highest accuracy and F1 score among all evaluated approaches, surpassing even the full end-to-end Roberta-large baseline while maintaining a comparable ROC-AUC. These results demonstrate that confidence-guided graph learning provides a complementary inductive bias that enhances subject-level classification beyond text-only modelling.

Figure 4 shows the ROC curves on the test set, where all Roberta-based models achieve competitive performance. While overall AUC values are comparable, the proposed method consistently attains higher sensitivity at low false-positive rates, which is clinically important for dementia screening where minimizing false alarms is critical.

Autism and RHD detection: In addition to dementia detection, we evaluate the generalization ability of the proposed model on two other brain-related disorder datasets: Autism and Right Hemisphere Damage (RHD). For all experiments, we employ the same model architecture and hyper-parameters described in the methodology section to ensure a fair comparison across tasks.

Table 9. Classification performance of the proposed model across different brain-related disorders.

Dataset	Accuracy	Precision	Recall	F1	ROC-AUC
Dementia (80:20)	0.9273	0.9697	0.9143	0.9412	0.9343
Dementia (10 fold cv)	0.9017 ± 0.0639	0.9204 ± 0.0455	0.9261 ± 0.0707	0.9223 ± 0.0518	0.9236 ± 0.0645
Autism	0.9104	0.9143	0.9143	0.9143	0.9464
Autism (10 fold cv)	0.9728 ± 0.0286	0.9775 ± 0.0372	0.9712 ± 0.0288	0.9739 ± 0.0272	0.9826 ± 0.0181
RHD	0.8125	0.8571	0.75	0.8	0.9531
RHD(10 fold CV)	0.8839 ± 0.0936	0.9267 ± 0.1474	0.8750 ± 0.1250	0.8836 ± 0.0831	0.9125 ± 0.0935

For the Autism and RHD datasets, 80% of the data is used for training, while the remaining 20% is reserved for testing. Table 9 reports the classification performance across the three disorders. The model achieves its highest 80:20 accuracy and F1-score on the Dementia dataset, while Autism and RHD show higher ROC-AUC values.

Performance on the Autism dataset remains competitive, yielding balanced precision and recall (both 0.9143) and a high ROC-AUC of 0.9464, indicating reliable discrimination despite the heterogeneity of autism-related language patterns. In contrast, performance on the RHD dataset is comparatively lower, particularly in recall(0.7500).

Overall, these results demonstrate that the proposed model generalizes well across multiple neurological and developmental disorders, while performance varies depending on dataset characteristics and linguistic variability.

We further evaluate the proposed model using 10-fold stratified cross-validation, reporting the mean ± standard deviation to assess robustness and stability across different subject splits.

The cross-validation results in Table 9 show that the proposed model maintains stable performance across different data splits, with relatively small standard deviations[3]. This consistency highlights the robustness of the graph-based learning framework and its ability to generalise beyond a single train–test split, even in low-resource settings such as RHD detection.

6.1.1 Comparative analysis. Table 10 presents a comparative evaluation of dementia detection models using Accuracy, Precision, Recall, and F1-score. The proposed framework demonstrates competitive performance across all metrics and achieves one of the highest F1-scores among graph- and transformer-based approaches. While Wen et al. [57] report a slightly higher F1-score (95.5%), their method relies on handcrafted part-of-speech features and assumes subject-independent samples, which may introduce information leakage when multiple transcripts from the same subject are present. Another study by Liu et al. [33] does not explicitly address subject-independent evaluation. The authors report results based on the total number of transcripts rather than the number of subjects, making it unclear whether multiple transcripts from the same subject may appear across training and testing splits.

In contrast, the proposed model operates at the subject level, constructs an explicit similarity graph, and enforces strict subject-level separation during training and evaluation. Other methods [6, 34, 47] achieve lower overall performance or rely on isolated document representations without modelling inter-subject relationships. These results highlight the benefit of incorporating subject similarity graphs and confidence-aware graph learning for robust dementia detection. Table 11 compares the proposed GNN-BD framework with the AD-GNN model by Cai et al. [7]. While both approaches employ graph neural networks for dementia detection, they differ substantially in design philosophy and computational requirements. AD-GNN constructs token-level graphs from speech transcripts and incorporates extensive data augmentation and multimodal audio processing. In contrast, GNN-BD operates solely on text transcripts and builds a subject-level similarity graph, resulting in a significantly lighter and more scalable pipeline.

[3]The RHD results show higher variability, likely due to the smaller dataset size.

Just Accepted

Table 10. Comparison of performance metrics across various dementia detection models.

Model	Accuracy (%)	Precision (%)	Recall (%)	F1-score (%)
[6]	81.54	87.34	83.13	85.19
[34]	90.16	92.81	88.60	90.60
[57]	92.20	93.50	97.10	95.50
[33]	93.50	94.00	89.00	91.19
[47]	88.08	90.50	84.34	87.23
Proposed (80:20)	**92.73**	**96.97**	**91.43**	**94.12**

Table 11. Comparison between the proposed GNN-BD and AD-GNN [7] for dementia detection

Category	Proposed GNN-BD	AD-GNN [7]	Advantage of Proposed Method
Scope of Application	Multi-disease (Dementia, Autism, RHD)	Single-disease (Dementia only)	Broader applicability
Data Modality	Text-only transcripts	Text + Audio (WavLM, TTS)	Lower computational cost
Graph Construction	Subject-level similarity graph (kNN)	Token-level dependency / dynamic graphs	Simpler and interpretable
Graph Optimisation	Fixed graph; entropy-guided optimisation	Static or dynamic graph learning (IDGL)	Stable training, no graph reconstruction
Model Architecture	GAT with value regularisation	GraphSAGE / GGNN	Lightweight yet effective
Performance	Competitive accuracy on dementia data (92.73)	Reported accuracy 84.84 (best accuracy)%	Comparable or improved performance
Reinforcement Learning	Entropy-guided value regression	Not used	Confidence-aware optimisation
Data Augmentation	Not required	Extensive augmentation (GPT, MLM, etc.)	Reduced pipeline complexity
Computational Efficiency	High (text-only pipeline)	Lower (audio + augmentation)	Practical deployment

Just Accepted

Moreover, GNN-BD introduces an entropy-guided optimisation mechanism that enhances prediction confidence without modifying graph topology or relying on dynamic graph reconstruction. Despite its simpler design, GNN-BD achieves competitive performance on dementia detection while offering improved generalisability to multiple neurological conditions and greater suitability for low-resource clinical settings.

Autism detection: We compare our method with Pal et al. [40] for autism detection, as shown in Table 12.

Although both studies utilize the TASD dataset, the proposed method and that of Pal et al. [40] differ substantially in objective and modelling strategy. Pal et al. focus on fine-grained multi-class prediction of ten behavioral traits using a single BERT-based classifier to support therapy recommendation. In contrast, our work addresses binary autism detection within a unified graph-based framework that also extends to other neurological disorders. Rather than relying on standalone sequence classification, we construct a subject similarity graph over text embeddings and apply a Graph Neural Network augmented with entropy-guided value regulariation. This design improves

Table 12. Comparison between the proposed GNN-BD framework and the autism detection model by Pal et al. [40]

Category	**Pal et al. [40] (2024)**	**Proposed GNN-BD**
Dataset	TASD (Toddler Autism Screening Dataset)	TASD (evaluated within a unified brain disease detection framework)
Detection Focus	Multi-class behavioral trait prediction (10 traits)	Binary autism vs. non-autism classification
Input Modality	Textual behavioral descriptions	Textual behavioral descriptions
Model Backbone	BERT (fine-tuned)	Roberta-based contextual embeddings
Architecture	Single sequence classifier	Graph Attention Network over subject similarity graph
Graph Construction	Not applicable	k-nearest neighbor graph based on cosine similarity of text embeddings
Graph optimisation	Not applicable	Fixed graph; entropy-guided value regulariation during training
Reinforcement Learning	Not used	Entropy-guided value regression (no policy learning or graph modification)
Performance	Accuracy: 0.88	Accuracy: 0.9104, F1-score: 0.9143 (80:20 setting)
Primary Objective	Behavioral assessment and therapy recommendation	Unified detection of multiple neurological disorders
Computational Cost	Moderate (single-model inference)	Low (text-only, no data augmentation or audio processing)

Just Accepted

classification robustness without requiring additional modalities or data augmentation, making it suitable for low-resource clinical NLP settings.

Given that text-based methods for detecting ASD or non-ASD using the TASD dataset remain underexplored in the literature, our work contributes a novel direction by introducing graph neural networks within this context.

RHD detection: We compare our proposed text-based graph neural network framework with the work of Cui et al. [11], who investigated right hemisphere damage (RHD) using a classical machine learning pipeline that integrates both acoustic and textual features. Their study reports an accuracy of 93.91% on the RHDBank dataset for distinguishing patients from non-patients.

In contrast, our approach operates exclusively on textual transcripts and performs subject-level classification using sentence embeddings and graph-based learning. Under a strict subject-independent evaluation protocol, our method achieves an accuracy of 81.25% on RHD detection (Table 13). While this accuracy is lower than that reported by Cui et al., it reflects a more conservative and clinically realistic setting that avoids clip-level inflation and does not rely on acoustic cues.

Importantly, Cui et al. segment speech into multiple short samples per subject and incorporate handcrafted acoustic features, which substantially increases the number of training instances and simplifies the classification task. Our framework instead models each subject as a single node in a semantic similarity graph, capturing higher-level discourse and pragmatic patterns associated with RHD using text alone. This design emphasizes scalability, interpretability, and applicability in scenarios where clean audio data is unavailable or impractical to obtain.

Table 13. Comparison of RHD detection approaches between Cui et al. [11] and the proposed method

Method	Accuracy (%)	Input Modality	Key Characteristics
Cui et al. [11]	93.91	Text + Audio	Classical ML models (e.g., SVM, RF); handcrafted lexical and acoustic features; clip-level evaluation
Proposed method	81.25 (80:20 setting)	Text-only	Sentence embeddings from transcripts; subject-level semantic graph construction; GNN with entropy-guided Q-regulariation

7 Data Leakage Prevention

To ensure a fair and reliable evaluation, the proposed framework is designed to prevent data leakage at all stages of training and inference. All speech transcripts are first aggregated at the subject level, such that each subject contributes a single sample regardless of the number of recordings. Dataset splitting is then performed exclusively at the subject level using a stratified 80/20 train–test split, ensuring that no subject appears in more than one partition. The validation set is further drawn only from the training subjects and is never used during model fitting or graph construction for the test set.

Graph construction is also performed in a leakage-free manner. The subject similarity graph used for training is built solely from training subjects, and cosine similarity is computed only among training embeddings. During validation and testing, each target subject is evaluated using an inference-time graph that consists of all training nodes plus a single test node, which is connected only to its k nearest training neighbors. Importantly, test subjects are never connected to one another, and their embeddings are not used to update the training graph topology or model parameters. Reinforcement-style regulariation is applied strictly during training and depends only on predictions and labels from training nodes. As a result, the proposed framework maintains strict isolation between training, validation, and test data, guaranteeing that reported performance reflects genuine generalization rather than information leakage.

8 Limitations

While the proposed methodology shows strong performance across the three datasets, several limitations should be acknowledged.

- **ASR dependency.** Although our models operate on text transcripts, these transcripts are still derived from audio recordings. Automatic Speech Recognition (ASR) errors which is caused by noise, atypical speech patterns, or low-quality recordings that can propagate into the text input and affect classification. Therefore, the approach is not entirely free from audio-related challenges, particularly in uncontrolled environments.

Just Accepted

- **Separate disease-specific models.** We apply the same methodology to dementia, autism, and right-hemisphere damage datasets, but we train *separate* models for each condition. Thus, the framework is generalisable, but it does not represent a single unified model capable of detecting multiple disorders simultaneously.
- **Dataset size and imbalance.** All datasets are relatively small and imbalanced, which may limit generalisability to broader clinical populations, diverse dialects, and real-world conversational settings. Furthermore, given the relatively small dataset sizes, there is a risk of overfitting, particularly for the dementia and RHD datasets. The use of cross-validation and subject-level stratified splitting partially mitigates this risk, but results should be interpreted with caution when generalising to larger, more diverse clinical populations.
- **Value regulariation complexity and explainability** The value regulariation component improves the graph structure but introduces additional hyper-parameter sensitivity and computational cost. This may present challenges when deploying the model in real-time or resource-limited clinical scenarios. the current study focuses primarily on performance-based evaluation and does not include an extensive comparative analysis against all recent large-scale language models. In addition, while the graph structure improves relational modelling, deeper explainability of individual decision pathways remains an important direction for future investigation. Future work will therefore explore multimodal extensions, improved interpretability mechanisms, and real-time diagnostic settings, particularly in healthcare environments with limited linguistic and computational resources. It should be noted that the current interpretability is limited to confidence-aware prediction weighting through the entropy-guided mechanism. Feature-level or clinically grounded interpretability, such as identifying which specific linguistic features drive individual predictions, is not yet provided and remains an important direction for future work.

9 Conclusion

This research introduces GNN-BD, a transcript-based Graph Neural Network framework for brain disease classification that integrates semantic graph construction with an entropy-guided, reinforcement-style optimisation mechanism. The proposed model achieves strong performance across multiple clinical datasets, including dementia, autism, and right hemisphere damage, while maintaining favorable generalisability and computational efficiency. Rather than performing sequential reinforcement learning or modifying graph topology, the optimisation strategy acts as a lightweight regulariation that shapes node representations during training. By operating directly on text transcripts, GNN-BD avoids complex audio processing pipelines, offering a scalable and clinically accessible alternative to prior approaches. Future work will explore extensions to multimodal data and real-time diagnostic settings, particularly in healthcare environments with limited linguistic and computational resources.

Just Accepted

References

[1] Samad Amini, Boran Hao, Lifu Zhang, Mengting Song, Aman Gupta, Cody Karjadi, Vijaya B Kolachalama, Rhoda Au, and Ioannis Ch Paschalidis. Automated detection of mild cognitive impairment and dementia from voice recordings: a natural language processing approach. *Alzheimer's & Dementia*, 19(3):946–955, 2023.

[2] Meysam Asgari, Liu Chen, and Hiroko Dodge. Topic-based measures of conversation for detecting mild cognitiveimpairment. In *Proceedings of the First Workshop on Natural Language Processing for Medical Conversations*, pages 63–67, 2020.

[3] Flavio Bertini, Davide Allevi, Gianluca Lutero, Laura Calzà, and Danilo Montesi. An automatic alzheimer's disease classifier based on spontaneous spoken english. *Computer Speech & Language*, 72:101298, 2022.

[4] Laiss Bertola, Natália B Mota, Mauro Copelli, Thiago Rivero, Breno Satler Diniz, Marco A Romano-Silva, Sidarta Ribeiro, and Leandro F Malloy-Diniz. Graph analysis of verbal fluency test discriminate between patients with alzheimer's disease, mild cognitive impairment and normal elderly controls. *Frontiers in aging neuroscience*, 6:185, 2014.

[5] Margaret Lehman Blake. Perspectives on treatment for communication deficits associated with right hemisphere brain damage. *Clinical Focus*, 2007.

[6] Mondher Bouazizi, Chuheng Zheng, and Tomoaki Ohtsuki. Dementia detection using language models and transfer learning. In *Proceedings of the 2022 5th International Conference on Software Engineering and Information Management*, pages 152–157, 2022.
[7] Hongmin Cai, Xiaoke Huang, Zhengliang Liu, Wenxiong Liao, Haixing Dai, Zihao Wu, Dajiang Zhu, Hui Ren, Quanzheng Li, Tianming Liu, et al. Exploring multimodal approaches for alzheimer's disease detection using patient speech transcript and audio data. *arXiv preprint arXiv:2307.02514*, 2023.
[8] Rupayan Chakraborty, Meghna Pandharipande, Chitralekha Bhat, and Sunil Kumar Kopparapu. Identification of dementia using audio biomarkers. *arXiv preprint arXiv:2002.12788*, 2020.
[9] Michail Chatzianastasis, Loukas Ilias, Dimitris Askounis, and Michalis Vazirgiannis. Neural architecture search with multimodal fusion methods for diagnosing dementia. In *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5. IEEE, 2023.
[10] Lei Chi, Arav Sharma, Ari Gebhardt, and Joseph Colonel. Predicting cognitive decline: A multimodal ai approach to dementia screening from speech. In *2025 IEEE International Conference on AI and Data Analytics (ICAD)*, pages 1–8. IEEE, 2025.
[11] Xia Cui, Amila Gamage, Terry Hanley, and Tingting Mu. Identifying indicators of vulnerability from short speech segments using acoustic and textual features. *Proceedings of Interspeech 2021*, pages 1569–1573, 2021.
[12] E Ron De Kloet, Marian Joëls, and Florian Holsboer. Stress and the brain: from adaptation to disease. *Nature reviews neuroscience*, 6(6):463–475, 2005.
[13] Michael D Devous. Functional brain imaging in the dementias: role in early detection, differential diagnosis, and longitudinal studies. *European Journal of Nuclear Medicine and Molecular Imaging*, 29(12):1685–1696, 2002.
[14] Francesco Di Gregorio, Giada Lullini, Silvia Orlandi, Valeria Petrone, Enrico Ferrucci, Emanuela Casanova, Vincenzo Romei, and Fabio La Porta. Clinical and neurophysiological predictors of the functional outcome in right-hemisphere stroke. *NeuroImage*, 308:121059, 2025.
[15] Johnny Downs, Sumithra Velupillai, Gkotsis George, Rachel Holden, Maxim Kikoler, Harry Dean, Andrea Fernandes, and Rina Dutta. Detection of suicidality in adolescents with autism spectrum disorders: developing a natural language processing approach for use in electronic health records. In *AMIA annual symposium proceedings*, volume 2017, page 641, 2018.
[16] Mahmoud Elbattah, Jean-Luc Guérin, Romuald Carette, Federica Cilia, and Gilles Dequen. Nlp-based approach to detect autism spectrum disorder in saccadic eye movement. In *2020 IEEE Symposium Series on Computational Intelligence (SSCI)*, pages 1581–1587. IEEE, 2020.
[17] Azra Emekci. Predictive modeling for early alzheimer's disease using natural language processing. In *2024 IEEE International Conference on Big Data (BigData)*, pages 4632–4636. IEEE, 2024.
[18] Muhammad Shoaib Farooq, Rabia Tehseen, Maidah Sabir, and Zabihullah Atal. Detection of autism spectrum disorder (asd) in children and adults using machine learning. *scientific reports*, 13(1):9605, 2023.
[19] Kathleen C Fraser, Jed A Meltzer, and Frank Rudzicz. Linguistic features identify alzheimer's disease in narrative speech. *Journal of Alzheimer's disease*, 49(2):407–422, 2015.
[20] Gloria Gagliardi and Fabio Tamburini. Linguistic biomarkers for the detection of mild cognitive impairment. *Lingue e linguaggio*, 20(1):3–31, 2021.
[21] Romane Gauriau, Bernardo C Bizzo, Felipe C Kitamura, Osvaldo Landi Junior, Suely F Ferraciolli, Fabiola BC Macruz, Tiago A Sanchez, Marcio RT Garcia, Leonardo M Vedolin, Romeu C Domingues, et al. A deep learning–based model for detecting abnormalities on brain mr images for triaging: preliminary results from a multisite experience. *Radiology: Artificial Intelligence*, 3(4):e200184, 2021.
[22] Fasih Haider, Sofia De La Fuente, and Saturnino Luz. An assessment of paralinguistic acoustic features for detection of alzheimer's dementia in spontaneous speech. *IEEE Journal of Selected Topics in Signal Processing*, 14(2):272–281, 2019.
[23] Loukas Ilias and Dimitris Askounis. Multimodal deep learning models for detecting dementia from speech and transcripts. *Frontiers in aging neuroscience*, 14:830943, 2022.
[24] Zerin Jahan, Surbhi Bhatia Khan, and Mo Saraee. Early dementia detection with speech analysis and machine learning techniques. *Discover Sustainability*, 5(1):65, 2024.
[25] William Jarrold, Bart Peintner, David Wilkins, Dimitra Vergryi, Colleen Richey, Maria Luisa Gorno-Tempini, and Jennifer Ogar. Aided diagnosis of dementia type through computer-based analysis of spontaneous speech. In *Proceedings of the workshop on computational linguistics and clinical psychology: from linguistic signal to clinical reality*, pages 27–37, 2014.
[26] Saugat Kannojia, Anirudh Praveen, Danish Vasdev, Saket Nandedkar, Divyansh Mittal, Sarthak Kalankar, Shaurya Johari, and Vipul Arora. Dementia detection using multi-modal methods on audio data. *arXiv preprint arXiv:2501.00465*, 2024.
[27] Gajendra Jung Katuwal. *Machine learning based autism detection using brain imaging*. Rochester Institute of Technology, 2017.
[28] Yusera Farooq Khan, Baijnath Kaushik, and Bilal Ahmed Mir. Computational intelligent models for alzheimer's prediction using audio transcript data. *Computing & Informatics*, 41(6), 2022.
[29] Yeo Jin Kim, Hye Yeong Jeong, Hui-Chul Choi, Jong-Hee Sohn, Chulho Kim, Sang-Hwa Lee, Joon Soo Shin, So Ra Chin, Yoon Kyoung Lee, So Jung Oh, et al. Effect of right hemispheric damage on structured spoken conversation. *Plos one*, 17(8):e0271727, 2022.
[30] JA Kosmicki, V Sochat, M Duda, and DP Wall. Searching for a minimal set of behaviors for autism detection through feature selection-based machine learning. *Translational psychiatry*, 5(2):e514–e514, 2015.

Just Accepted

[31] Maider Lehr, Emily Tucker Prud'hommeaux, Izhak Shafran, and Brian Roark. Fully automated neuropsychological assessment for detecting mild cognitive impairment. In *Interspeech*, volume 2012, 2012.
[32] XiuLi Li, FuBiao Huang, TieJun Guo, MengChen Feng, and Shan Li. The continuous performance test aids the diagnosis of post-stroke cognitive impairment in patients with right hemisphere damage. *Frontiers in Neurology*, 14:1173004, 2023.
[33] Ning Liu, Zhenming Yuan, and Qingfeng Tang. Improving alzheimer's disease detection for speech based on feature purification network. *Frontiers in Public Health*, 9:835960, 2022.
[34] Lovro Matošević and Alan Jović. Accurate detection of dementia from speech transcripts using roberta model. In *2022 45th Jubilee International Convention on Information, Communication and Electronic Technology (MIPRO)*, pages 1478–1484. IEEE, 2022.
[35] Sahar Sinene Mehdoui, Abdelhamid Bouzid, Daniel Sierra-Sosa, and Adel Elmaghraby. Dementia insights: A context-based multimodal approach. *arXiv preprint arXiv:2503.01226*, 2025.
[36] Jamila Minga, Melissa Johnson, Margaret Lehman Blake, Davida Fromm, and Brian MacWhinney. Making sense of right hemisphere discourse using rhdbank. *Topics in Language Disorders*, 41(1):99–122, 2021.
[37] Pierre Missotten, Gilles Squelard, Michel Ylieff, David Di Notte, Louis Paquay, Jan De Lepeleire, Frank Buntinx, and Ovide Fontaine. Relationship between quality of life and cognitive decline in dementia. *Dementia and geriatric cognitive disorders*, 25(6):564–572, 2008.
[38] Prasenjit Mukherjee, RS Gokul, Sourav Sadhukhan, Manish Godse, and Baisakhi Chakraborty. Detection of autism spectrum disorder (asd) from natural language text using bert and chatgpt models. *International Journal of Advanced Computer Science and Applications*, 14(10), 2023.
[39] Mark Olfson, Melanie Wall, Shang-Min Liu, Michael Schoenbaum, and Carlos Blanco. Declining health-related quality of life in the us. *American journal of preventive medicine*, 54(3):325–333, 2018.
[40] Monalin Pal et al. An intelligent fusion-based behavioral trait prediction for autistic spectrum disorder with artificial intelligence. *Fusion: Practice & Applications*, 15(1), 2024.
[41] Yuval Palgi, Amit Shrira, and Oleg Zaslavsky. Quality of life attenuates age-related decline in functional status of older adults. *Quality of Life Research*, 24(8):1835–1843, 2015.
[42] Jacqueline Peng, Mengge Zhao, James Havrilla, Cong Liu, Chunhua Weng, Whitney Guthrie, Robert Schultz, Kai Wang, and Yunyun Zhou. Natural language processing (nlp) tools in extracting biomedical concepts from research articles: a case study on autism spectrum disorder. *BMC Medical Informatics and Decision Making*, 20(Suppl 11):322, 2020.
[43] Farida Far Poor, Hiroko H Dodge, and Mohammad H Mahoor. A multimodal cross-transformer-based model to predict mild cognitive impairment using speech, language and vision. *Computers in biology and medicine*, 182:109199, 2024.
[44] Ryan A Rava, Samantha E Seymour, Meredith E LaQue, Blake A Peterson, Kenneth V Snyder, Maxim Mokin, Muhammad Waqas, Yiemeng Hoi, Jason M Davies, Elad I Levy, et al. Assessment of an artificial intelligence algorithm for detection of intracranial hemorrhage. *World Neurosurgery*, 150:e209–e217, 2021.
[45] Jamie Reilly, Amy D Rodriguez, Martine Lamy, and Jean Neils-Strunjas. Cognition, language, and clinical pathological features of non-alzheimer's dementias: an overview. *Journal of communication disorders*, 43(5):438–452, 2010.
[46] Nils Reimers and Iryna Gurevych. Sentence-bert: Sentence embeddings using siamese bert-networks. In *Proceedings of the 2019 conference on empirical methods in natural language processing and the 9th international joint conference on natural language processing (EMNLP-IJCNLP)*, pages 3982–3992, 2019.
[47] Alireza Roshanzamir, Hamid Aghajan, and Mahdieh Soleymani Baghshah. Transformer-based deep neural network language models for alzheimer's disease risk assessment from targeted speech. *BMC Medical Informatics and Decision Making*, 21(1):92, 2021.
[48] Sergio Rubio-Martín, María Teresa García-Ordás, Martín Bayón-Gutiérrez, Natalia Prieto-Fernández, and José Alberto Benítez-Andrades. Enhancing asd detection accuracy: a combined approach of machine learning and deep learning models with natural language processing. *Health Information Science and Systems*, 12(1):20, 2024.
[49] Yamanki Santander-Cruz, Sebastián Salazar-Colores, Wilfrido Jacobo Paredes-García, Humberto Guendulain-Arenas, and Saúl Tovar-Arriaga. Semantic feature extraction using sbert for dementia detection. *Brain sciences*, 12(2):270, 2022.
[50] Ravi Shankar, Anjali Bundele, and Amartya Mukhopadhyay. A systematic review of natural language processing techniques for early detection of cognitive impairment. *Mayo Clinic Proceedings: Digital Health*, 3(2):100205, 2025.
[51] Reza Soleimani, Shengjie Guo, Katarina L Haley, Adam Jacks, and Edgar Lobaton. Dementia detection by in-text pause encoding. In *2024 46th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, pages 1–4. IEEE, 2024.
[52] Muhammed Talo, Ozal Yildirim, Ulas Baran Baloglu, Galip Aydin, and U Rajendra Acharya. Convolutional neural networks for multi-class brain disease detection using mri images. *Computerized Medical Imaging and Graphics*, 78:101673, 2019.
[53] Fadi Thabtah and David Peebles. A new machine learning model based on induction of rules for autism detection. *Health informatics journal*, 26(1):264–286, 2020.
[54] Barbara FC Van Heijst and Hilde M Geurts. Quality of life in autism across the lifespan: A meta-analysis. *Autism*, 19(2):158–167, 2015.
[55] Jochen Weiner, Miguel Angrick, Srinivasan Umesh, and Tanja Schultz. Investigating the effect of audio duration on dementia detection using acoustic features. In *Interspeech*, pages 2324–2328, 2018.

Just Accepted

[56] Jochen Weiner, Mathis Engelbart, and Tanja Schultz. Manual and automatic transcriptions in dementia detection from speech. In *Interspeech*, pages 3117–3121, 2017.
[57] Bingyang Wen, Ning Wang, Koduvayur Subbalakshmi, and Rajarathnam Chandramouli. Revealing the roles of part-of-speech taggers in alzheimer disease detection: scientific discovery using one-intervention causal explanation. *JMIR Formative Research*, 7(1):e36590, 2023.
[58] Chuheng Zheng, Mondher Bouazizi, and Tomoaki Ohtsuki. An evaluation on information composition in dementia detection based on speech. *IEEE Access*, 10:92294–92306, 2022.

Just Accepted