Discover Sustainability

Research

Early dementia detection with speech analysis and machine learning techniques

Zerin Jahan[1] · **Surbhi Bhatia Khan**[1] · **Mo Saraee**[1]

Received: 12 December 2023 / Accepted: 11 March 2024
Published online: 11 April 2024

Abstract
This in-depth study journey explores the context of natural language processing and text analysis in dementia detection, revealing their importance in a variety of fields. Beginning with an examination of the widespread and influence of text data. The dataset utilised in this study is from TalkBank's DementiaBank, which is basically a vast database of multimedia interactions built with the goal of examining communication patterns in the context of dementia. The various communication styles dementia patients exhibit when communicating with others are seen from a unique perspective by this specific dataset. Thorough data preprocessing procedures, including cleansing, tokenization, and structuring, are undertaken, with a focus on improving prediction capabilities through the combination of textual and non-textual information in the field of feature engineering. In the subsequent phase, the precision, recall, and F1-score metrics of Support Vector Machines (SVM), K-Nearest Neighbours (KNN), Random Forest, and Artificial Neural Networks (ANN) are assessed. Empirical facts are synthesized using text analysis methods and models to formulate a coherent conclusion. The significance of text data analysis, the revolutionary potential of natural language processing, and the direction for future research are highlighted in this synthesis. Throughout this paper, readers are encouraged to leverage text data to embark on their own adventures in the evolving, data-centric world of dementia detection.

Keywords Dementia · Speech transcript analysis · Text mining · Feature extraction · Word embedding · Machine learning · Classification

1 Introduction

Dementia is a complex and debilitating neurological condition that affects millions of people worldwide. It is not a specific disease but rather an umbrella term used to describe a range of cognitive impairments that significantly interfere with a person's ability to function independently in their daily life. As a person ages, the risk of developing dementia increases, and it is estimated that the prevalence of dementia will continue to rise as global populations age.

Dementia primarily affects memory, thinking, and communication skills, leading to a progressive decline in cognitive abilities. The most common form of dementia is Alzheimer's disease, which accounts for ~ 60–80% of all cases. Other types of dementia include vascular dementia, Lewy body dementia, frontotemporal dementia, and mixed dementia, among others.

The effects of dementia are far-reaching and not limited to the individual diagnosed with the condition. It profoundly impacts their families, caregivers, and society as a whole. As the disease progresses, those affected may experience

✉ Surbhi Bhatia Khan, s.khan138@salford.ac.uk; Zerin Jahan, z.jahan@edu.salford.ac.uk; Mo Saraee, m.saraee@salford.ac.uk | [1]School of Science, Engineering and Environment, University of Salford, Salford, UK.

personality changes, mood swings, difficulty recognizing familiar faces, and even challenges in performing simple tasks. Eventually, individuals with advanced dementia may lose their ability to communicate, become fully dependent on others for their care, and experience physical limitations.

Caregivers of individuals with dementia face significant emotional, physical, and financial burdens. The responsibility of providing around-the-clock care, managing behavioural changes, and coping with the progression of the disease can be overwhelming. Additionally, dementia places a considerable strain on healthcare systems and social support networks, demanding specialized services and resources.

Dementia is given top priority in the realm of public health by the World Health Organization (WHO). In May 2017, the World Health Assembly approved the Global Action Plan on the Public Health Response to Dementia for the period 2017–2025. This comprehensive plan outlines a roadmap for policymakers at various levels, including global, regional, and national partners, as well as WHO itself. The plan encompasses several key areas: elevating dementia to a public health priority, promoting awareness and inclusivity, reducing dementia risk, offering diagnosis, treatment, and care, establishing information systems for dementia, supporting caregivers, and fostering research and innovation. To aid in monitoring the progress of this global dementia action plan, WHO created the Global Dementia Observatory (GDO), a data platform that compiles vital dementia-related data from nations across the world. Furthermore, WHO introduced the GDO Knowledge Interaction Platform to facilitate cross-regional and cross-national learning and collaboration, promoting global action in the field of dementia. This platform serves as a repository of exemplary practices in dementia care and research.

Despite the immense challenges posed by dementia, there is ongoing research aimed at understanding the underlying causes and developing effective treatments and interventions. Early detection and intervention can improve the quality of life for individuals with dementia and their families, allowing for better planning and access to support services [1].

1.1 Motivation and contributions

In order to enhance the global dementia action plan's monitoring, WHO developed the Global Dementia Observatory (GDO), a data platform that aggregates country data on 35 crucial dementia indicators across the global action plan's seven strategic areas. The WHO added the GDO Knowledge Interaction Platform to the GDO with the intention of fostering reciprocal learning and multidirectional interaction across regions, countries, and individuals to enable global action. It serves as a repository for dementia-related best practises [1].

Dementia is a growing public health concern with a profound impact on individuals, families, and societies worldwide. As the global population ages, the prevalence of dementia is expected to rise, necessitating innovative approaches to early detection and intervention. In recent years, advancements in natural language processing (NLP) and machine learning techniques have opened up exciting opportunities to detect dementia from textual data, such as written or spoken language.

The objectives of the work are given below:

- Applying text mining techniques different groups of people to extract relevant linguistic features from the transcribed speech data. Identify specific language patterns, semantic changes, vocabulary usage, and syntactic structures that might be indicative of early cognitive decline or the presence of dementia.
- Investigate the potential of the developed system and text mining methods to detect signs of cognitive decline associated with dementia.
- Evaluate the accuracy and reliability of the speech transcription and text mining approach for dementia detection. Compare its performance against existing diagnostic methods to assess its potential as a complementary or alternative screening tool.

The paper is organized as follows. Section 1 outlines the introduction with the research objectives. Section 2 discusses the comprehensive analysis of the research done in the area. Section 3 lists the proposed methodology and the steps undertaken to complete the objectives. Section 4 present the critical evaluation and analysis of the results and the last section concludes.

2 Literature review

A broad range of study has evolved in the constant pursuit of understanding and minimising the terrible impact of dementia on individuals and societies, pushed by advances in technology, languages, and the field of machine learning. This section takes a detailed look at the growing landscape of dementia research, focusing on a varied body of work that uses speech and language analysis to diagnose this devastating disorder.

Alzheimer's disease and other conditions that fall under the umbrella term "dementia" pose a serious threat to global health. Finding subtly linguistic indicators that signal the beginning of cognitive decline has become a focus of research since language is a complex window into cognition. The studies covered in this literature review come together at the nexus of deep learning, NLP, and conventional linguistic analysis, coming together in their quest to harness the complex patterns of human speech to identify alterations associated with dementia.

2.1 Background

Many researchers investigate a wide range of approaches, from neural networks to conventional linguistic features, with the common aim of identifying the linguistic hallmarks of dementia. As we examine the results of these research, a complex picture starts to take shape. While some studies focus on the finer points of grammatical constructions, others explore the depth of semantic meaning. Few consider the temporal aspect, tracking linguistic developments across time. The use of multimodal techniques, which include voice, text, and maybe other modalities, is showing promise to identify dementia more thoroughly and accurately.

Here Zheng et al. [2] focused on automatic dementia detection using speech samples of dementia patients. The study incorporated advancements in Deep Learning as well as NLP to develop language models that can accurately identify dementia patterns. The authors proposed a novel approach by exploring the use of stop words, typically removed during text preprocessing, as they may provide non-contextual information useful for dementia detection. Three language models were implemented: one processing context words only, another was stop words with Parts of Speech (PoS) tag sequences, and a third model combining both. The experiments were carried out with the help of a dataset of transcribed texts from dementia patients and control individuals. The findings revealed that both grammar and vocabulary have a substantial role in dementia classification. Under 10-fold cross-validation, the model integrating context words and stop words got the maximum accuracy of 81.54%, while the model processing context words only scored 70.00%. The literature review highlights potential areas for future research, suggesting the use of sophisticated parsers to extract finer features from sentences and exploring data augmentation techniques to address limited data challenges. The breakdown of sentences into lower-level representations, such as PoS tag sequences, is considered a promising approach for generating meaningful data for training. In this work Comuni [3] compared the effectiveness of two machine learning algorithms. In order to identify signs of Alzheimer's disease in conversation transcripts, used the Bayesian Network and the LSTM Recurrent Neural Network. The number of seniors suffering from Alzheimer's disease is rising as a result of the increased life expectancy around the world, making early detection essential for effective treatment. The study discovered encouraging findings in the use of machine learning to identify suspected Alzheimer's symptoms and mild dementia in conversation transcripts. On the chosen dataset, the LSTM

performed extremely well, achieving an accuracy of 86.5%, whereas the Bayesian Network did so at 72.1%. The empirical results confirmed the hypotheses formulated in the research questions, demonstrating that machine learning can effectively detect Alzheimer's disease using text samples. The study highlighted that a combination of lexical, semantic features, and the age of the subject contributed to successful detection across all three algorithms. The LSTM, with its complex structure and ability to handle long-term dependencies, outperformed the Bayesian Network as the state-of-the-art algorithm in natural language processing. Here Pulido et al. [4] explored the application of automatic speech and voice analysis techniques for monitoring patients with Alzheimer's disease (AD). As the number of AD cases continues to rise globally, early detection and efficient healthcare systems are essential. Currently, AD lacks a cure, and diagnosis often occurs late when irreversible damage has already occurred. Therefore, early detection becomes critical for effective treatment and longitudinal studies. The review highlighted the importance of finding accessible biomarkers for AD detection and proposes voice or speech analysis as a powerful indicator of cognitive state. Several studies have shown that voice analysis can detect early AD symptoms years before clinical

diagnosis, making it a promising avenue for early detection. However, longitudinal studies and large databases with longitudinal information are lacking in the field, hindering the progress in understanding AD progression. The review stresses the importance of public database challenges and the need for substantial longitudinal studies to improve diagnostic accuracy. Advancements in technology, especially in automatic systems, offer faster and less computationally burdensome results for speech acoustic analysis. The concept of Health 4.0, which enhances security for dementia patients through sensing devices and therapeutic tools, opens new avenues for linguistic analysis in healthcare. The review categorized studies based on conventional and non-conventional feature extraction processes, emphasizing the significance of combining both linear and non-linear aspects of the voice signal for comprehensive results. Machine learning algorithms, such as MLP or CNN classifiers, show promise for complex classification systems on voice and speech signals. However, the review noted the need for larger datasets and standardized tasks to ensure robust and accurate classification. Multimodal analysis involving various biomarkers and assessment methods could further improve early AD detection. Nambiar et al. [5] explored the early detection of dementia in patients using English speech transcript files through the application of deep learning and natural language processing techniques. The study highlights the importance of early detection in managing dementia symptoms and improving access to medical attention and treatment. The paper employed various vector embeddings (GloVe, Word2Vec, Doc2Vec) and transformer models (BERT, RoBERTa, ALBERT) in combination with deep classifiers (LSTM, BiLSTM, GRU) to detect individuals with and without dementia using the DementiaBank dataset. The best-performing models achieved an accuracy of 0.812 and an F1 score of 0.81, using the BERT+BiLSTM and ALBERT+BiLSTM combinations, respectively. The study acknowledged limitations in being monolingual and trained solely on English speech transcripts. However, it suggested future improvements by overcoming hardware restrictions and developing multilingual models for dementia detection using speech transcripts. Here Zhu et al. [6] proposed WavBERT models present a promising approach to automatic dementia detection by integrating Wav2vec ASR with BERT, leveraging both semantic and non-semantic information. The extended WavBERT models show improved performance compared to the basic model, achived by inclusion of non-semantic information preservation techniques. The authors indicated future goals, including exploring the transformer encoder of BERT for pre-training the embedding conversion network, to further enhance the models' performance and effectiveness in dementia detection. This research paper explores the use of Wav2vec and Bidirectional Encoder Representations from Transformers (BERT) for dementia detection by exploiting both semantic and non-semantic information from patient's speech data. The authors propose a basic WavBERT model that extracts semantic information using Wav2vec and analyzes it with BERT for dementia detection. Additionally, extended WavBERT models are introduced to preserve non-semantic information by determining inter-word pauses using Wav2vec's blank tokens and designing a pre-trained embedding conversion network. The evaluation on the ADReSSo dataset shows that the WavBERT models achieved a high accuracy of 83.1% in the classification task, the lowest root-mean-square error (RMSE) score of 4.44 in the regression task, and a mean F1 score of 70.91% in the progression task. The effectiveness of WavBERT models, which exploit both semantic and non-semantic speech information, was confirmed, and they outperformed the baseline linguistic model. In their recent journal article, Burke et al. [7] delved into the potential of lexical-semantic characteristics in spontaneous speech as predictive indicators of cognitive status in individuals with Alzheimer's clinical syndrome (ACS) and those without cognitive impairments (healthy controls). Building on previous studies that had demonstrated the efficacy of these characteristics in predicting cognitive issues in individuals with mild

cognitive impairment (MCI), this research expanded its scope by investigating four additional speech indices linked to language processing research. The study involved the analysis of speech transcripts from the Cookie Theft Task in 81 individuals with ACS and 61 healthy controls, using random forest and logistic machine learning techniques. The primary aim was to ascertain whether individual-level lexical-semantic features could effectively distinguish between individuals with ACS and their healthy counterparts. The outcomes indicated that the lexical- semantic features, referred to as Ostrand and Gunstad's (OG) features, exhibited strong performance across various machine learning models, achieving a commendable classification accuracy of 78.4%. Interestingly, the newly introduced "New" lexical-semantic features also displayed potential in discerning between ACS and healthy control groups. This research suggests that features related to spontaneous speech, previously identified as sensitive to MCI, can similarly serve as indicators of ACS in the studied population. The introduction of novel markers highlighting lexical and semantic distinctiveness further enhances the identification of individuals with ACS. These findings encourage further exploration of automated lexical-semantic analyses as a valuable tool in the early detection of both MCI and Alzheimer's disease (AD). Guerrero-Cristancho et al. [8] discusses advances in pattern recognition approaches for detecting linguistic problems in Alzheimer's disease patients. The study makes use of numerous language variables derived from transcripts of Alzheimer's patients and healthy controls,

such as word embeddings, word frequency, and grammar aspects. The study used transcripts from the Dementia-Bank Pitt Corpus, which included data from 98 Alzheimer's disease patients and 98 healthy controls. The researchers retrieved three types of features: 1200 word embedding features, 1408 Term Frequency-Inverse Document Frequency (TF-IDF) features, and 8 grammar features. Four unique models were proposed: three models based on independent feature sets and a fourth model combining all three feature sets using an early fusion method. A Leave-One-Out cross-validation strategy was used during the optimisation phase, which is a reliable technique for model evaluation. The purpose was to assess the efficiency of the suggested models in discriminating between AD patients and healthy controls based on language features. The results illustrate the effectiveness of the proposed methodology. The early fusion technique, which merges the three feature sets, achieved the highest accuracy, reaching up to 81.7%. This shows that incorporating multiple language elements improves the overall accuracy of AD detection. Surprisingly, even a limited set of linguistic traits produced encouraging results, with accuracy values reaching up to 72.8%. This emphasises the significance of grammatical patterns in detecting AD-related language problems. The comprehensive analysis of the work is done and is shown in Table 1.

2.2 Research gap

Based on the future directions and limitations identified in the previous research on dementia detection, several research gaps and areas for further investigation emerge:

1. One significant research gap lies in the need for larger datasets to enhance the generalization of algorithms, particularly complex models like LSTM. Expanding the dataset size could help improve the accuracy and robustness of dementia detection models, enabling more reliable predictions.
2. Another area of research could involve the exploration of additional features beyond those proposed in the previous work. Augmenting the feature set could potentially lead to better model performance and more informative insights into dementia-related speech patterns.
3. Addressing the research gap regarding sample size and the progression of diseases over time, longitudinal studies are recommended. Long-term observations of dementia patients can provide valuable insights into how speech patterns change throughout the course of the disease, facilitating more accurate diagnosis and monitoring.
4. An interesting avenue for future research involves investigating the relationship between speech patterns and apathetic symptoms in dementia patients. Prior research suggests a connection, and further exploration could shed light on whether speech characteristics align with apathetic symptoms and their implications for diagnosis and care.
5. To strengthen the understanding of dementia, future research could delve into the association between speech and language characteristics and neuropathological changes. Incorporating validated biomarkers and postmortem follow-up could provide insights into the relationship between speech patterns and specific underlying pathology, potentially leading to more precise diagnostic tools.
6. Addressing the issue of small and imbalanced datasets, future work should aim to collect samples from diverse sources and clinical sites. This approach would enhance the generalizability of findings and reduce potential biases associated with small or unrepresentative datasets.

3 Proposed methodology

This section discusses the methodology development for dementia detection through text analysis. Feature extraction entails identifying and quantifying salient linguistic attributes within the text data. These features serve as the foundation for training machine learning models. Given the nuanced nature of language, careful consideration is given to selecting features that capture meaningful linguistic patterns indicative of cognitive health. Simultaneously, a robust methodology is devised to ensure the effectiveness and reliability of the dementia detection process. This involves the selection of appropriate algorithms, model architectures and evaluation metrics. The whole methodology is depicted in Fig 1.

Data is loaded and processed linguistically in this data-driven pipeline, after which participant lines are extracted and sentences are merged into tokenized chunks. In order to classify the data into binary categories (0 for healthy, 1 for dementia), linguistic characteristics like hesitation and word repetition are taken into consideration. Stopword elimination, word stemming, and tokenization for feature extraction are steps in the further refining process. The machine learning algorithm is trained on the training set, which is separated from the testing set in the dataset. 5-fold cross-validation

Table 1 Research comparison of previous works

Study (year)	Methodology	Accuracy (%)	Key findings and notes
Zheng et al. (2022)	Deep learning and NLP	81.54	Explored stop words for dementia detection. Integrating both context words and stop words improved accuracy
Comuni (2019)	Bayesian network vs. LSTM	LSTM: 86.5 Bayesian: 72.1	LSTM outperformed Bayesian Network for Alzheimer's detection. Combined lexical, semantic features, and age
Pulido et al. (2020)	Speech and voice analysis	N/A	Voice analysis shows promise for early AD detection. Emphasized the need for larger datasets
Nambiar et al. (2022)	Deep learning and NLP	81.2	Explored various embeddings and transformer models. Achieved high accuracy with BERT + BiLSTM
Zhu et al. (2021)	Wav2vec ASR with BERT	83.1	Leveraged both semantic and non- semantic information for dementia detection
Burke et al. (2023)	Lexical-semantic features	78.4	Lexical-semantic features performed well in distinguishing ACS patients from healthy controls
Juan et al. (2020)	Pattern recognition	72.8 Up to 81.7	Explored various language variables and feature sets. Early fusion of features achieved the highest accuracy

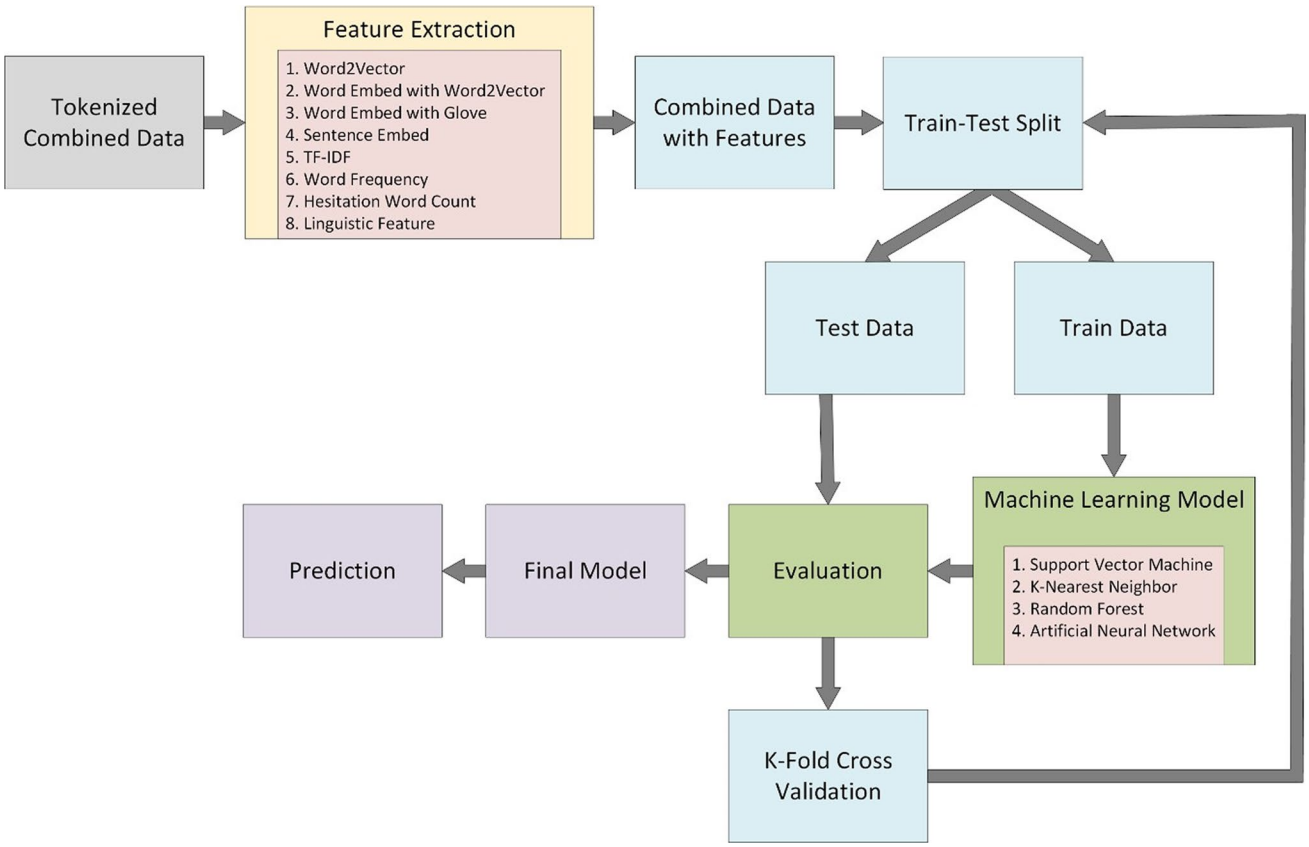

Fig. 1 Diagram of methodology of the dementia detection model

is used to ensure robust assessment when the model is evaluated on the test set. By comparing predicted and actual labels, the accuracy of the trained model's final prediction on the training data is determined. The comprehensive results, including the final prediction, cross-validation scores, and accuracy, are displayed to provide a holistic overview of the model's performance.

The algorithm used is given below:

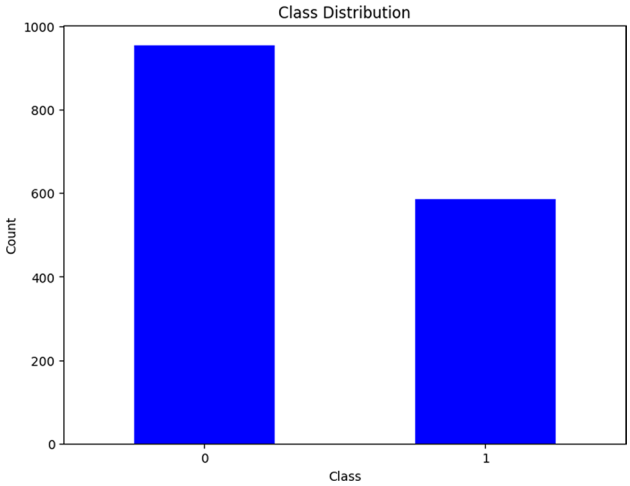

Fig. 2 Class distribution for dementia detection

Fig. 3 Here the wordcloud presentations which represents the most said word

Algorithm 1 Dementia detection

1. Input text data x
2. Perform training x_train
3. Obtain various model for prediction. Like:
   KNN -> $\hat{y} = majority(neighbours)$
   SVM-> Decision function: $f(x) = sign\ (w * x + b)$
   Objective function: $\frac{1}{2}||w||^2 + C\ \sum_{i=1}^{N} \xi_i$ subject to $y_i(w * x + b) \geq 1 - \xi_i$ and $\xi_i \geq 0$
   for i = 1,2,..N
   Random Forest-> $\hat{y} = mode(trees)$
   ANN-> Forward pass: $a^{(l+1)} = g(z^{(l+1)})$, where $(z^{(l+1)}) = W^{(l)}a^{(l)} + b^{(l)}$
   Loss function: $J(\theta) = -\frac{1}{m}\sum_{i=1}^{m}[y^{(i)}\log(a^{(L)}) + (1 - y^{(i)})\log(1 - a^{(L)})]$
   Backward pass: $\frac{\partial J}{\partial W^{(l)}} = \frac{1}{m} dZ^{(l)}a^{(l-1)T}$, $\frac{\partial J}{\partial b^{(l)}} = \frac{1}{m}\sum_{i=1}^{m} dZ^{(l)}$
4. Accuracy evaluation with cross validation
   $\bar{P} = \frac{1}{k}\sum_{i=1}^{k} P_i$, where k=5
5. prediction= $\hat{y} = sign\ (f(x))$
6. Accuracy calculation y_pred[*i*], y_test[*i*]
7. Accuracy loss calculation
8. If accuracy< *m*, *m*= standard accuracy value
   Change parameter, increasing accuracy
9. Else prediction =$\hat{y}$
10. End

3.1 Dataset and data preprocessing

The TalkBank project, directed by Brian MacWhinney at Carnegie Mellon University, includes the collaborative and specialised database known as DementiaBank [9]. The database offers data in a variety of languages, including English. There are also numerous corpora of text and audio files. The DementiaBank dataset is accessed via a password-protected system for cooperative members. A concise, respectful request through the project supervisor to TalkBank led to permission, emphasizing responsible and ethical data handling. Pitt corpus has the greatest data that was collected among them. Pitt corpus was chosen for the research. There are two distinct text scripts within the Pitt corpus. One is for healthy people and is labelled 'Control,' while the other is labelled 'Dementia,' and is for dementia patients. Here, the script is more of a question-and-answer format. The data source used in this research comprises transcribed speech data stored in CHA (CHAT) files. These files are commonly used in linguistic and phonetic research for the transcription of spoken language. CHA files contain linguistic annotations and metadata, making them suitable for our analysis.

Figure 2 shows the class distribution of Dementia detection text. To begin, the dataset's basic characteristics was examined, including its size, structure, and data types. The dataset comprised then with two classes: 'Control' (denoted

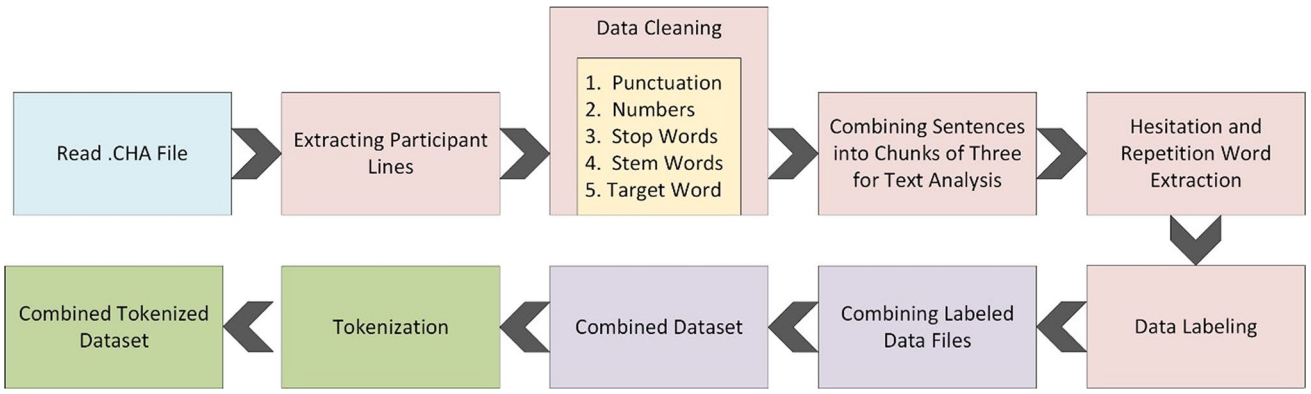

Fig. 4 Preprocessing workflow

by 0) and 'Dementia' (denoted by 1). In total, It includes 1541 data samples, with 955 belonging to the control group and 586 to the dementia group. Understanding the class distribution is essential as it can highlight potential class imbalances that may need to be addressed during modelling.

Some analysis on data was also done and the word cloud presentation are shown in Fig. 3.

3.2 Data preprocessing

The preprocessing workflow consists of several key steps, each designed to clean and prepare the raw transcribed data for subsequent analysis. Below, a detailed theoretical description was provided which is a crucial step in understanding and gaining insights from any dataset. In this case, this dealing with a dataset related to dementia detection through natural language processing of each step which is depicted in Fig 4.

In the process of preparing textual data for analysis, the initial objective is to extract and retain participant lines denoted as '*PAR:' lines, excluding investigator lines to focus solely on speech and language patterns. Subsequently, punctuation marks, numbers, and specified target words such as 'exc' and 'gram' are removed, ensuring a clean and relevant dataset for in-depth analysis. An additional step involves combining consecutive sentences into chunks of three to preserve contextual information for more accurate text analysis. The extraction of hesitation words (e.g., "uh," "um") and repeated words, significant linguistic markers of cognitive impairment, is crucial, with a meticulous filtering process applied to eliminate noise elements. Without using a predefined dictionary, the hesitation words were manually extracted by analyzing the dataset. After comprehensive preprocessing, the dataset comprises 955 healthy control samples and 586 dementia-affected samples, totaling 1541 samples. The final step involves data labeling, assigning '0' to healthy control and '1' to dementia-affected data, resulting in a structured dataset ready for various text analysis approaches and machine learning models, facilitating subsequent studies such as dementia diagnosis and linguistic pattern exploration [10].

Following the initial preprocessing steps, further cleaning of the data was carried out. The process included tokenizing words, which involves separating the text into individual words or tokens to facilitate more effective analysis, enabling tasks such as word frequency counting and pattern analysis. Subsequently, stopwords, common but contextually less meaningful words like "the," "is," and "and," were removed to reduce data dimensionality and enhance the relevance of language patterns associated with dementia in later text analysis. Additionally, stemming was applied to reduce words to their root or base form.

After that, the text data underwent further cleaning and transformation through various techniques. Initially, a Word2Vec model was trained on the preprocessed text data, generating 100-dimensional word vectors using a Continuous Bag of Words (CBOW) model with a window size of 5 and a minimum word count of 1. Sentence embeddings were then obtained by averaging the word vectors within each sentence, creating a NumPy array for subsequent analysis and model training. The process continued with the utilization of the Universal Sentence Encoder model for encoding sentences into fixed-length vectors, providing a high-dimensional numerical representation of semantic content. TF-IDF vectorization with N-grams was employed to quantify term importance within sentences, considering

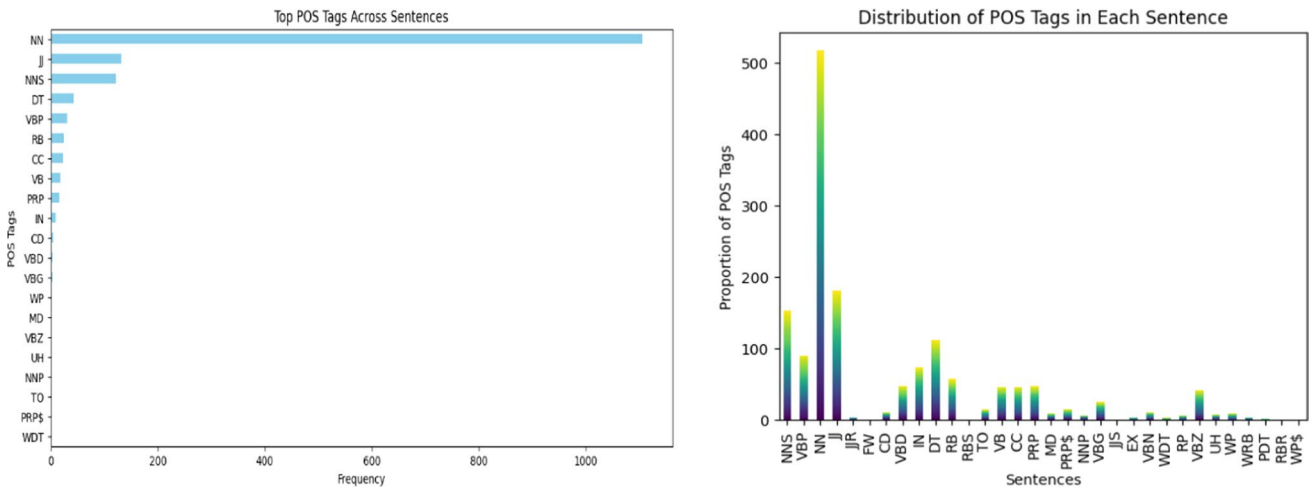

Fig. 5 Top and distribution of Pos Tag pos tags used

both unigrams and bigrams. The code initialized a TF-IDF vectorizer, retaining 100 features for efficiency, and applied it to sentences from the "Sentence" column. GloVe embeddings, capturing semantic relationships between words, were loaded from a pre-trained model, allowing the calculation of sentence embeddings for each sentence in the dataset. Linguistic features such as word frequency and document length were then calculated, with word frequency providing insights into the prominence of specific terms, and document length indicating the size or complexity of sentences. Finally, the counting of hesitation words, known linguistic markers of cognitive impairment, was performed using a custom function applied to each sentence in the "Sentence" column.

In the Linguistic Features Extraction phase, crucial for dementia detection and text analysis tasks, various linguistic features were meticulously calculated. These features, including the Type Token Ratio (TTR) [7] for lexical diversity, prepositional ratio, verb-to-noun ratio, and content density, contribute significantly to understanding linguistic characteristics. TTR, computed by dividing unique words by total words, gauges vocabulary richness, with a CountVectorizer preserving word case. Grammatical features such as prepositional ratio and verb-to-noun ratio, as indicated by Eq. (1), were derived to assess sentence structure and language usage. The 'calculate_content_density' function, defined by Eq. (2), determined the ratio of content words to total words, reflecting sentence informativeness [8]. The distribution od the POS tags is shown in Fig. 5.

$$\text{verb\_to\_noun\_ratio} = \text{num\_verbs}/\text{num\_nouns}, \tag{1}$$

$$\text{CD} = (\text{verbs} + \text{nouns} + \text{adjectives} + \text{adverbs})/\text{words}. \tag{2}$$

Additionally, a 'calculate_pos_ratios' function provided insights into noun and verb distribution. Results from these calculations were integrated into a new array, encompassing TTR, noun ratio, verb ratio, prepositional ratio, content density, and verb-to-noun ratio. This array, combined with Word2Vec and GloVe embeddings, Universal Sentence Encoder-based sentence embeddings, TF-IDF Vectorization, and other features, formed a comprehensive feature matrix capturing linguistic, structural, and semantic information.

4 Result and discussion

The train-test split applied after that, dividing data into training and testing sets using the 'Label' column as the target variable. A parameter of test_size = 0.2 reserves 20% for testing and 80% for training, with shuffle = True and random_state = 42 ensuring randomness and reproducibility**.**

The results are carried out on different state of the art machine learning algorithms such as SVM, RF, KNN, ANN.

Support vector machine (SVM): Support Vector Machine is used for classification, finding a hyperplane to separate classes. SVM classifiers, with an RBF kernel and balanced class weights, are employed. Cross-validation assesses accuracy, and the model is evaluated on the test set, providing a comprehensive classification report.

K-nearest neighbors (KNN): KNN is an intuitive algorithm for classification. A model with neighbors set to 5 is created, balancing local trends and computational efficiency. After fitting to the training set, accuracy is calculated, and a classification report is generated on the test set.

Table 2 Result table for all algorithms with word2vec

Algorithm	Accuracy	Precision	Recall	F1-score
SVM	97.41	97	98	97
KNN	98.7	98	99	98
Random forest	99.35	99	99	99
Algorithm	Val_loss	Val_acc	Loss	Test Accuracy
ANN	1.66	99.68	1.66	99.67

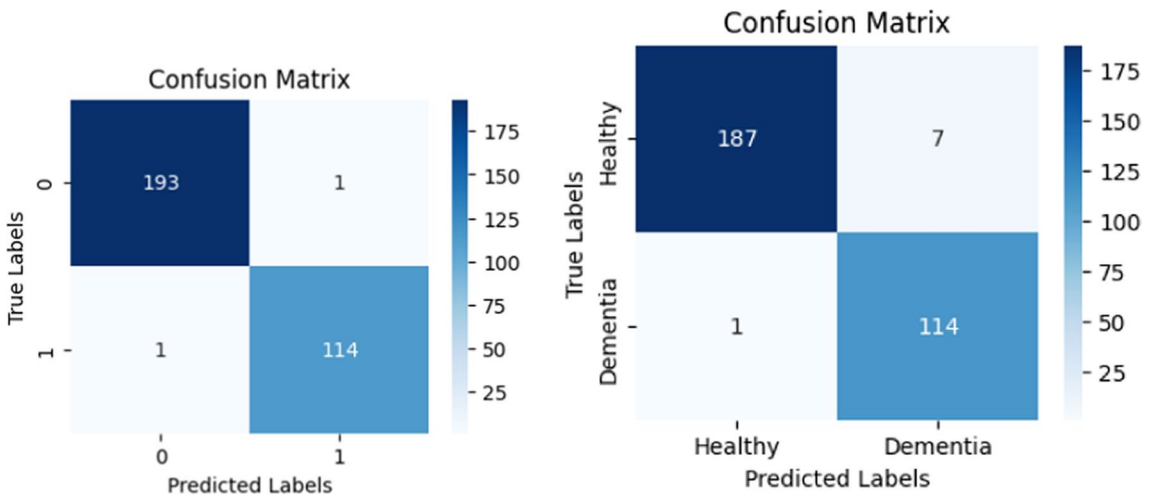

Fig. 6 Confusion matrix for all the algorithms with word2vector

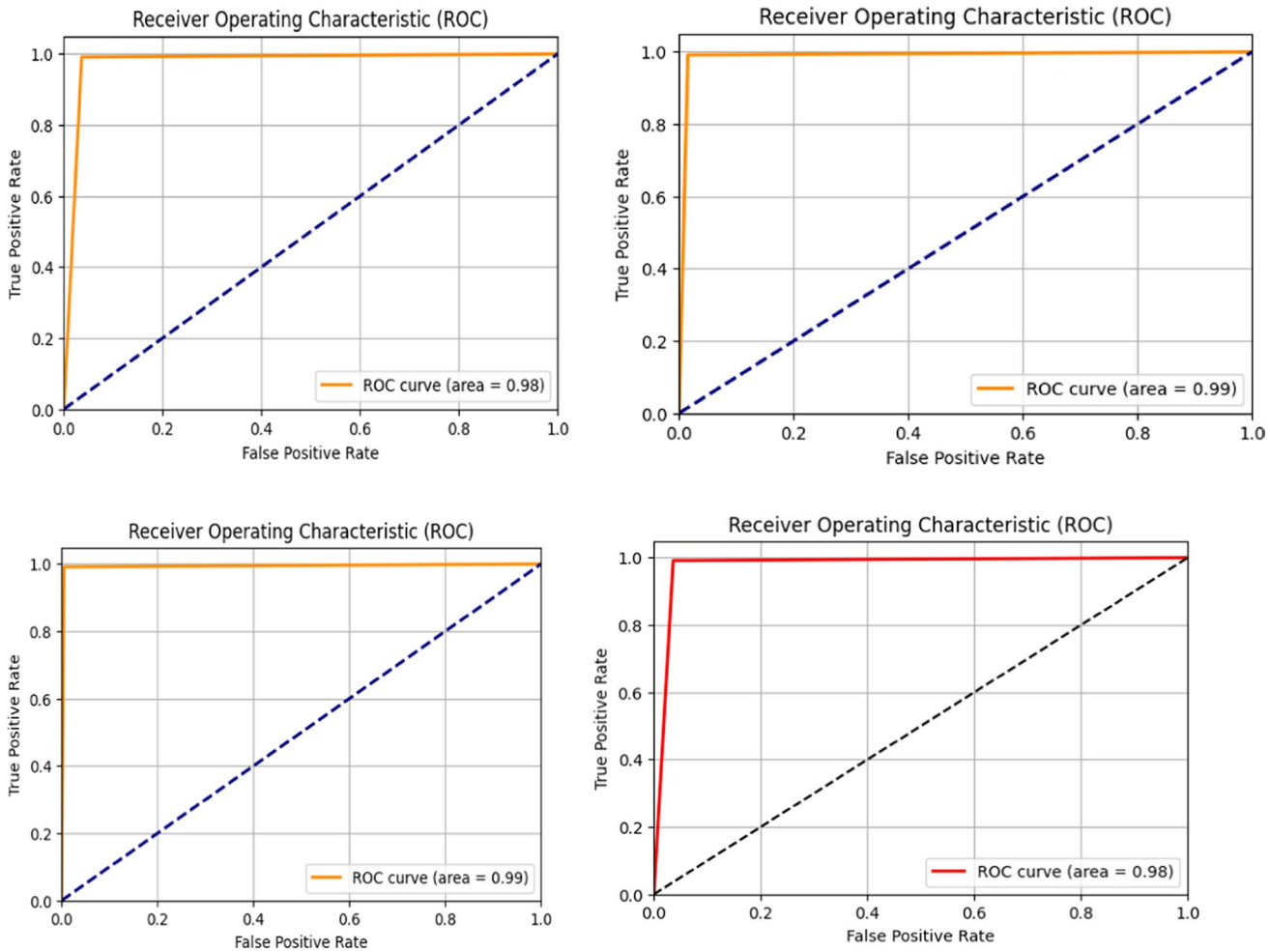

Fig. 7 ROC for all the algorithms with word2vector

Table 3 Result Table for all algorithms with glove

Algorithm	Accuracy	Precision	Recall	F1-score
SVM	97.41	97	96	97
KNN	98.70	98	98	99
Random forest	99.02	99	99	99
Algorithm	Val_loss	Val_acc	Loss	Accuracy
ANN	2.39	98.71	2.39	98.70

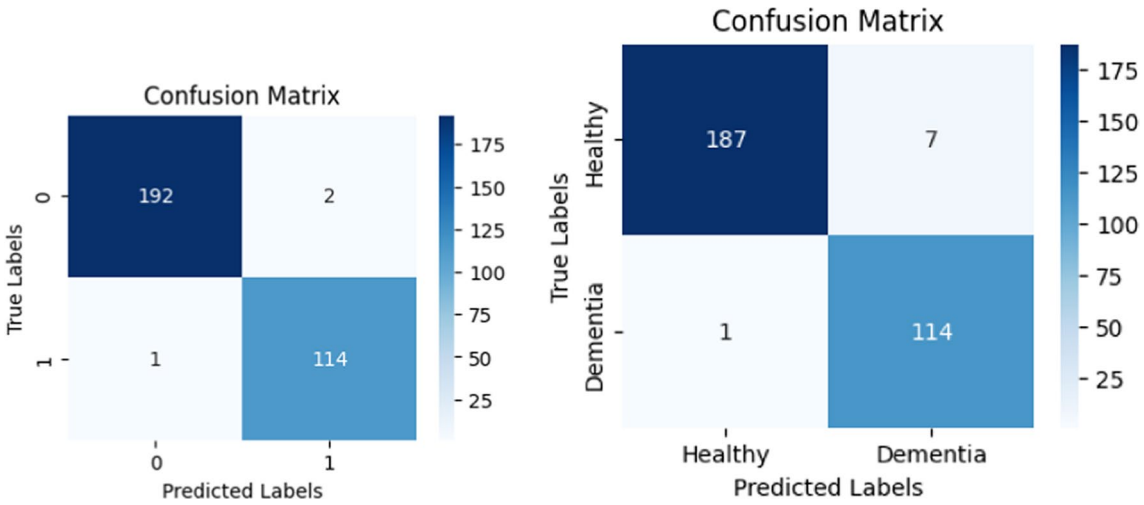

Fig. 8 confusion matrix of all algorithms with Glove

Random forest (RF): Random Forest, an ensemble method, is used for classification. The code handles class imbalance using `class_weight = 'balanced'`. After fitting the full training set, predictions are made on the test set. Accuracy calculations and a classification report assess the model's performance.

Artificial neural networks (ANN): ANNs, inspired by the brain's structure, are created with three layers. ReLU activation mitigates the vanishing gradient problem. The model is compiled with 'adam' optimizer and 'binary_crossentropy' loss. Trained for 5 epochs, it's evaluated on the test set, providing accuracy as a measure of performance on unseen data.

A cross-validation technique with 5 folds is employed to robustly assess every classifier's performance. It aids in detecting overfitting or underfitting, guides hyperparameter tuning, and ensures reliable performance metrics for trustworthy classification models.

4.1 Evaluation

To gain an understanding of the outcome, evaluation procedures such AS accuracy, precision recall, F1 score, confusion matrix, and roc curve were applied. Accuracy, Precision, recall, and the F1 score are critical metrics used to evaluate the performance of classification models. They are especially relevant when dealing with imbalanced datasets, where one class significantly outnumbers the other. These metrics provide a more nuanced understanding of model performance beyond accuracy [11].

$$\text{Accuracy} = (\text{TP} + \text{TN})/(\text{TP} + \text{TN} + \text{FP} + \text{FN}). \tag{3}$$

$$\text{Precision} = \text{TP}/(\text{TP} + \text{FP}). \tag{4}$$

$$\text{Recall} = \text{TP}/(\text{TP} + \text{FN}). \tag{5}$$

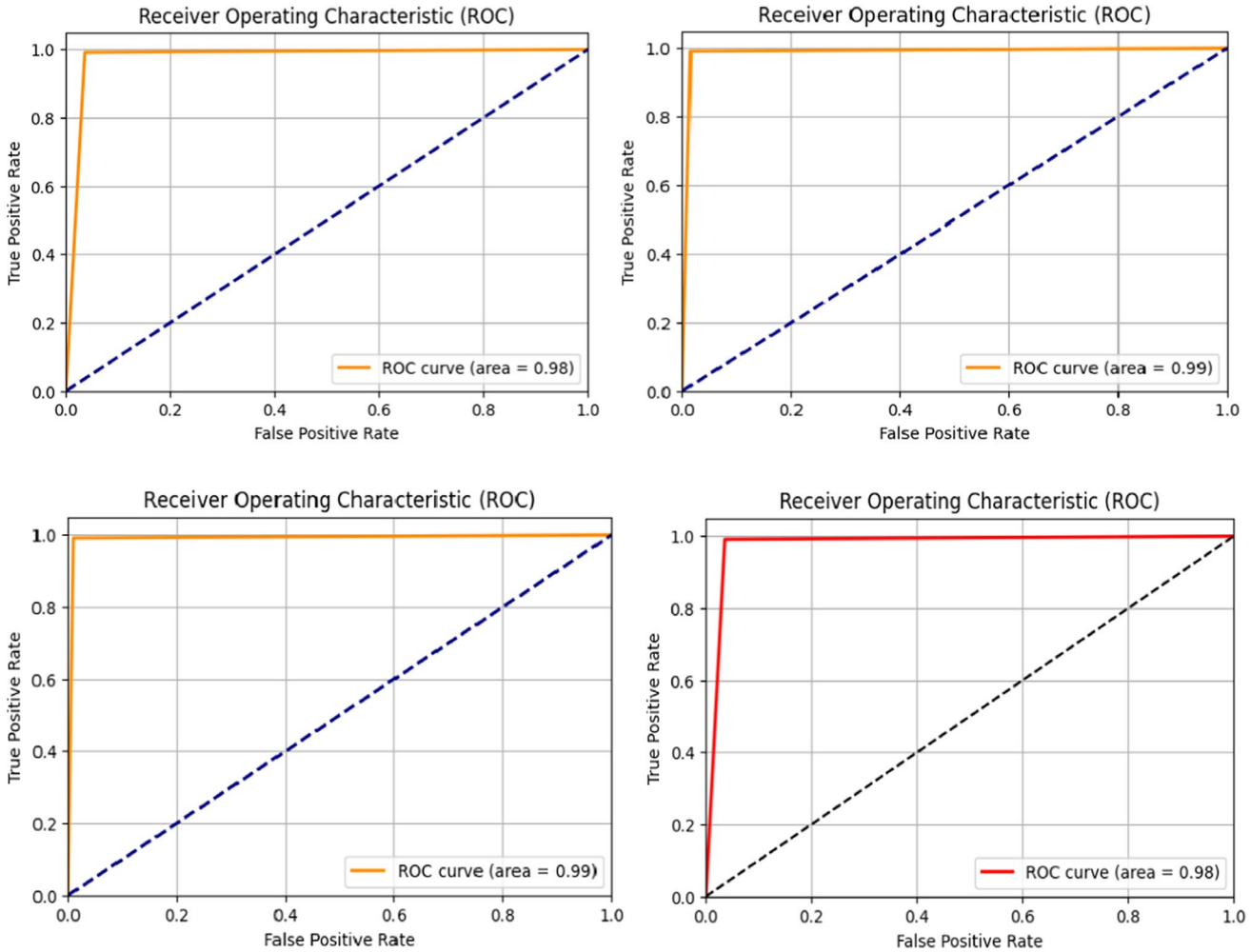

Fig. 9 Roc of all algorithms with glove

Table 4 Result Table for all algorithms with sentence embedding

Algorithm	Accuracy	Precision	Recall	F1-score
SVM	97.41	97	98	97
KNN	98.70	98	99	99
Random forest	98.05	99	97	99
Algorithm	Val_loss	Val_acc	Loss	Accuracy
ANN	2.31	99.35	2.31	99.35

$$F1 = (2 \cdot \text{Precision} \cdot \text{Recall})/(\text{Precision} + \text{Recall}). \quad (6)$$

Both the confusion matrix and the receiver operating characteristic (ROC) curve are essential tools for evaluating model performance and offering insightful information on its behaviour in the field of classification modelling. Each of these evaluation techniques are shown in three different parts as the analysis is conducted on word2vec, glove, and sentence embedding. All of these are Word frequency, TF-IDf, hesitation count, and linguistic traits are the four features that remain consistent across all sections. With these, a fifth feature—different forms of embeddings such—will be added one at a time to differentiate how embedding affects other features and the model as a whole.

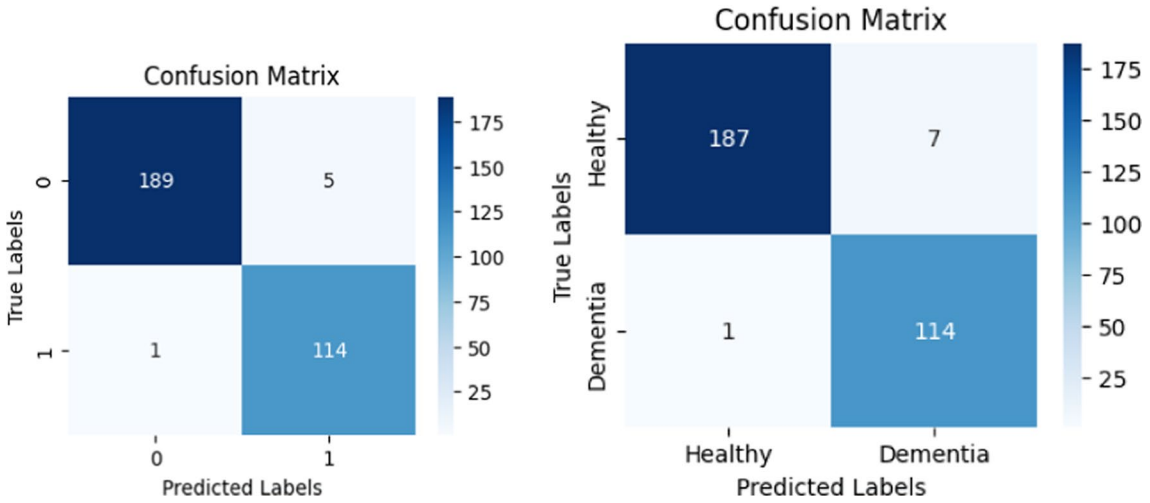

Fig. 10 Confusion matrix of all algorithms with sentence embedding

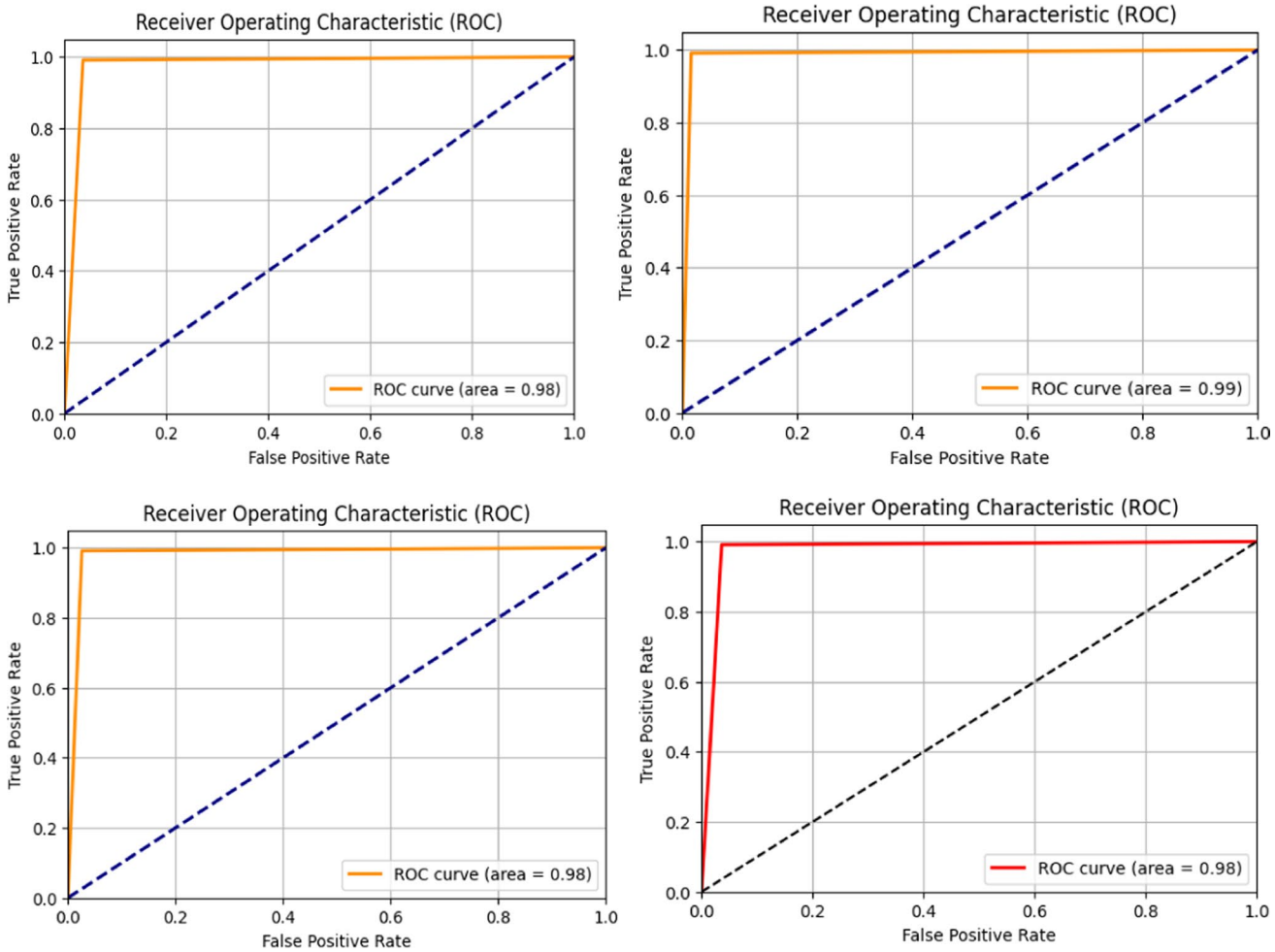

Fig. 11 Roc of all the algorithms with sentence embedding

4.1.1 With Word2Vec

Building on Word2Vec, I explore the application of Word Embeddings with Word2Vec, a technique that extends word embeddings to the entire vocabulary. This technique aims to capture contextual information and semantic meaning in a broader context. The combination of Word2Vec embeddings with the constant features at a word- level granularity enhances our understanding of the features' influence within the context of surrounding words. Insights may include more nuanced feature interactions, context-specific patterns, and improved feature representations. These findings could lead to a deeper understanding of how word embeddings affect the analysis of the constant features. The results are shown in Table 2.

The confusion matrix is shown in Fig. 6 and the ROC curve is shown in Fig. 7.

4.1.2 With glove

Glove (Global Vectors for Word Representation) is another word embedding technique that focuses on capturing global word co-occurrence statistics. Unlike Word2Vec, which is based on local context, Glove embeddings consider the entire corpus to generate word representations. The integration of Glove embeddings with the constant features allows us to explore the impact of global context on feature analysis. Results may reveal different patterns, such as how word co-occurrence statistics influence the constant features. Insights may include the effectiveness of Glove embeddings in capturing semantic relationships and how this affects feature interpretation. The results are shown in Table 3.

The confusion matrix is shown in Fig. 8 and the ROC curve is shown in Fig. 9

4.1.3 Sentence embeddings

Sentence embeddings represent entire sentences or documents as vectors, summarizing their content and meaning. They are used to capture the overall context and semantic information of text data. The inclusion of sentence embeddings with the constant features enables us to analyse features at a higher level of abstraction—

entire sentences or documents. Insights may include the ability of sentence embeddings to capture document-level semantics, identify similarities or differences between documents, and reveal overarching themes or topics in the data. The results are shown in Table 4.

The confusion matrix is shown in Fig. 10, the ROC curve is shown in Fig. 11.

4.1.4 Comparative discussion

This discussion is done considering all the result tables, confusion matrix and ROC given in Tables 2, 3, 4 and Figs. 6, 7, 8, 9, 10, 11. Heres the detailed discussion.

Word embedding techniques: Word2Vec and GloVe both performed exceptionally well across all machine learning algorithms. Sentence embeddings also delivered strong results, demonstrating their effectiveness in capturing semantic information from text.

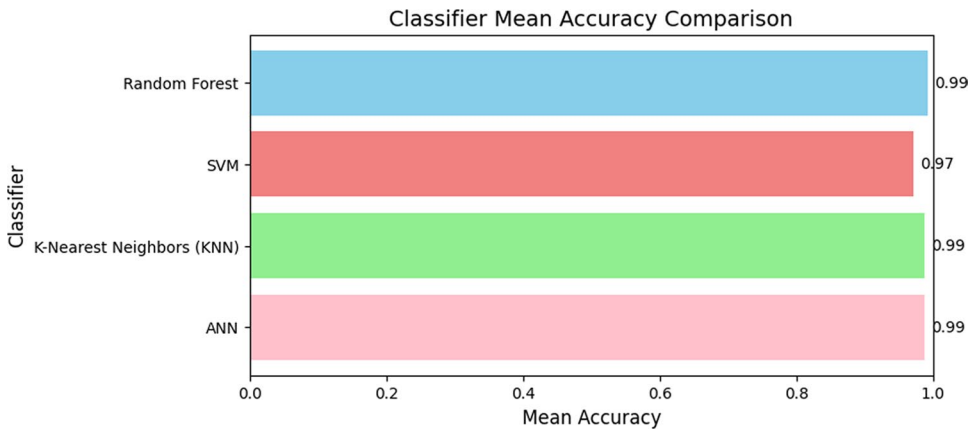

Fig. 12 Accuracy of the model for all the algorithms

Machine learning algorithms: SVM, KNN, and Random Forest consistently delivered strong performance with all three embedding techniques. Neural networks (ANN) also performed well, suggesting that they can effectively leverage different embeddings.

Accuracy: Random Forest consistently achieved the highest accuracy across all embeddings, while KNN and ANN also consistently delivered high accuracy.

Precision, recall, and F1-score: Most models achieved balanced precision, recall, and F1-score values, indicating their ability to classify both positive and negative cases effectively.

Choice of embedding: The choice between Word2Vec, GloVe, or sentence embeddings may depend on factors like dataset size, computational resources, and specific domain requirements. GloVe and Word2Vec are both strong options, with GloVe showing a slight preference for precision, while Word2Vec excels in recall.

The accuracy is shown in Fig. 12.

A total depiction of accuracy for all the algorithms is given in Fig. 11.

5 Conclusions

This paper outlined the importance of text analysis in the contemporary world. The pervasiveness of textual information and its wide range of uses, including sentiment analysis, document classification, and recommendation systems, was emphasized. The trip was embarked upon with a clear goal to understand its complexities since the potential for text data to influence decision-making processes was acknowledged. The arena of machine learning models, each a powerful instrument in the orchestration of text categorization in dementia diagnosis, was entered with enriched characteristics. To negotiate the complexities of the text data, Support Vector Machines (SVM), K-Nearest Neighbours (KNN), Random Forest, and Artificial Neural Networks (ANN) were used. Good accuracy, precision, recall, and F1-score values were consistently delivered by these algorithms, demonstrating their ability to recognize patterns in the textual landscape. An in-depth examination of our findings has been conducted. The accuracy, recall, and F1-score measurements provided a more nuanced view of model performance, which was critical in scenarios involving imbalanced datasets or specific trade-offs. The confusion matrices were thoroughly studied, shedding light on the models' ability to produce correct predictions while minimizing errors. Furthermore, the transformative impact of word embeddings on feature analysis was revealed, with Word2Vec, GloVe, and sentence embeddings, along with the other four significant linguistic characteristics, all providing new insights. The remarkable performance of machine learning algorithms was also lauded, with Random Forest coming out on top. In the future, this work on dementia detection through text analysis opens several exciting possibilities and avenues for further research and development. These include expanding the dataset for increased diversity and size, exploring multimodal approaches, hybrid deep learning algorithms that combine text and speech data, conducting longitudinal studies to track language changes over time, and adopting advanced machine learning models. Real-time monitoring, ethical considerations, clinical integration, cross- cultural research, interdisciplinary collaboration, and public awareness initiatives are also promising directions for future work.

Author contributions Conceptualization: Z.J., S.B.K. and M.S.Methodology: Z.J. and S.B.K.Data Curation: S.B.KFormal Analysis: Z.J., S.B.K. and M.SWriting original draft: .Z.J.Visualization and Validation: .Z.J.Review and Editing: S.B.K. and M.S.Supervision: S.B.K.

Data availability The data is available in the repository and is accessible from Dementia.talkbank.org on request.

Code availability The code is available on request.

Declarations

Competing interests The paper declares no competing interests.

References

1. World Health Organization. Dementia. In: World Health Organization. 2023. https://www.who.int/news-room/fact-sheets/detail/dementia. Accessed 08 Dec 2023.
2. Zheng C, Bouazizi M, Ohtsuki T. An evaluation on information composition in dementia detection based on speech. IEEE Access. 2022;10:92294–306. https://doi.org/10.1109/ACCESS.2022.3203068.
3. Comuni F. A natural language processing solution to probable Alzheimer's disease detection in conversation transcripts. Semantic Scholar. 2019. https://www.semanticscholar.org/paper/A-natural-language-processing-solution-to-probable-Comuni/38957bc3d4a25d82bc9f4ea3a7716d78a081b0be. Accessed 08 Dec 2023.
4. Pulido MLB, Hernández JBA, Ballester MÁF, González CMT, Mekyska J, Smékal Z. Alzheimer's disease and automatic speech analysis: a review. Expert Syst Appl. 2020;150: 113213. https://doi.org/10.1016/j.eswa.2020.113213.
5. Nambiar AS, Likhita K, Pujya KVSS, Gupta D, Vekkot S, Lalitha S. Comparative study of deep classifiers for early dementia detection using Speech Transcripts. IEEE Xplore. 2022. https://doi.org/10.1109/INDICON56171.2022.10039705.
6. Zhu Y, Obyat A, Liang X, Batsis JA, Roth RH. WavBERT: exploiting semantic and non-semantic speech using Wav2vec and BERT for dementia detection. Interspeech. 2021. https://doi.org/10.21437/interspeech.2021-332.
7. Burke E, Gunstad J, Pavlenko O, Hamrick P. Distinguishable features of spontaneous speech in Alzheimer's clinical syndrome and healthy controls. Neuropsychol Dev Cogn. Sect B Aging Neuropsychol Cogn. 2023;1–12. https://doi.org/10.1080/13825585.2023.2221020.
8. Guerrero-Cristancho JS, Vásquez-Correa JC, Orozco-Arroyave JR. Word-embeddings and grammar features to detect language disorders in Alzheimer's disease patients. TecnoLógicas. 2020;23(47):63–75. https://doi.org/10.22430/22565337.1387.
9. DementiaBank. (n.d.). In: Dementia.talkbank.org. 2023. https://dementia.talkbank.org/. Accessed 08 Dec 2023.
10. Orimaye SO, Wong JS-M, Golden KJ, Wong CP, Soyiri IN. Predicting probable Alzheimer's disease using linguistic deficits and biomarkers. BMC Bioinformat. 2017;18(1). https://doi.org/10.1186/s12859-016-1456-0.
11. Kanstrén T. A Look at Precision, Recall, and F1-Score. Medium. 2020. https://towardsdatascience.com/a-look-at-precision-recall-and-f1-score-36b5fd0dd3ec. Accessed 08 Dec 2023.