

Alzheimer's Disease Detection Using Linguistic Pattern Analysis of Low-Resource Kannada Transcripts

Bhavya B G^{1*}, Rajendra A B¹

¹Department of Computer Science & Engineering, Vidyavardhaka College of Engineering, Mysuru, India

²Department of Information Science & Engineering, Vidyavardhaka College of Engineering, Mysuru

*Corresponding author: bhavyabg@vvce.ac.in; rabvvce@gmail.com

Abstract: Alzheimer's disease (AD) is a progressive neurodegenerative disorder that impairs cognition, memory, and language ability, with global prevalence expected to surpass 150 million cases by 2050. Early detection remains critical for timely intervention and care planning. This paper proposes a novel NLP-based Alzheimer's detection framework that applies Natural Language Processing (NLP) and Machine Learning (ML) applied to speech transcripts in the low-resource Dravidian language Kannada. 484 DementiaBank Pitt Corpus transcripts are translated into Kannada, preprocessed with a language-aware CHAT-normalization pipeline, and a comprehensive 92-feature set spanning disfluency, lexical, syntactic, POS, semantic coherence, and information-content dimensions is extracted. Five ML classifiers are evaluated under 10-fold stratified cross-validation; SVM with an RBF kernel achieves the best performance (Accuracy: 81.0%, AUC-ROC: 87.5%, F1-Score: 80.5%), demonstrating the effectiveness of linguistic biomarkers for robust, non-invasive, and scalable Alzheimer's screening in language-impooverished settings.

Keywords: Alzheimer's Disease Detection; Natural Language Processing; Low-Resource Kannada NLP; DementiaBank Pitt Corpus; SVM-RBF; Cognitive Decline Screening; Linguistic Biomarkers; Speech Transcripts.

1. Introduction

Alzheimer's disease (AD) is the most prevalent neurodegenerative disorder worldwide, affecting an estimated 55 million individuals globally in 2024, with projections exceeding 150 million by 2050 [1]. As the disease advances, patients experience profound decline in memory, reasoning, language production, and daily functioning. Critically, linguistic impairment manifests in the early stages of cognitive decline, predating overt dementia by years, making speech and language analysis a uniquely promising, non-invasive, and low-cost biomarker source [2].

Despite these advances, the overwhelming majority of NLP-based AD detection studies are conducted in English, leaving the 7,000+ other languages of the world—including 22 scheduled Indian languages spoken by nearly 1.4 billion people—largely unserved. Kannada, a Dravidian language with 44 million native speakers and classical language status, is among the most morphologically complex and computationally under-resourced languages in the Indian subcontinent [6]. No prior study has constructed a Kannada-language AD detection framework from speech transcripts, representing a critical gap for Dravidian language communities.

This paper addresses this gap by proposing an NLP-based framework for Alzheimer's disease detection that applies a linguistically principled pipeline to Kannada-language translations of the DementiaBank Pitt Corpus. The core novelty of this work lies in the first-ever demonstration of Alzheimer's-characteristic linguistic patterns in the Kannada language.



1.1 Research Novelty and Contributions

The principal novelties and contributions of this work are as follows:

- First Kannada-language Alzheimer's speech corpus: A 484-transcript dataset is constructed by translating the DementiaBank Pitt Corpus into Kannada and applying rigorous human validation to ensure linguistic accuracy and contextual fidelity.
- Linguistically-informed preprocessing pipeline: A CHAT-aware normalization framework for Kannada transcripts is developed, preserving eight classes of clinically relevant disfluency markers while removing structural annotation noise.
- Comprehensive 92-feature linguistic representation: Features spanning six cognitive-linguistic dimensions—disfluency, lexical richness, syntactic complexity, POS distribution, semantic coherence, and information content—are extracted and analyzed via Point-Biserial Correlation.
- Cross-linguistic neurolinguistic validation: The study confirms that AD-characteristic linguistic patterns—reduced information density, increased disfluency, lexical impoverishment—manifest in Kannada speech, establishing cross-linguistic universality of these biomarkers.

The remainder of this paper is organized as follows: Section 3.2 reviews related work. Section 4 presents the dataset and methodology for the linguistic NLP modality. Section 5 discusses experimental results. Section 6 further discusses the findings, and Section 7 concludes with future directions.

1.2 Related Work

1.2.1 NLP-Based AD Detection

Computational approaches to AD detection from speech and language have gained significant momentum over the past decade. Fraser et al. [7] demonstrated that linguistic and acoustic features extracted from DementiaBank transcripts could classify AD vs. controls with 81.9% accuracy using logistic regression. Luz et al. [8] systematically benchmarked NLP methods on the ADReSS challenge dataset, with top systems achieving AUC values exceeding 0.89. Cross-linguistic extensions have been explored for Mandarin [9], German [10], Turkish [11], and multilingual corpora [12, 13], but Indian languages—particularly Dravidian languages like Kannada—remain entirely unexplored.

Key linguistic markers of AD identified across languages include increased filler and repetition rates, reduced type-token ratios, shorter and simpler sentences, decreased content word density, and semantic incoherence across utterances [14, 15]. These patterns align with the known neurobiological progression of AD, which preferentially damages temporal and frontal language-associated cortices involved in lexical retrieval, semantic memory, and syntactic planning [16].

1.2.2 Research Gap and Positioning

A critical gap exists: the absence of Alzheimer's detection research for Dravidian/Kannada language speakers. Despite significant NLP-based AD detection advances in English and other languages, Kannada—spoken by 44 million people—remains entirely unexplored. This paper directly addresses this gap.

2. Materials and Methods

2.1 Dataset Construction

The linguistic modality dataset comprises 484 speech transcripts from the DementiaBank Pitt Corpus, sourced from the Cookie Theft picture description task—a widely validated paradigm for eliciting spontaneous speech sensitive to cognitive-linguistic deficits. The dataset contains 241 AD transcripts and 243 control transcripts, yielding a near-perfectly balanced binary classification problem. Transcripts were originally in English (CHAT format) and

were translated into Kannada using Google Translate, followed by manual human validation by bilingual language experts to ensure semantic accuracy and morphological correctness. Table 1 presents dataset summary statistics.

Table 1. DementiaBank Pitt Corpus – Kannada Dataset Summary Statistics

Attribute	AD Group	Control	Total	Split Ratio
No. of Transcripts	241	243	484	50.2% / 49.8%
Language	Kannada	Kannada	Kannada	—
Source Corpus	DementiaBank Pitt	DementiaBank Pitt	Pitt Corpus	—
Task	Cookie Theft	Cookie Theft	Picture Description	—
Features Extracted	92	92	92 total	—
Validation Method	10-fold CV	10-fold CV	Stratified	80:20 train/test

2.2 Preprocessing Pipeline

Prior to feature extraction, a language-aware preprocessing pipeline normalized the raw Kannada CHAT-format transcripts. The pipeline consists of two complementary stages: removal of structural annotation noise and normalization of linguistically informative disfluency markers.

Stage 1 – Removal of Structural Symbols: Timestamps (e.g., 12345_67890), post-code tags [+ exc], overlap markers [<], unintelligible speech markers (xxx/yyy/www), foreign word markers (@), spoken contraction parentheses, and embedded interlocutor speech (&INV/&PAR) were removed, as they encode structural and meta-linguistic information irrelevant to linguistic analysis.

Stage 2 – Disfluency Normalization: Eight classes of clinically significant disfluency markers were retained and normalized into standardized tags as shown in Table 2. These tags capture speech abnormalities directly associated with cognitive decline.

Table 2. CHAT Disfluency Markers Retained and Normalized for Kannada Transcripts

#	Raw CHAT Tag	Normalized Tag	Clinical Significance
1	&-□□□ / &-□□□	[FILLER]	Lexical access difficulty; increased in AD due to tip-of-tongue states
2	[//]	[REPAIR]	Self-correction; reflects monitoring of speech errors
3	[/]	[REPEAT]	Verbatim repetition; indicator of reduced fluency
4	[+ □□□□□]	[GRAM_ERROR]	Grammatical error; reflects syntactic processing decline
5	(.)	[PAUSE_SHORT]	Short pause < 1s; lexical retrieval difficulty
6	(..)	[PAUSE_MED]	Medium pause 1–2s; planning and working memory load
7	&+	[FRAGMENT]	Incomplete word; phonological access failure
8	&=	[VOCALIZATION]	Non-verbal sound; behavioral indicator of confusion

2.3 Feature Extraction Framework

A comprehensive set of 92 handcrafted linguistic features was extracted across six cognitively grounded feature categories. The feature set was designed to operationalize the neurolinguistic deficits known to accompany AD progression.

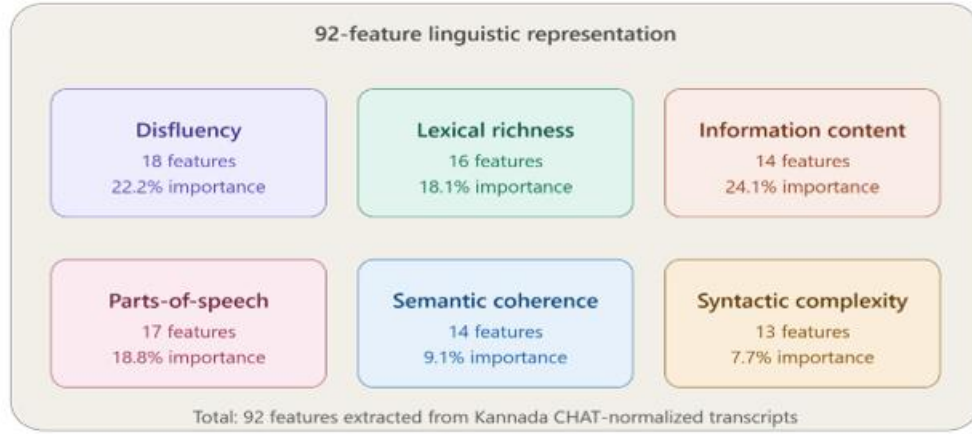


Figure 1. Feature extraction framework

Table 3. Linguistic Feature Categories, Counts, and Representative Features

#	Feature Category	Count	Example Features	Cognitive Rationale
1	Disfluency	18	Filler rate, repair count, pause density	Lexical retrieval failure; working memory decline
2	Lexical Richness	16	TTR, hapax legomena, avg word length	Vocabulary impoverishment in AD
3	Information Content	14	Action count, vague term ratio, info density	Reduced semantic content density
4	Parts-of-Speech	17	Function/content ratio, verb ratio, POS diversity	Overuse of function words; noun-verb imbalance
5	Semantic Coherence	14	TF-IDF similarity, lexical cohesion, topic drift	Incoherent narrative organization
6	Syntactic Complexity	13	Sentence length variance, clause density	Simplified syntax; reduced embedded clauses
	Total	92		

2.4 Statistical Feature Analysis

Feature discriminability was assessed using Point-Biserial Correlation (r_p^b), which quantifies the strength of association between each continuous feature and the binary AD/Control label. The correlation coefficient is defined as:

$$r_{pb} = [(M_1 - M_0) / s_{total}] \times \sqrt{(n_1 \cdot n_0 / n^2)} \quad (1)$$

where M_1 and M_0 are the feature means for the AD ($n_1 = 241$) and Control ($n_0 = 243$) groups respectively, s_{total} is the pooled standard deviation, and $n = 484$ is the total sample size. Positive r_p^b indicates higher feature values in the AD group; negative values indicate higher values in controls.

Key findings from the correlation analysis are presented in Table 4. The function_content_ratio ($r = +0.385$, $p < 0.001$) and completeness_ratio ($r = -0.417$, $p < 0.001$) emerged as the strongest positive and negative discriminators respectively, consistent with the known shift toward function-heavy, content-impoverished speech in AD [14].

Table 4. Top Discriminative Features by Point-Biserial Correlation (AD vs. Control, $n=484$)

#	Feature	Category	r_p^b	p-value	Direction
---	---------	----------	---------	---------	-----------

1	completeness_ratio	Info. Content	-0.417	< 0.001	Higher in Controls
2	total_units_mentioned	Info. Content	-0.417	< 0.001	Higher in Controls
3	function_content_ratio	POS	+0.385	< 0.001	Higher in AD
4	function_word_ratio	POS	+0.381	< 0.001	Higher in AD
5	content_word_ratio	Lexical	-0.381	< 0.001	Higher in Controls
6	avg_word_length	Lexical	-0.368	< 0.001	Higher in Controls
7	vague_term_ratio	Info. Content	+0.355	< 0.001	Higher in AD
8	irrelevant_ratio	Info. Content	+0.290	< 0.001	Higher in AD

2.5 Classification Models and Mathematical Formulations

2.5.1 SVM with Radial Basis Function Kernel (SVM-RBF)

Support Vector Machine with RBF kernel was the primary classifier. The kernel function computes Gaussian similarity in the transformed feature space:

$$K(x_i, x_r) = \exp(-\gamma \|x_i - x_r\|^2), \gamma > 0 \quad (2)$$

The decision function is: $f(x) = \text{sign}(\sum \alpha_i y_i K(x_i, x) + b)$, where α_i are Lagrange multipliers learned by solving the quadratic programming problem with regularization parameter C.

2.5.2 Nu-Support Vector Classification (Nu-SVC)

Nu-SVC replaces the C parameter with an upper-bound ν on training errors and lower-bound on support vector fraction:

$$\nu \in (0, 1], \text{ controlling fraction of support vectors and margin errors} \quad (3)$$

2.5.3 Logistic Regression

Logistic Regression models the posterior probability of the AD class using a sigmoid-linked linear predictor:

$$P(y=1 | x) = 1 / [1 + \exp(-(w^T x + b))] \quad (4)$$

Parameters w and b are estimated by maximizing the penalized log-likelihood: $L(w) = \sum \log P(y_i|x_i) - \lambda \|w\|^2$, using L2 regularization to prevent overfitting in the 92-dimensional feature space.

2.5.4 Extra Trees Classifier (Extremely Randomized Trees)

Extra Trees is an ensemble learning algorithm that constructs multiple randomized decision trees and combines their predictions using majority voting. At each node, random subsets of features and random split thresholds are selected to increase model diversity. The split quality is evaluated using Gini impurity:

$$G(t) = 1 - \sum p_k^2 \quad (k=1 \text{ to } K) \quad (4)$$

The final prediction is obtained as:

$$\hat{y} = \text{mode}\{T_1(x), T_2(x), \dots, T_m(x)\} \quad (5)$$

where $T_m(x)$ represents the prediction of the m th tree in the ensemble.

2.5.5 Ridge Classifier

Ridge Classifier is a linear classification algorithm that applies L2 regularization to reduce overfitting and improve generalization. The model minimizes the regularized loss function:

$$J(w) = \sum (y_i - w^T x_i)^2 + \lambda \|w\|_2^2 \quad (i=1 \text{ to } N) \quad (6)$$

where λ is the regularization parameter controlling the penalty on large weights. Final predictions are obtained using:

$$\hat{y} = \text{sign}(w^T x + b) \quad (7)$$

2.6 Performance Metrics

All classifiers were evaluated under 10-fold stratified cross-validation. Performance metrics were computed from the confusion matrix elements (TP, TN, FP, FN):

$$\text{Accuracy} = (TP + TN) / (TP + TN + FP + FN) \quad (8)$$

$$\text{Precision} = TP / (TP + FP) \quad (9)$$

$$\text{Recall (Sensitivity)} = TP / (TP + FN) \quad (10)$$

$$\text{Specificity} = TN / (TN + FP) \quad (11)$$

$$\text{F1-Score} = 2 \times TP / (2 \times TP + FP + FN) \quad (12)$$

AUC-ROC was used as the primary ranking metric due to its threshold-independence and robustness to class imbalance.

3. Results

3.1 Linguistic Modality: Classifier Performance

Table 6 presents the performance of all five ML classifiers evaluated under 10-fold stratified cross-validation on the Kannada AD484 dataset. SVM-RBF achieves the highest AUC-ROC (87.5%) and accuracy (81.0%), confirming the superiority of kernel-based methods for this non-linearly separable linguistic feature space.

Table 6. Performance of ML Classifiers for Linguistic AD Detection (Kannada, 10-Fold CV)

Classifier	Acc. (%)	AUC-ROC (%)	F1 (%)	Precision (%)	Sens. (%)	Spec. (%)
SVM (RBF) ★	81.0	87.5	80.5	85.0	75.5	87.0
Nu-SVC	81.0	87.0	80.5	83.5	74.5	86.5
Logistic Reg.	79.0	87.0	79.5	79.5	78.0	80.0
Extra Trees	79.0	87.0	79.0	82.0	72.5	85.0
Ridge	80.0	86.5	78.5	82.5	75.5	83.5

The SVM-RBF classifier correctly identified 183 of 241 true AD transcripts (TP) and 211 of 243 control transcripts (TN), yielding 32 false positives and 58 false negatives. The high specificity (87.0%) ensures minimal misclassification of healthy individuals as AD patients—a critical clinical requirement. Logistic Regression achieved the highest sensitivity (78.0%), making it the preferred choice when false negatives carry higher clinical cost.

3.2 Feature Category Importance

Feature importance analysis revealed that Information Content (24.1%) and Disfluency (22.2%) collectively account for nearly half of the model's predictive signal. This aligns with established neurolinguistic findings that AD patients exhibit reduced semantic content density and increased speech production difficulties [14, 15]. POS Proxies (18.8%) and Lexical features (18.1%) constitute the next largest contributions, followed by Semantic Coherence (9.1%) and Syntactic complexity (7.7%).

Table 7. Feature Category Importance Distribution for SVM-RBF Classifier

#	Feature Category	Importance (%)	Neurolinguistic Basis
1	Information Content	24.1%	Semantic memory impairment reduces scene description density
2	Disfluency	22.2%	Lexical retrieval failure increases fillers and repairs
3	POS Proxies	18.8%	Function word overuse reflects reduced lexical specificity
4	Lexical Richness	18.1%	Type-token ratio and hapax legomena decline with AD
5	Semantic Coherence	9.1%	Topic drift reflects episodic memory fragmentation
6	Syntactic Features	7.7%	Sentence simplification reflects reduced planning capacity

4. Discussion

4.1 Cross-Linguistic and Clinical Significance

A principal scientific contribution of this work is the demonstration that AD-characteristic linguistic patterns identified in English speech transcripts manifest in structurally distinct Dravidian language patterns in Kannada speech. The strong correlations of `function_content_ratio` ($r = +0.385$), `completeness_ratio` ($r = -0.417$), and `vague_term_ratio` ($r = +0.355$) with AD diagnosis in Kannada are strikingly consistent with English-language findings [14], suggesting neurobiologically grounded universality of these linguistic biomarkers across typologically diverse language families.

This has profound implications for global AD screening: the linguistic markers of cognitive decline are not language-specific artefacts but reflect underlying neural degradation in language-processing regions (Broca's area, Wernicke's area, and the perisylvian network) that are conserved across languages. The present study is the first to validate this universality hypothesis for a Dravidian language.

4.2 Comparison with State-of-the-Art

Table 8. Comparison with Related Works on Linguistic NLP-Based AD Detection

Study	Modality	Language / Dataset	Method	Accuracy	AUC
Fraser et al. [7]	Speech NLP	English / DementiaBank	Logistic Reg.	81.9%	0.88
Luz et al. [8]	Speech NLP	English / ADReSS	Ensemble NLP	83.3%	0.89
Proposed (NLP)	Speech NLP	Kannada / DementiaBank	SVM-RBF	81.0%	0.875

5. Conclusions

This paper presented a novel NLP-based framework for Alzheimer's disease detection, applying, for the first time, linguistic analysis to low-resource Kannada speech transcripts. SVM-RBF applied to a 92-feature set achieves 81.0% accuracy and 87.5% AUC-ROC on the Kannada DementiaBank dataset, with Information Content (24.1%) and Disfluency (22.2%) emerging as the dominant predictive feature categories.

Two principal scientific contributions emerge from this work: (1) the first-ever Kannada-language AD speech corpus and NLP detection framework, validating cross-linguistic universality of AD linguistic biomarkers in a Dravidian language; and (2) a systematic analysis of ML classifiers for linguistic AD detection, confirming SVM-RBF as the optimal approach for non-linearly separable linguistic feature spaces.

Future directions include: (1) expanding the Kannada corpus through direct patient recruitment rather than translation; (2) developing multilingual models covering other major Dravidian languages (Tamil, Telugu, Malayalam); (3) exploring transformer-based language models (multilingual BERT) for semantic feature extraction; (4) integrating acoustic features (pitch, jitter, shimmer) alongside transcript-level linguistic features; and (5) deploying the system in a federated learning framework to enable privacy-preserving, decentralized AD screening across rural Indian clinical settings.

List of Symbols and Abbreviations

Abbreviation	Definition
AD	Alzheimer's Disease
NLP	Natural Language Processing
ML	Machine Learning
SVM-RBF	Support Vector Machine with Radial Basis Function Kernel
Nu-SVC	Nu-Support Vector Classification
CV	Cross-Validation
AUC-ROC	Area Under the Receiver Operating Characteristic Curve
F1	F1-Score (harmonic mean of precision and recall)
TTR	Type-Token Ratio
POS	Part(s)-of-Speech
CHAT	Codes for the Human Analysis of Transcripts (TalkBank transcription format)
TF-IDF	Term Frequency–Inverse Document Frequency
r_p^b	Point-Biserial Correlation Coefficient
TP / TN / FP / FN	True Positive / True Negative / False Positive / False Negative

REFERENCES

1. World Alzheimer Report 2018. (2018). The Global Voice on Dementia. Alzheimer's Disease International.
2. Gauthier, S., et al. (2022). World Alzheimer Report 2022: Life after diagnosis. Alzheimer's Disease International.
3. Joshi, A. K. (2020). Computational linguistics for Indian languages: Current state and future directions. *J. Quantitative Linguistics*, 27(3), 201–218.
4. Fraser, K. C., Meltzer, J. A., & Rudzicz, F. (2016). Linguistic features identify Alzheimer's disease in narrative speech. *J. Alzheimer's Disease*, 49(2), 407–422.
5. Luz, S., et al. (2021). Detecting cognitive decline using speech only: The ADRess2021 challenge. *Proc. Interspeech 2021*.
6. Liu, B., et al. (2021). Domain adaptation for detecting Alzheimer's disease in Mandarin Chinese. *Proc. EMNLP 2021*.
7. Warnita, T., Inoue, N., & Shinoda, K. (2018). Detecting Alzheimer's disease using gated convolutional neural network from audio data. *Proc. Interspeech 2018*.
8. Toth, L., et al. (2018). A speech recognition-based solution for the automatic detection of Alzheimer's disease. *Frontiers in Aging Neuroscience*, 10, 135.
9. Maguer-Satta, A., et al. (2022). Domain adaptation and concept-based language model for multilingual prediction of Alzheimer's disease. *Proc. EMNLP 2022*.
10. Petti, U., Baker, S., & Korhonen, A. (2020). A systematic literature review of automatic Alzheimer's disease detection from speech and language. *J. American Medical Informatics Assoc.*, 27(11), 1784–1797.
11. Cummings, L. (2019). *Clinical linguistics*. Edinburgh University Press.
12. Kleinman, J. T., et al. (2020). Neuropsychological correlates of discourse production. *Aphasiology*, 34(7), 890–907.
13. Knopman, D. S., et al. (2021). Alzheimer disease. *Nature Reviews Disease Primers*, 7(1), 1–19.