Computing and Informatics, Vol. 41, 2022, 1589–1624, doi: 10.31577/cai_2022_6_1589

COMPUTATIONAL INTELLIGENT MODELS FOR ALZHEIMER'S PREDICTION USING AUDIO TRANSCRIPT DATA

Yusera Farooq Khan*, Baijnath Kaushik

School of CSE, Shri Mata Vaishno Devi University
Katra, Jammu and Kashmir, India
e-mail: {18dcs006, baijnath.kaushik}@smvdu.ac.in

Bilal Ahmed Mir

Graduate School of Science and Engineering
University of Toyama, Toyama 930-8555, Japan
e-mail: mirb940@gmail.com

Abstract. Alzheimer's dementia (AD) is characterized by memory loss, which is one of the earliest symptoms to develop. In this study, we investigated audio transcript data of patients with Alzheimer's dementia. The study involved the use of three intelligent computational approaches: conventional machine learning (Support Vector Machine, Random Forest, Decision Tree), sequential deep learning (LSTM, bidirectional LSTM, CNN-LSTM), and transfer learning (BERT, XLNet) models for automatic detection of linguistic indicators for early diagnosis of Alzheimer's dementia. These models were trained on the DementiaBank clinical transcript dataset. The grid search tuning approach is used for tuning the values of the hyperparameters. Text vectorization is done using the Term Frequency-Inverse Document Frequency (TF-IDF) information retrieval approach. TF-IDF is based on the Bag of Words (BoW) paradigm, which deals with the less and more relevant words in a transcript. Results were evaluated and compared using several performance metrics. The state-of-the-art techniques implemented on DementiaBank dataset in our methodology achieved better performance in terms of accuracy. Transfer learning models showed better classification results in comparison to sequential deep learning models. However, sequential deep learning models outperformed traditional

* Corresponding author

machine learning models. Overall, in terms of accuracy, BERT and XLNet were the most accurate, with accuracy of 93 % and 92 %, respectively.

Keywords: Dementia, memory loss, transcript data, deep learning, convolutional neural network, Bi-LSTM

1 INTRODUCTION

The most prevalent form of dementia is AD. Around 50 million people worldwide have AD, and other forms of dementia are anticipated to triple in prevalence within 30 years because of the ageing population [1]. AD accounts for 60 % to 70 % of all dementia cases, affecting 1 in every 14 people over the age of 65 and one in every 6 people over the age of 80 [2]. AD is incurable and cannot be healed or reversed. However, a medicine may be used to delay or stop the progression [3]. Cognitive impairment, which may include difficulties with word retrieval and impaired reasoning or decision-making, is one of the initial symptoms associated with Alzheimer's disease. Memory loss is one of the initial symptoms, followed by problems with language use, and everyday tasks, and, in more advanced stages, struggles with simple bodily functions such as walking. Neurons in other areas of the brain are affected and killed as the disease progresses [4]. Once an essential part of an individual's daily life, those activities that may be out of reach, such as organizing family gatherings or engaging in sports, is now out of the question. The functional areas of the brain eventually degenerate. Finally, the patient loses full control and becomes dependent on the continuous attention of a caregiver [5].

1.1 Speech Difficulties in Alzheimer's Disease Patients

Cognitive impairment is a direct and inevitable result of language difficulties, making it one of the most recognisable symptoms of AD. During patient-neurologist interactions, language and communication deficiencies are visible, and interactional remarks can be employed to distinguish between cognitive difficulties produced by neurodegenerative disorders and cognitive difficulties caused by functional memory disorders [6]. Prior research has identified that AD has a major effect on the speech signal, and numerous methods for detecting AD using only speech or spoken text information have been published. Typical AD memory disorder causes all of these problems [7]. For example, it is possible that the initial indicators of AD will be word retrieval issues, which can present themselves in changes in a variety of language elements such as verbal naming, density and amount of words, correct meaning communication, pauses, and speaking pace. Word recovery is assessed frequently with the help of image description tasks, in which participants are guided to explain what they see. These tasks enable the evaluation of the lexical and syntactic complexity, which is also declining in dementia, in addition to word retrieval [8]. Memory

deficiency also leads to repetitive terms and concepts, leading to communication mistakes, less consistency, and density of knowledge [9].

1.2 Motivation

Alzheimer's speech processing using machine learning (ML) and natural language processing (NLP) motivates this study. There is no cure for AD, but the progression can be delayed and, in some cases, halted by treatment if diagnosed early. The pathology of AD most likely starts several years, if not decades, before symptoms appear. As a result, there is a chance for prevention if potential developments can diagnose the disease using linguistic biomarkers before symptoms appear. Narratives are analysed to compare the language abilities of persons with and without Alzheimer's disease. This results in investigating the link between cognitive and language abilities, and to establish an early predictor. Through the use of a mobile application, people with limited access to medical care will be able to screen for early indicators of dementia. While these innovations will be useful, they are still in progress and are not yet available to the public. The motivation of this work is to implement the state-of-the-art techniques for automatic speech analysis to monitor Alzheimer's disease patients and to provide light on prospective future research issues. In the proposed study, a comprehensive analysis has been done to test and validate the efficiency of conventional machine learning models (SVM, Decision Tree and Random Forest), sequential deep learning models (LSTM, Bi-LSTM and CNN-LSTM) and with pre-trained transfer learning models (BERT and XLNet) for Alzheimer's prediction.

1.3 Contributions

In automatic detection of Alzheimer's dementia with computational intelligence models and NLP, promising results were achieved. Potentially, automated language processing could lead to an efficient and non-intrusive way of detecting clinical problems and the accessibility and affordability of dementia testing. The following is a brief description of the article's main contributions:

- This study evaluated the comparative performance of three computational intelligence approaches for the classification of AD and non-AD by identifying linguistic patterns.
- First, we investigated the role of conventional ML models: Decision Tree, Random Forest, and Support Vector Machine for the early detection of linguistic characteristics of Alzheimer's patients.
- Furthermore, we investigated the performance of sequential DL models: LSTM, Bi-LSTM, and a hybrid of CNN-LSTM for automatic detection of linguistic indicators of cognitive memory loss in AD.
- Additionally, we explored the significance of two TL models: XLNet and BERT for predicting AD.

- Two embedding techniques like TF-IDF and pre-trained embedding are used. The objective is to test the efficiency and robustness of the models used in the study with a corpus of training, test and validation data.
- Also, the efficiency and accuracy of the models used in this study were compared with state-of-the-art existing models to show that our approach has better and more robust performance in comparison to the earlier models.
- Further, the performance of the models in the study was tested with statistical validation technique (cross-validation) and has shown that the performance of the models was consistent over all folds of training and test sets.

2 VOICE PROCESSING AND ALZHEIMER'S DEMENTIA

Language evaluation has seen an increase in the usage of automated speech signal processing techniques in recent years, particularly in diagnosing cognitive pathologies [10, 11]. Such techniques can be used to identify signal features that are essential for diagnosing certain disorders. Subsequently, using intelligent computational techniques, the process of sample classification is carried out in accordance with the prior findings achieved. The most often deployed and traditional aspects of AD detection from audio transcripts are linear, as they are the most simply interpretable clinically. Other recent and innovative methods, on the other hand, have included non-linear characteristics. Both of these characteristics are significant markers of the language's expressive architecture and are dependent on the activities accomplished. Intelligent computational techniques are used in this work to classify speech features in addition to the feature extraction procedure, which uses statistical analysis, and mathematical models. Several of these are conventional machine learning (ML) [12], sequential deep learning (DL) [13] and transfer learning (TL) [14, 15, 16] models, each of which has distinct architecture and operating features.

2.1 NLP in Computer-Aided Diagnosis

The outbreak of disease diagnostic has called attention to the enhancement of remote diagnosis systems and early detection of diseases, and the discovery of new antidotes and drugs. Today, the healthcare industry is expanding at a speed never seen before. In this scenario, the role of NLP becomes extremely prominent in utilizing digital data (e.g., Electronic Health Records (EHR) [17], Weibo User Depression Detection Data Set [18], etc.) to detect serious diseases at the earliest. These data are primarily unstructured and are time-consuming to determine the stage of disease. An NLP model can help to remove irrelevant text and highlight the medical keywords along with their numeric values, if any, to provide a quick summary of the diagnostic report. This will save time in identifying the stage of a patient from a pile of text. NLP in computer-aided diagnosis has high susceptibility for enhanced medical decision-making [19].

3 RELATED WORK

Earlier work on language-based AD detection focused primarily on hand-crafted features extracted from transcripts, with some acoustic data included. In König et al. study, data was collected when participants completed a series of brief cognitive vocal tasks [11]. Detecting initial vocal cues (such as short answers, repeated requests for clarification about the past, and starting with interjections) was accomplished using the speech processing methods. In addition, the vocal markers were examined for their "ability" to discriminate between patients with healthy control (HC), mild cognitive impairment (MCI), and AD. Second, the vocal indicators were tested for their "ability" to differentiate between HC, MCI, and AD. Automated audio analysis was able to distinguish between HC and MCI by 79 % ± 5 % and HC with AD by 87 % ± 3 %. Orimaye et al. used the DementiaBank language transcript clinical dataset, which included 99 patients with suspected AD and 99 HC, to construct machine learning models [20]. The models discovered a variety of syntactic, lexical, and n-gram linguistic biomarkers that could be used to differentiate the likely AD group from the healthy group. They found that people who might have AD used a lot less syntactic parts of their language and a lot more lexical parts of their language. Fraser and his team used the recordings of 264 people who talked about the Cookie Theft picture from the DementiaBank corpus. To find out how accurate the automation classification from healthy to AD was, machine learning techniques were used. In this case, the standard accuracy of 81 % was achieved [21]. Clark et al. studied the data of 107 people with MCI and 51 HC for the study. The tests were transcribed, and linguistic characteristics were extracted, comprising raw word count, intrusions, repeats, groups, and shifts, mean word frequency, average sentence frequency, and algebraic connection. The study combined linguistic dimensions with data from MRIs, allowing for the creation of novel results by achieving an accuracy of 83.20 %. In the analysis, the classifiers trained (Novel + Brain) in new ratings outperformed those trained in rough rating [22]. Karlekar et al. [23] LSTM-RNNs, and their combinations were employed in three neural models based on CNNs and LSTM-RNNs to distinguish between AD and control patient language samples. The accuracy of the CNN, LSTM, and CNN-LSTM models was 82.8 %, 83.7 %, and 84.9 %, respectively. Verbal fluency (VF), spontaneous speech (SS), and other tasks are the three main types of language tests used to determine many health issues associated with AD [24]. Meghanani et al. [25] experimented with fastText, and CNN models with a single convolutional layer were used to extract n-grams from the input phrase while initializing word embeddings with GloVe vectors. CNN and fastText text models reach 79.16 and 83.33 % accuracy for predicting AD. JabaSheela et al. evaluated AD patients' language using DL. AD and CN transcripts were used to train neural networks. CNN and CNN+bidirectional LSTM were compared where CNN + bidirectional LSTM obtained 72 % accuracy in experiments [26]. Sarawgi et al. [27] evaluated ADReSS dataset using multimodal inductive transfer learning with temporal features to detect AD and its severity. 83.3 % accuracy was recorded in their study, which was further evaluated on the Pitt database with an accuracy of

88 %. Balagopalan et al. studied the usefulness of speech transcript representations derived from more clinically applicable language feature-based approaches and TL models (e.g., BERT). Feature-based techniques and fine-tuned BERT models worked well with a small set of linguistic variables, indicating the necessity of comprehensive linguistic information for identifying AD cognitive deficits. Fine-tuned BERT models detected AD with 85.14 % accuracy [28]. A typical method of attempting SS is to ask participants to explain an image or to engage in conversation with the participants. Also, it could be employed to recall a movie, a day, a case, or a dream. These activities can be used to examine various linguistic features, including word retrieval ability, syntactic and semantic difficulties, and communication errors.

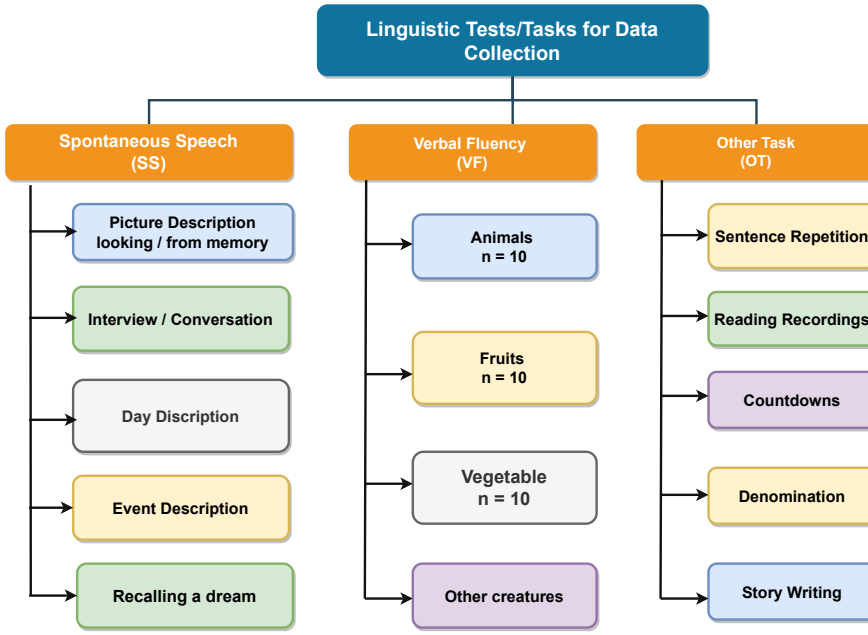


Figure 1. Categories of language exams for AD prediction

In VF, one minute is allotted for participants to come up with a list of terms beginning with the letter F. Typically, success in fluency exams has been determined by an estimate of the number of right words generated in one minute. A wide range of measurements falls outside the boundaries of either the short-term or long-term memory tests. Using these tasks, researchers can study many aspects of cognition, semantic processing, and linguistic and auditory processes. Datasets related to Alzheimer's dementia contain audio or video recordings of these tasks that were transcribed and used in the literature. A list of the methods and tasks that are used to acquire language and speech data is shown in Figure 1. In Figure 1, "n" represents the number of words participants are asked to name in fluency tasks

within 1 minute that are either from the same semantic category or begin with the same letter. The linguistic data of all these tests were recorded on video or audio and then converted into transcription.

4 METHODOLOGY

In this section, the methodology for the prediction of AD is discussed. The outline of the framework proposed in this study is shown in Figure 2. The abbreviations used in Figure 2 are listed in Table 1.

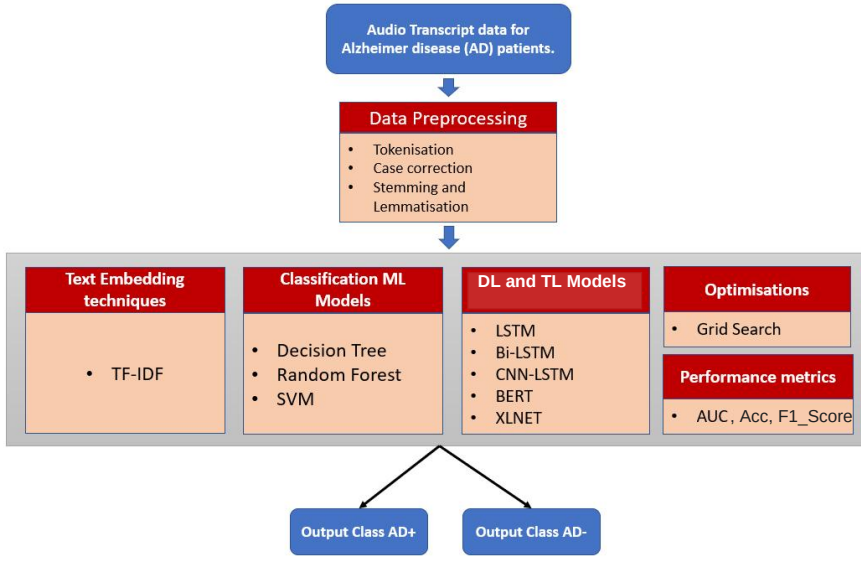


Figure 2. Proposed methodology framework

S. No.	Abbreviation	Full Form
1.	TF-IDF	Term Frequency-Inverse Document Frequency
2.	SVM	Support Vector Machine
3.	LSTM	Long short-term memory
4.	Bi-LSTM	Bidirectional Long short-term memory
5.	CNN-LSTM	Convolutional Neural Network – Long short-term memory
6.	BERT	Bidirectional Encoder Representations from Transformers
7.	AUC	Area Under the Curve
8.	Acc	Accuracy

Table 1. Description of abbreviations used in Figure 2

4.1 Dataset Description

DementiaBank is a research corpus that collects speech and language samples of Alzheimer's dementia and other dementia-related disorders. The dataset contains transcripts of interviews with Alzheimer's patients.

The Pitt Corpus, a dataset containing records of 264 participants explaining the image of Cookie Theft attainable on DementiaBank Corpus. Cookie Theft picture is a popular test for language and cognitive impairment evaluation and was used for all experiments performed in this research. Numerous exercises are documented and transcribed in the DementiaBank dataset. We used transcriptions from a descriptive assignment for both the language model and the bag of words classification. Cookie Theft was an image that participants were asked to interpret, a speech-based medical examination for neurological diseases. The linguistic data were collected either on audio or video in all of these tests and subsequently transcribed. The linguistic features that are identified in the dataset are listed in Table 2 which are further converted into vectors using the vectorization approach. Based on these linguistic features identified, the data is labelled as 0 and 1, representing control normal (CN) and Alzheimer's dementia (AD). Further, the set of data was partitioned into training, testing, and validation data, and several classifiers were used to evaluate the effectiveness of the automatic identification between CN and AD category. The dataset consists of a total of 3272 sample sentences of normal and Alzheimer's patients. Out of these figures, 1676 are from class 0, and 1596 are from class 1. Here, 0 signifies control normal patient (AD−) and 1 signifies Alzheimer's patient (AD+), as shown in Figure 3.

S. No.	Categories of Linguistic features	Clusters of Linguistic Features
1.	Short Answers and Bursts of Speech	Short responses clustered by speech pauses. i.e., "uh", "okay", "s", "um", "hm", "xxx", "oh", "shh".
2.	Repeated Requests for Clarification	The cluster includes clarification inquiries and confusion about the task, specifically in the past tense. i.e., "did I tell?", "Whether that's more than what I said?", "will he hit the bottom ?", "uh everything that's going to happen, huh ?", "does that have enough ?", "I?".
3.	Starting with Interjections	Clusters include utterances that begin with interjections. i.e., "oh", "well", "and", "but", "may be". { "maybe that was an apron and um maybe this was the um", "oh there's a girl & uh reaching(g) for a cookie", ... }

Table 2. Categories of linguistic features identified in dataset

	Transcripts	AD
0	there's &um a young boy that's getting a cooki...	1
1	and it he's uh in bad shape because uh the thi...	1
2	and in the picture the mother is washing dishe...	1
3	and the dishes might get falled over if you don't	1
4	fell fall over there there if you don't get it	1
...	...	...
3267	the mother's standing there doing the dishes	0
3268	she's washing the dishes looking out the open ...	0
3269	and the water's runnin(g) down over the sink o...	0
3270	and <there are> [//] she's dryin(g) a dish	0
3271	summer of the year	0

3272 rows × 2 columns

Figure 3. Samples of AD and non-AD linguistic transcript after pre-processing and class labelling

4.2 Data Pre-Processing

The dataset created initially is labeled as a .tsv file. Each entry of the patient was used as an input. Patients with possible AD were marked as 1, and the entries of control patients were marked as 0. Since the dataset has an approximately equal sample count (both AD+ and AD−), the need for data balancing is insignificant. Apart from this, medical data is very sensitive (patients' personal information) and needs to be processed with domain expertise. In non-medical NLP problems, we usually reduce words to their respective present tenses or stem the words in order to get the original root word.

However, in the case of medical data, the use of such techniques would not be a good idea. It is possible that an Alzheimer's patient may incur grammatical or linguistic mistakes and repeat that pattern in the subsequent sentences. Therefore, the data pre-processing is done using the following steps:

- Removing punctuations,
- Removing left out spaces,
- Removing stop words,
- Removing non-ascii characters,
- Keeping only alphabet characters by removing all the symbols,
- Converting the whole sentence into lower case for uniformity.

4.3 Vectorization Method: Term Frequency – Inverse Document Frequency (TF-IDF)

This approach is employed by the concept that a AD+ and AD− individual would utilise a specific set of words in the phrases. TF-IDF captures those specific set of words and generates vectors putting high weights on those words. TF IDF is a statistical method that is widely used in information retrieval from a collection of documents.

This will represent the frequency of a term (word) in a particular document such that whether it appears multiple times or it is a rare term. The inverse term frequency will tell how dominant a term is in the whole collection of documents [29]. This approach captures the semantic relevance of words by identifying which words are irrelevant and which are important. Each word in the collection will have its TF and IDF value [30]. Common words like "is", "the" will have very high TF but will be penalized by the IDF as this will appear in most of the document.

TF-IDF captures those specific set of words and generates vectors putting high weights on those words. The TF-IDF have several relevant parameters that we may send to the algorithm which will handle the preprocessing. Following are some of the important parameters:

- Stop words: A list of strings can be passed to the algorithm so that these strings will be dropped from actual collections of documents. Typical stop words from English are: "is", "an", "this", "the" etc.
- n-gram range: By defining the boundary of minimum and maximum corresponding n-grams can be extracted.
- Maximum and minimum document frequency: We can set the upper and lower cut-off for the most and least occurring words. Like Maximum document frequency of 80 will ignore all the words that are present in more than 80 % of the document collection.
- Lower-casing: We can opt for all the characters into lowercase only.
- Maximum features: This will allow to select only top maximum features according to the term frequency score.
- TF-IDF parameters: Vector dimensions = 316, max_df = 0.9, min_df = 5.

The step-wise sequence of generating word embeddings for a sentence is demonstrated in Figure 4.

Using the word embeddings approach, we calculated numerical vectors for each pre-processed data point. First, we generated word indexes by converting all of the sample text into sequences. The Keras text tokenizer is utilised to extract these indices. We have assured that the tokenizer does not issue a zero index to any term (because of padding), and we have also adjusted the vocabulary length correspondingly. Following that, each distinct word in the dataset is allocated a unique index, which is utilised to build numeric vectors of all text samples. The vector,

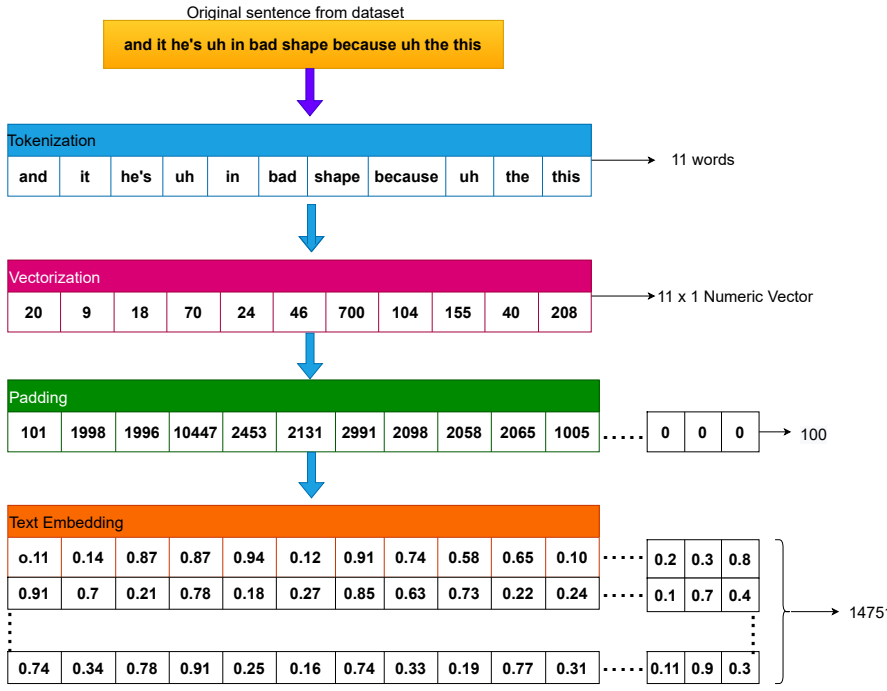


Figure 4. word embedding for a single sentence

or embedding dim in our case, has a length of 186. By doing a search within the vector space, each and every one of the top 14 751 unique words is transformed into vectors. The supplied vector thus has 100-dimensional columns of features with a vocabulary of 14 751, and each vector is filled as a row in the embedding matrix. We built a 25×128 output embedding vector for each tokenized vector using a 128-element embedding layer on our DNNs. In order to reconstruct the linguistic context of words, an EMBEDDING_FILE is created. This creates a vector space with typically several hundred dimensions, where each unique word in the corpus has its own vector.

At last, the data is divided into three sets: Train, Validation, and Test. Train set contains 70 % of the data, and validation set contains 20 %, and the test set has the remaining 10 % of the data. The vectors generated are fed to various models for classification. Below are the statistics shown for three classification approaches.

5 METHODS: COMPUTATIONAL INTELLIGENCE TECHNIQUES

In this study, we implemented three intelligent computational approaches for the classification of Alzheimer's dementia and non-dementia by detecting linguistic indicators of cognitive memory loss.

5.1 First Approach: Machine Learning Models

5.1.1 Decision Tree (DT)

DT is a non-linear supervised and most prevalent algorithm that takes a set of attributes and can map them into target class, value, and probability [31, 32].

The study used criterion = 'gini', min_sample_leaf = 1, min_sample_split = 2, presort = 'deprecated'. The training is done for every instance of the training set. The algorithm also lacks robustness towards noisy data and usually overfits the noise.

5.1.2 Random Forest (RF)

To have a more robust classifier than a decision tree, many ensemble techniques because they are hard to overfit, give better accuracy, and can even estimate missing data. RF [33] is one such model that constructs multiple decision trees, called base models or weak learners in ensemble technique, simultaneously in the training phase, and the majority of the decision class is chosen as the outcome.

In this methodology, various parameters of random forest, such as n_estimators = 100, min_sample_leaf = 1, min_sample_split = 2, criterion = gini, are used. Of all the outcomes, either mean or median was taken depending upon the distribution of the original database.

5.1.3 Support Vector Machine (SVM)

SVM [34] assigns a category to new record on the basis of its decision boundary (Hyperplane) and works well for linearly separable data points, but for non-linear datasets, it needs to transform the data into high-dimensional feature spaces, such that the transformed data become linearly separable [35, 36]. To perform this, SVM classifier uses a mathematical algorithm popularly known as kernel trick. It is a function that eases to work with the given data points without the need to find corresponding data points in the transformed feature space. Parameters set for SVM are: degree = 3, Kernel = 'rbf' and pre_dispatch = $2 * n\_jobs$.

5.2 Second Approach: Sequential Deep Learning Models

In this approach, we used sequence models and sequence + convolution models for training our data. Sequential neural networks [37] are a variant of feed-forward neural networks with a recurrent loop from the previous time step to the next time step for capturing the sequential dependency in the data [38].

5.2.1 Long Short-Term Memory (LSTM)

For capturing the long-term dependencies in sequential data and solving the issue of vanishing gradient problems in recurrent neural networks, LSTMs were intro-

duced [39]. The underlying architecture of LSTM used in this methodology as shown in Figure 5 represents the architecture consisting of input_length = 100, output_dim = 100, we are using input_dim = size of the vocab in the embedding layer, one LSTM layer of 128 unit with 0.2 dropout rate and output dense layer with 1 unit. At each time step $t = T$, LSTM takes input of the cell state/memory state and hidden state from the previous time step $t = T - 1$, and input of the current state.

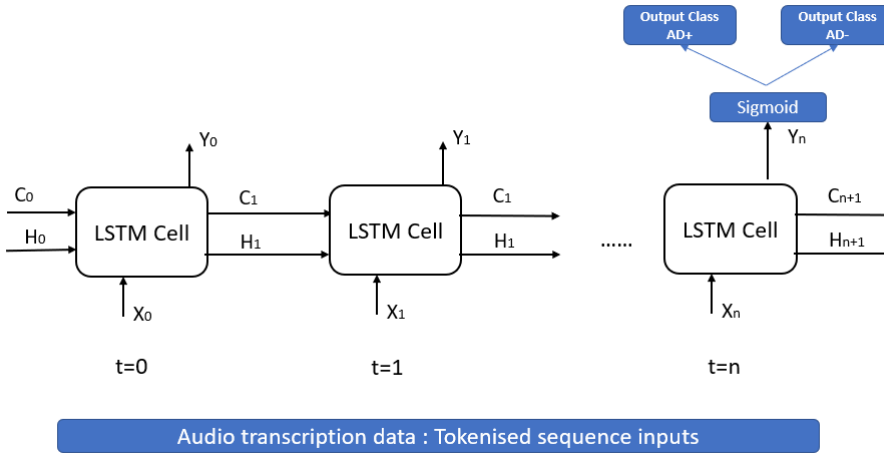


Figure 5. LSTM architecture for the classification of AD+ and AD−

The output gate filters out the information that should be passed to the next LSTM cell in the sequence at time $t = T + 1$.

LSTM model is trained with loss = binary_crossentropy, optimizer = Adam, metrics = accuracy, epochs = 50, batch_size = 256 and validation_split = 0.1. One of the major shortcomings of LSTM is that it only captures unidirectional context in a sentence and therefore a more robust architecture like Bi-LSTM had to be designed.

5.2.2 Bidirectional Long Short-Term Memory (Bi-LSTM)

Improving over the shortcomings of unidirectional LSTM, we experimented with a bidirectional LSTM. Bi-LSTM is a powerful Recurrent Neural Network, able to capture a word's semantics in a sentence in both directions simultaneously [23]. Bidirectional LSTM architecture implemented in this is shown in Figure 6, where we used input length = 100 and output dim = 100; input dim = the size of the vocabulary in the embedding layer; one LSTM layer with 128 units and a 0.2 dropout rate; and a dense output layer with 1 unit. Using the Bi-LSTM model, the loss

was set to binary cross-entropy and used Adam as our optimizer; the metrics are accuracy, epochs are 50, the batch size is 256, and validation split is 0.01.

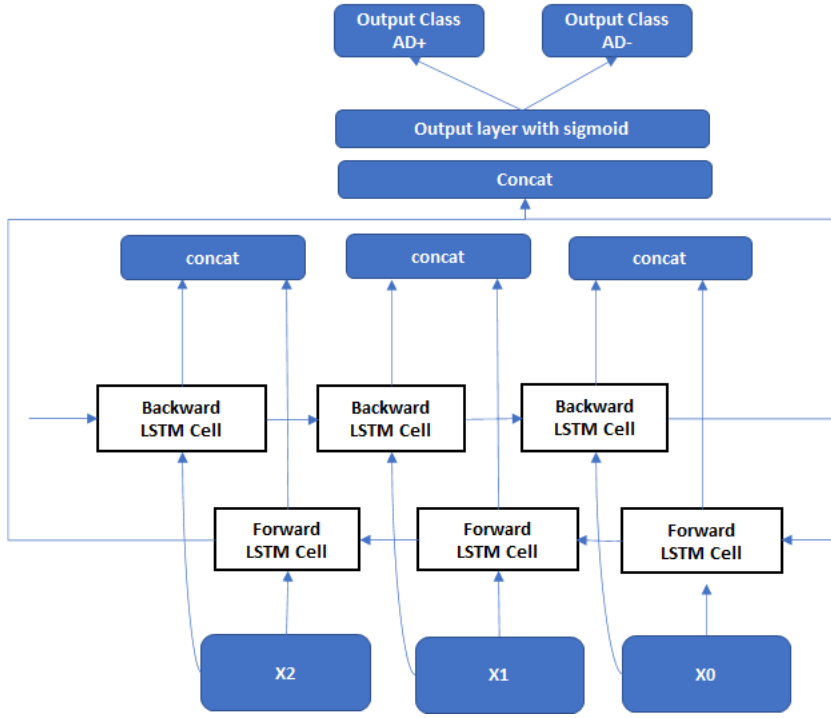


Figure 6. Architecture of a bidirectional RNN with a single hidden layer

The complete text sequence $[X_0, X_1, \ldots, X_n]$ is passed as input with the first token of the sequence in forward LSTM's first cell and the last token of the sequence in backward LSTM's first cell. The classification task is performed by concatenating the output of the last LSTM cells of both the forward and backward LSTM layers and then passing the values through an output layer. In the output layer, a sigmoid function is used as an activation to make the probabilities of both classes of events occurring. Bi-LSTM showed enhanced performance on transcript data than unidirectional LSTM.

5.2.3 Hybrid of CNN-LSTM

Convolutional Neural Network (CNN) has already been proved to be a significant neural network architecture for the extraction of meaningful features from the image data. The CNN model is used for feature extraction from the sequential data and the LSTM model for interpreting the features [40]. This model can be described as an input layer followed by a CNN layer and an LSTM layer and then the dense output

layer, as shown in Figure 7. Hybrid of CNN and LSTM is implemented for AD classification using the embedding layer with input_length = 100, output_dim = 100, input_dim = size of the vocab and dropout rate = 0.1, a convolutional layer with 32 units with pool size = 4, one LSTM layer with 64 units and output dense layer with 1 unit with sigmoid activation function. This hybrid model is trained with loss = binary_crossentropy, optimizer = Adam, metrics = accuracy, epochs = 50, batch_size = 256 and validation_split = 0.1. The result is Max pooled and flattened before passing it to through an LSTM layer for creating a context vector in the final LSTM unit [41, 42]. The context vector passes through a fully connected sigmoid-activated output layer for bi-class predictions.

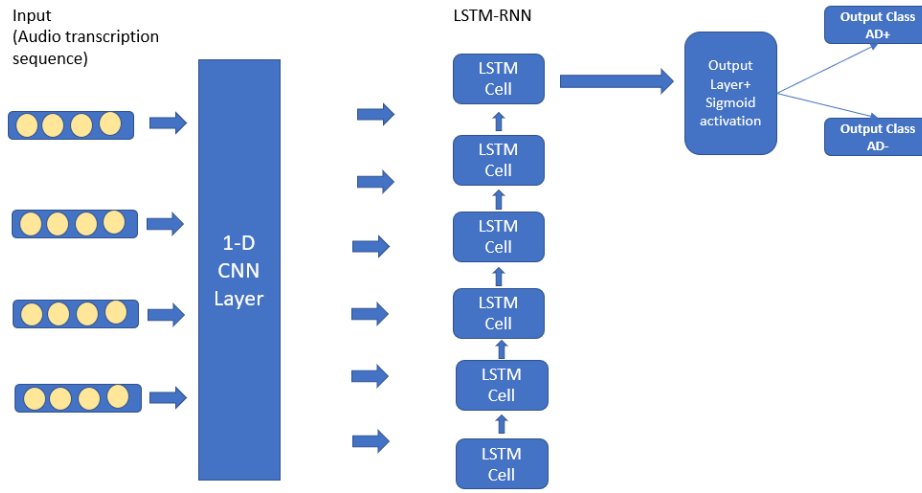


Figure 7. CNN-LSTM hybrid model architecture for classification of AD+ and AD−

5.3 Third Approach: Transfer Learning Models

Transfer learning models have brought a breakthrough in the field of NLP and are continuing to do so [43]. To train transcript data for AD classification, the study employed Bidirectional Encoder Representations from Transformers (BERT) and XLNet. BERT develops a language representation using concurrent Masked Language Modeling (MLM) and together with Next Sentence Prediction (NSP). MLM allows bidirectional training which was not possible in previous methods.

5.3.1 XLNet

XLNet is an autoregressive language model trained using Transformer-XL architecture. XLNet solves the existing shortcomings of BERT. Though BERT learns a language model using MLM, the masked tokens never learn the context with each

other and hence can result in irregular token predictions. We used embedding layers and dropout with learning rate 2e−5.

Yang et al. introduced the XLNet model: Generalized Autoregressive Pretraining for Language Understanding [44]. For all input sequence factorization orders, it uses an autoregressive approach to learn bidirectional contexts. This improves bidirectional linguistic competence and word associations. The context word is used to forecast the following word in an autoregressive model. As a result, the subsequent token is dependent on all preceding tokens. XLNet is generalised because it utilises a method called permutation language modelling to capture bidirectional context. It combines auto-regressive and bidirectional context modelling techniques for the purposes of natural language inference, text analytics, and document evaluation. XLNet learns the true bidirectional context using unsupervised representation learning using autoregressive language modeling. Deep learning concepts like recurrence and attention are combined in the transformer architecture, which enables the model to learn long-term dependencies.

5.3.2 BERT: Bidirectional Encoder Representations from Transformers

Building over the powerful transformer architecture, Bidirectional Encoder Representation from Transformers (BERT) utilizes the encoder block of transformer to learn language representation. It uses the transformer, a system that identifies contextual relationships between words in a text, to accomplish this decoding and encoders are both necessary parts of a transformer in order to make predictions based on the input data. The fine-tuning of BERT is illustrated in Figure 8.

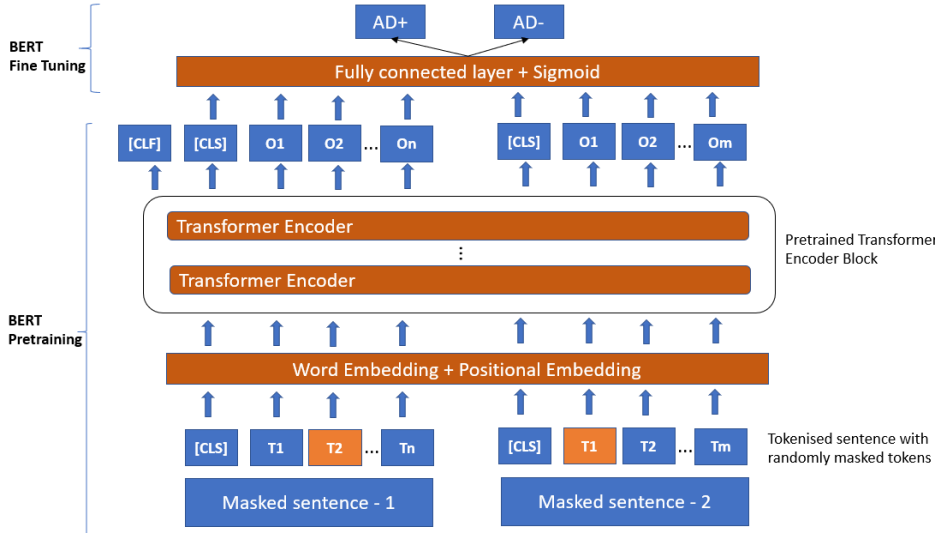


Figure 8. BERT training and fine-tuning shown in a single architecture

Transformers architecture has two key components associated: Encoder and Decoder. The encoder-decoder system leverages the attention mechanism to attain state-of-the-art effectiveness on the majority of natural language-based activities [45].

- The Encoder: The encoder simultaneously processes the English words and generates embeddings [BERT Base = 768 dimension]. These embeddings are the vectors that encapsulate the meaning of the word. Also, these vectors are not uniform and can change based on the context.
- The Decoder: The decoder uses the previously generated words and uses the embeddings produced by the Encoder to generate the next word.
- Input Embeddings: The input embeddings of a transformer network is generated by adding a positional encoding of each word i with a pre-trained embedding vector, where positional encoder is a vector that defines the position of words in a sentence typically using a sine and cosine vectors to generate a function.
- The final encoding will look like: Initial embedding (Word2vec) + positional encoding = vectors with positional context information.
- Multi-Head Self-Attention: Attention mechanism provides a way to learn the inherent relation between words within a single sentence or different related sentences. Self-attention finds the relevance of the i^{th} word in the input sentence with other words in the same input sentence. For each word, self-attention captures the relation and context in which that word is related to other words. To obtain more robust attention vectors for each word, we determine numerous attention vectors for each word and compute the final attention vector per word using a weighted average. To achieve multi-head self-attention we need V (Value), K (Key), Q (Query) vectors to extract different components of an input word. While using multi-head attention there will be multiple attention vectors for each word and a weighted matrix is created to convert 8 vectors to a single vector.
- Feed-Forward Layer: Feed-forward network is used in the encoder block to convert the output of the attention network to be used as an input to the next encoder/decoder blocks.
- Masked Multi-Head Self-Attention: On the decoder side, masked multi-head attention is needed because while generating the attention vectors, we can only use the previous words and not all the words, and therefore masking the future words is important.
- Multi-Head Encoder-Decoder Attention works with self-attention vectors of the masked multi-head self-attention and word's attention vectors from the Encoder block. This attention layer is responsible for establishing relationship between tokens of encoder sentence and decoder sentence and generates an attention vector for each word representing the relationship with words in other sentences.

- ADD and NORM: Batch normalisation and layer normalisation are performed after each layer to ensure that the weights do not become too high or low given the values and also to increase the training speeds considerably. In this study, we are using embedding layer, transformer and output layer with learning rate 2e−5, Adam optimizer and epsilon 1e−8, the attention vectors from the encoder block of a transformer can be used with a sigmoid activation function on the output layer to generate the results.

The concept of transfer learning gained a very strong foot with the advent of BERT. Further we leveraged the pretrained models and did the fine-tuning of these models.

6 PERFORMANCE EVALUATION

Quantitative assessment of three approaches implemented in this study is reported in this section. The evaluation quantitatively measures the performance of ML, DL and TL models on audio transcript data from DementiaBank.

6.1 Grid Search Techniques for Parameter Optimization

In classification models and estimators, there are a few parameters (constants and conditions) like learning rate, constraints, batch size, optimizers parameters etc. which are not tuned during the training phase.

The parameters provided before training as arguments are known as hyper-parameters, and the optimal value selection process is known as hyperparameters tuning. To find out optimal hyperparameters for these models, we applied the Grid search technique [46]. It is a primitive approach that exhaustively searches among all the possible values of parameters and gives the best set. To perform the grid search, an initial set of desired parameters is given, and an optimal set of hyperparameters that best reduces the cost function of the problem is selected for further process. The process is sometimes repeated or cross-validated to gain a generalized set of hyper-parameters. Optimized parameters for all three approaches are listed in Table 3.

6.2 Performance Measures for Classification

Examine the number of instances in which the class was accurately predicted but not assigned (true negatives) and the number of instances where the class was correctly predicted but not assigned (false positives) (false negatives). These four numbers form the confusion matrix in Table 4 for the binary classification.

The confusion matrices [47] are required to provide an analysis of the number of samples that were correctly classified as either a True Positive (TP), a False Positive (FP), or a True Negative (TN).

We investigated five performance indicators: Testing accuracy, validation accuracy, F1_score, area under the ROC curve and confusion matrices for the three

Conventional ML Models	**Hybrid Sequential DL Models**	**Transfer Learning Models**
SVM [params: c = 1.0, kernel=RBF, gamma=scale]	LSTM [Input length = 100, LSTM layer (128 hidden units), recurrent dropout = 0.2, Epochs = 50]	BERT [epochs: 8, tokenizer = BERT base, max tokenizer length = 64, batch size = 32, Optimizer = Adam (lr = 2e−5, epsilon = 1e−8)]
Random Forest [params: n_estimators = 100, min_sample_leaf = 1, min_sample_split = 2, criterion = gini]	Bidirectional LSTM [Input length = 100, BiLSTM layer (128 hidden units), recurrent dropout = 0.2, Epochs = 50]	XLNET [epochs: 12, tokenizer = XLNet base cased, max tokenizer length = 128, batch size = 32, optimizer = Adam (lr = 2e−5)]
Decision Tree [params: criterion = 'gini', min_sample_leaf = 1, min_sample_split = 2]	CNN-LSTM [Input length = 100, Conv1D (128 hidden units), Max-Pooling layer (4 × 4 filter), LSTM layer (64 hidden units), dropout = 0.1, Epochs = 50]	

Table 3. Optimized parameters of each model implemented in this study

	Predicted Condition		
Actual Condition	Total Population	Positive	Negative
	Positive	True Positive (TP)	False Negative (FN)
	Negative	False Positive (FP)	True Negative (TN)

Table 4. Confusion matrix for binary classification

approaches applied for classification tasks in this study [48]. Results obtained using these five performance measures for binary class (AD+ and AD−) classification task are given in the Results subsection.

6.3 Results

For NLP challenges, text classification has risen to the ML, DL, and TL applications in recent years. Classification outcomes evaluation is essential in healthcare for determining the existence of pathology and ruling out disease in healthy peo-

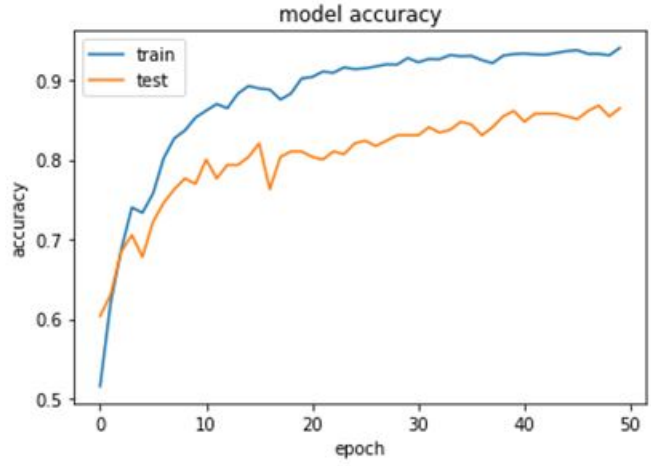
Model	Testing Accuracy (%)	Validation Accuracy (%)	F1_Score
First Approach: Conventional Machine Learning techniques			
Random Forest	84.4	85	0.84
Decision Tree	84.5	85	0.85
Support Vector Machine	85	86	0.85
Second Approach: Sequential Deep Learning Models			
LSTM	85.4	87	0.84
Bidirectional LSTM	85.3	87	0.85
CNN-LSTM	90	89	0.90
Third Approach: Transfer Learning Models			
XLNet	92	90	0.92
BERT	**93**	**95**	**0.93**

Table 5. Performance of three approaches implemented on DementiaBank dataset for the classification of AD+ and AD−

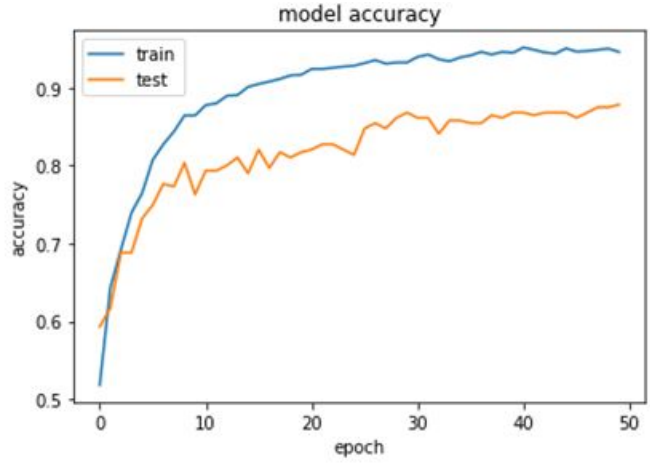
ple. Table 5 lists the classification results for three approaches: conventional ML, sequential DL, and pre-trained TL architecture on the DementiaBank dataset for comparison.

Using first approach, Decision Tree, Random Forest, and Support Vector Machine achieved an accuracy of 84.4 %, 84.5 % and 85 %, respectively. SVM among machine learning models showed better classification performance. Hybrid CNN-LSTM, among our second approach, has shown significant classification performance by achieving an accuracy of 90 %. However, the third approach has outperformed machine learning and deep learning models. XLNet and BERT model with optimized parameters have achieved the highest testing accuracy of 92 % and 93 %, respectively. When fed the vectors generated by TF-IDF to classification models using optimized parameters, the best-performing BERT model attained testing accuracy of 93 % and validation accuracy of 95 %, setting a new milestone for this challenge. For each layer, BERT uses the self-attention mechanism, and the result is passed through a feed-forward network and then to the next encoder. Moreover, BERT develops word representations that are dynamically influenced by the words surrounding them in order to capture various types of information, resulting in more accurate feature representations in dementia prediction. Hence, BERT's word-preprocessing and word-embedding characteristics improve the models performance, hence achieving a better performance.

The model's loss and accuracy information is stored in the object history for each epoch. Training and validation accuracy is plotted across the number of epochs in the training process. This will be effective in interpreting important decisions in the model's performance. Training vs. testing accuracy plots for conventional ML approaches (DT RF and SVM) are shown in Figure 9. The loss function is computed over all data items throughout an epoch and is assured to yield the quantitative loss measure at the given epoch. However, visualizing the curve over iterations provides

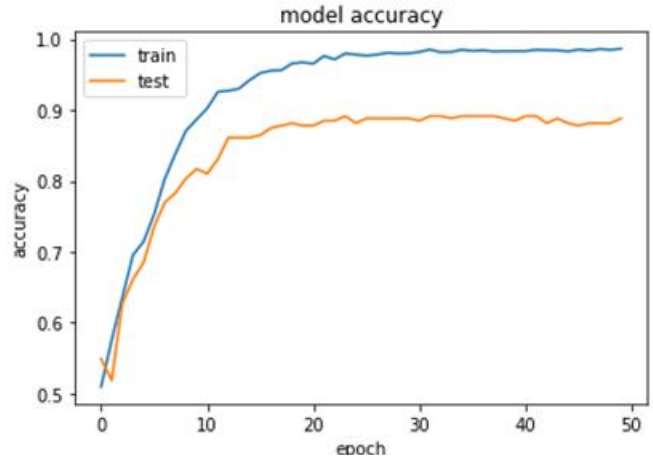


a)



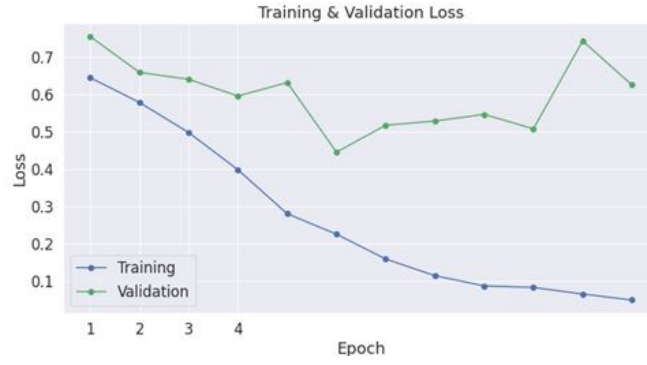
b)

information about the loss of a portion of the dataset. The training loss vs. validation loss over epochs for XLNet and BERT models is depicted in the graphs shown in Figure 10. BERT has shown higher validation accuracy and low validation loss compared to XLNet. Each epoch terminates after all training data is sent once. Whereas, data sent in batches, each epoch may contain numerous backpropagations steps which improves performance. Each backpropagation step in BERT enhanced the performance significantly while reduced validation loss. Further, the confusion matrix for the above discussed models is shown in Figure 11.

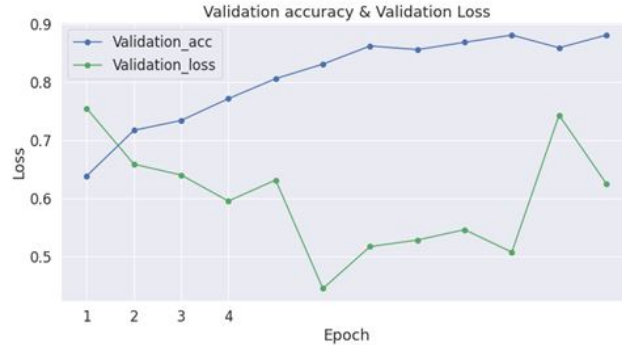


c)

Figure 9. Plots showing training vs. testing accuracy across the number of epochs using ML models for AD+ and AD− classification task for a) Decision Tree, b) Random Forest, and c) Support Vector Machine

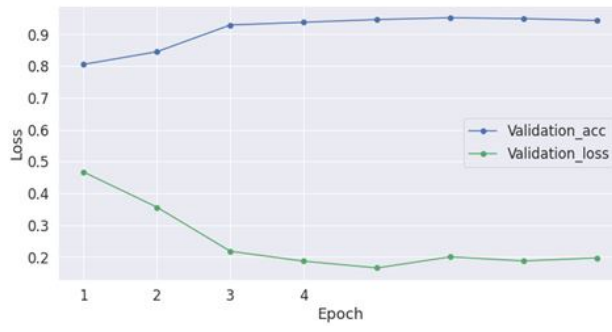


a)



b)

c)



d)

Figure 10. Plots showing a) training vs. validation loss for XLNet across the number of epochs, b) validation accuracy vs. validation loss for XLNet across the number of epochs, c) training vs. validation loss for BERT across the number of epochs, and (d) validation accuracy vs. validation loss for BERT across the number of epochs for AD+ and AD− classification task

A bar graph plotted to present the validation and testing accuracy for three approaches applied in this study (ML vs. DL vs. TL) is shown in Figure 12. The plot depicts that BERT has achieved the highest testing accuracy of 93 % and validation accuracy of 95 % for AD+ and AD− classification task.

6.4 Performance Evaluation

Area Under the Curve (AUC) is a performance measure to select the best from a set of different classification algorithms given a classification problem statement. It is basically an area covered under the Receiver's Operating Curve (ROC) [49]. Therefore AUC is the area under the ROC.

The classification performance of models is analyzed properly, distinguishing between the classes, and this metric is suitable only for binary classification problem

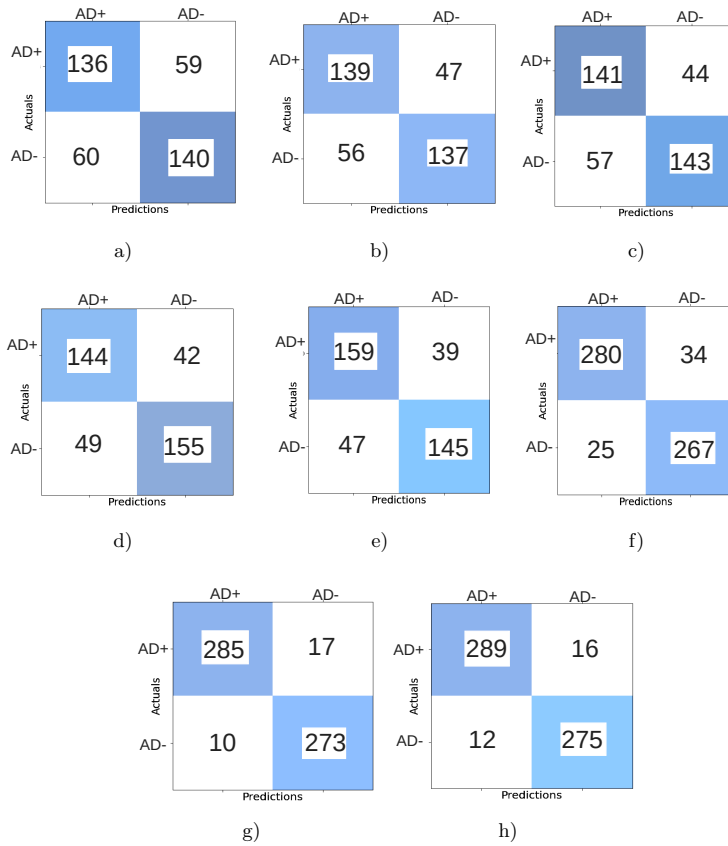


Figure 11. Confusion matrix showing accurately predicted and false predicted AD+ and AD− for a) DT, b) RF, c) SVM, d) LSTM, e) Bi-LSTM, f) CNN-LSTM, g) XLNet, h) BERT

statements. ROC is plotted for model performance with different threshold settings considering true positive rate (TPR) and false positive rate (FPR) on the X and Y axes, respectively. The plot of TPF (sensitivity) vs. FPF (1-specificity) for various cut-off values produces a ROC curve in the unit square [50]. The AUC score of algorithms is compared for a binary classification task, and the model with a higher AUC score is preferred. The higher the AUC, the better the ability of the machine learning algorithm to distinguish/classify the points in their correct classes. As an illustration, the corresponding ROC curves for the ML, DL, and TL techniques were drawn in Figure 13, 14, and 15, respectively. In most cases, derived indices like the area under the whole curve are used to measure the diagnostic accuracy. As diagnostic classification capability improves, the ROC curves associated with this capability move closer to the upper left-hand corner of the ROC space. The ratio

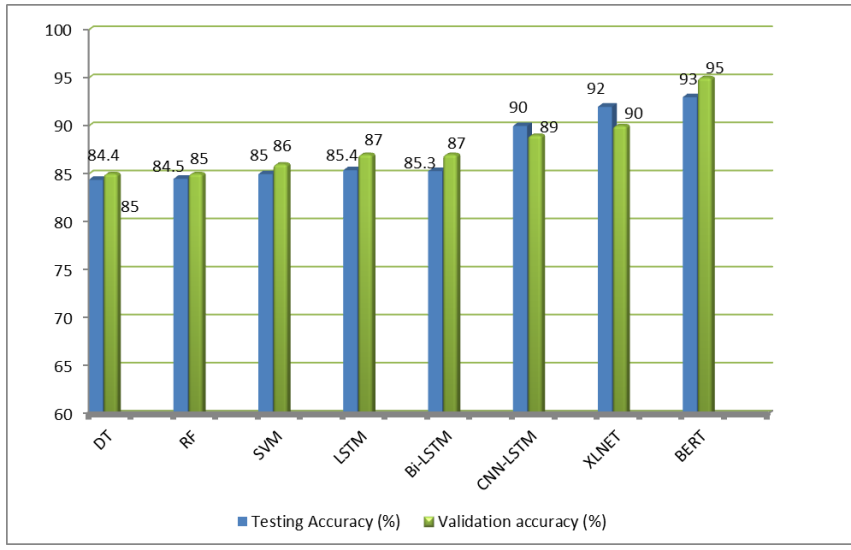
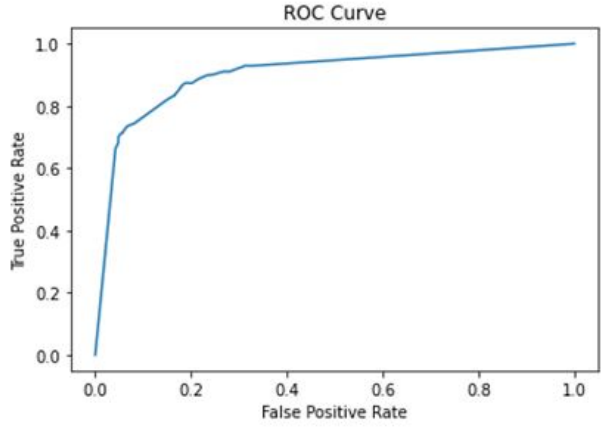
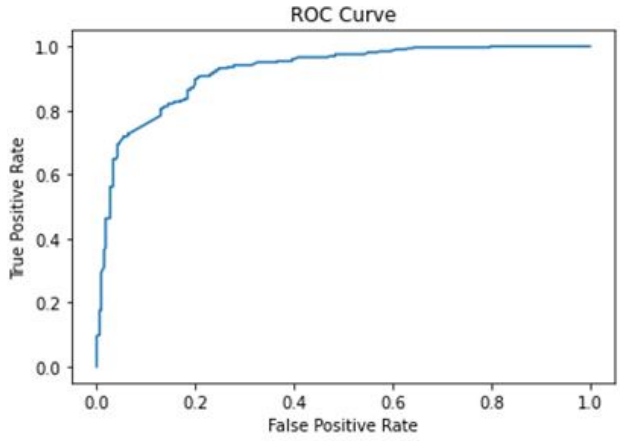


Figure 12. Graph plotting testing and validation accuracy for ML models (DT, RF, and SVM), DL models (LSTM, Bi-LSTM, CNN-LSTM) and TL models (XLNet and BERT)

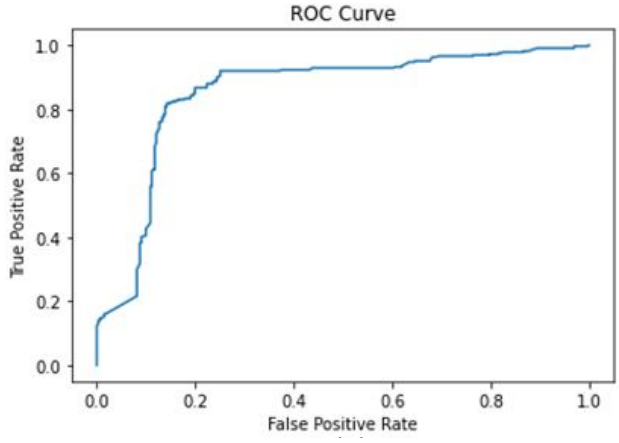
of the two density functions shows how many people have AD and how many do not have AD. The slope at each point on the ROC curve is equal to this ratio. The AUC values for XLNet and BERT are 0.97 and 0.97, respectively. BERT's ROC curve and associated AUC demonstrate a better predictor of Alzheimer's dementia than healthy individuals.



a)



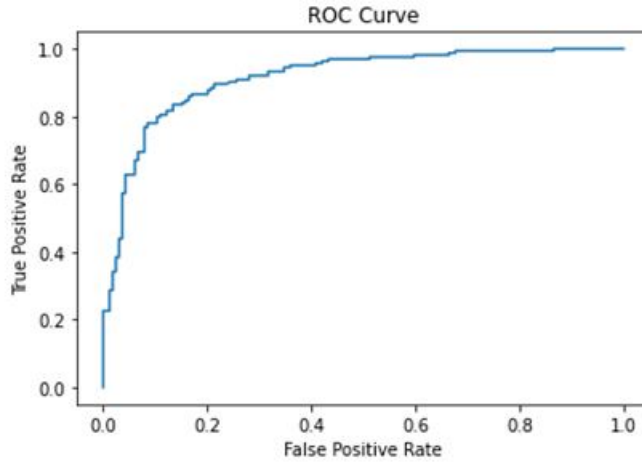
b)



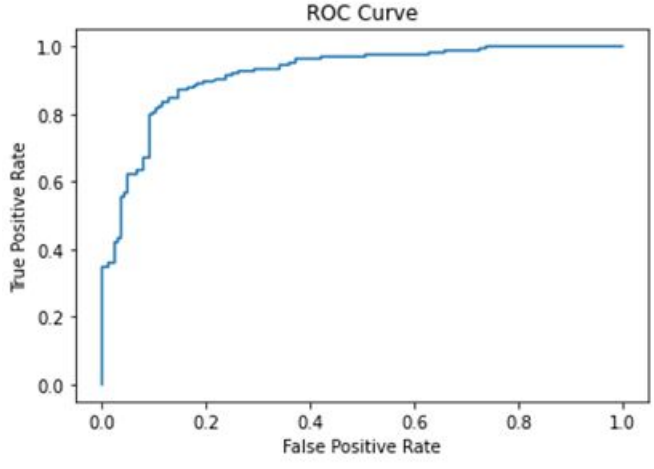
c)

Figure 13. AUC curve plots for AD+ and AD− classification task using ML approach: a) Decision Tree, b) Random Forest and c) Support Vector Machine model

Cross-validation (CV), a statistical technique, is implemented in this work to generalize the applicability of models in terms of performance. The CV approach effectively examines model performance scores. In this technique, the model is trained on several train-test splits a certain number of times, giving a better indication of how effectively the model would perform on unknown data. The procedure has a parameter named k that defines the number of groups into which a given data sample should be split. As a result, the technique is known as k-fold cross-validation. In this study, 10-fold cross-validation is performed ($k = 10$), where the data set is divided into ten parts at random. Nine of them are used for training, and one-tenth



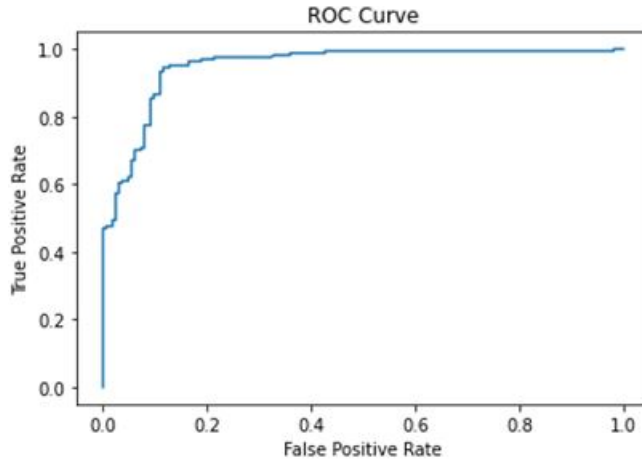
a)



b)

are used for testing and repeated the process ten times with a fresh holdout set and a new tenth to be tested. For $k-1$ training iterations, every data point is subjected to an exact test. The variance of the estimate diminishes as k increases. Moreover, CV has a higher potential of enabling the detection of over-fitting. Table 6 enlists the performance of eight models tested in this work for $k = 10$ folds, and the mean for each model is calculated over 10 folds.

The statistical mean following cross-validation illustrates the performance of the model over ten subgroups of the entire sampling. On average, 89.32 % of machine learning models, such as the DT, exhibit a strong comprehension and feature



c)

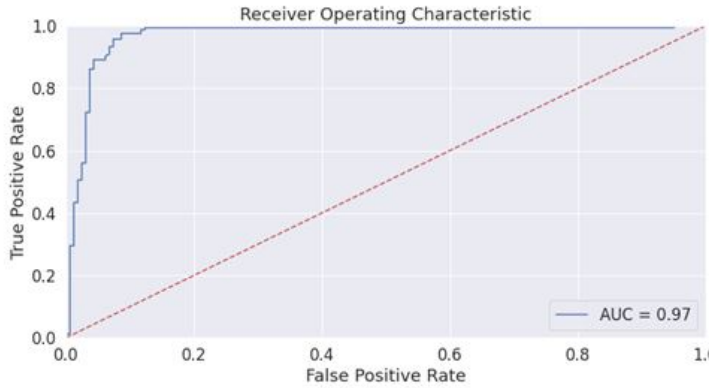
Figure 14. AUC curve plots for AD+ and AD− classification task using DL approach: a) LSTM, b) Bi-LSTM and c) CNN-LSTM model

Fold	DT	RF	SVM	LSTM	Bi-LSTM	CNN_LSTM	XLNet	BERT
1_fold	0.90909	0.90683	0.76442	0.83206	0.86641	0.91603	0.90992	0.95572
2_fold	0.89808	0.89490	0.71171	0.80916	0.85114	0.90458	0.90381	0.95572
3_fold	0.88535	0.86544	0.73333	0.81297	0.85114	0.91221	0.89923	0.95572
4_fold	0.88509	0.87009	0.71321	0.80534	0.83587	0.90458	0.90381	0.95572
5_fold	0.87087	0.87462	0.74407	0.82824	0.84351	0.90839	0.91297	0.95572
6_fold	0.89024	0.87425	0.73658	0.80534	0.83587	0.90458	0.90229	0.95572
7_fold	0.87341	0.86292	0.73157	0.83969	0.87404	0.91603	0.92366	0.95572
8_fold	0.92258	0.90909	0.76388	0.81297	0.84732	0.90458	0.91450	0.95572
9_fold	0.88165	0.87572	0.72972	0.83969	0.87022	0.92748	0.91145	0.95572
10_fold	0.91147	0.90189	0.72682	0.82442	0.86641	0.91603	0.91603	0.95572
Mean	0.89322	0.8917	0.66596	0.81626	0.850191	0.90801	0.90977	0.95572

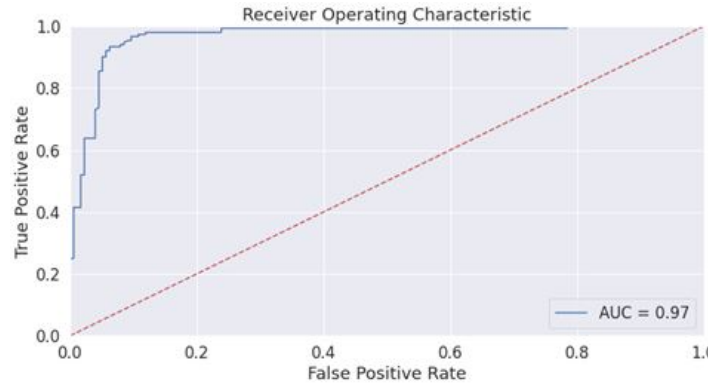
Table 6. 10-fold cross-validation performance evaluation of models with optimized parameters

simulation. A larger distributed mean for BERT (95.52) than for XLNet (90.97) and CNN-LSTM hybrid architecture (90.80) indicates the model's better stability.

Table 7 compares the benchmark performance achieved by the approaches used in the current study with the accuracy achieved by previous studies using these three approaches. Figure 16 shows a graphical representation of performance comparison in terms of accuracy. From the graph, it is concluded that TL, hybrid DL, DL, and most of the ML approaches used in the current study have shown a better



a)



b)

Figure 15. AUC curve plots for AD+ and AD− classification task using TL approach: a) XLNet and b) BERT Model

performance compared to state-of-the-art studies.

7 CONCLUSION AND FUTURE SCOPE

There is evidence that Alzheimer's dementia warning signs can be found in audio transcript data. We have attempted to provide an overview of existing state-of-the-art techniques in applying NLP to health outcomes research, with a specific emphasis on assessment methods. From the experimental analysis, it is concluded that AD can be diagnosed more quickly and accurately if linguistic biomarkers are analyzed through the verbal utterances of elderly persons. We implemented traditional ML, hybrid DL and TL approach to compare classification performance on

	Study	**Approach**	**Accuracy (%)**
Results of Previous Studies	Fraser et al. [21]	Machine Learning	81
	Orimaye et al. [20]		87
	König et al. [11]		87
	Meghanani et al. [25]	Deep Learning	83.33
	JabaSheela et al. [26]	Hybrid Deep Learning	72
	Sarawgi et al. [27]	Transfer Learning	88.0
	Balagopalan et al. [28]		85.14
Results of Current Study	Decision Tree (DT)	Machine Learning	84.4
	Random Forest (RF)		84.4
	Support Vector Machine (SVM)		85
	LSTM	Deep Learning	85.4
	Bi-LSTM		85.3
	CNN-LSTM	Hybrid Deep Learning	90
	XLNet	Transfer Learning	92
	BERT		93

Table 7. Comparative performance evaluation of approaches used in the proposed work with state-of-the-art studies and on benchmark dataset

linguistic data taken from DementiaBank. Statistical ML classifiers (such as SVM or RF) treat the sentence as a bag of word models (each word is treated independently, irrespective of its position in the sequence). Text data is temporal in nature, where the sequence of words defines the exact meaning of a sentence. Bi-LSTM or transformer-based architectures are aces in capturing the bidirectional semantic context of words in each sentence. As indicated, state-of-the-art techniques for assessing the early onset of Alzheimer's disease are quite varied, either by analyzing behavioural or language patterns in audio transcript data. An experimental comparative analysis of conventional ML, sequential DL, and TL approaches was carried out based on various performance evaluation metrics. Experimental results showed that transformer-based architectures were well performing in predicting AD with less validation loss and higher accuracy. The accuracy of fine-tuned BERT model is significantly outperformed in comparison with other models. Furthermore, the performance of predictive validation is estimated using k-fold cross-validation ($k = 10$), which better demonstrates how well a model will generalize to the test set. On the basis of these findings, future research will be conducted in order to develop better detection and prediction models for AD. We shall explore different advanced architectures, such as Embeddings from Language Model (ELMo), a deep contextualized word representation. Various experimentations can be performed using different tokenization, embedding, and encoding strategies for transcript representations. In addition to this, the use of feature-fusion techniques might improve the identification of dementia.

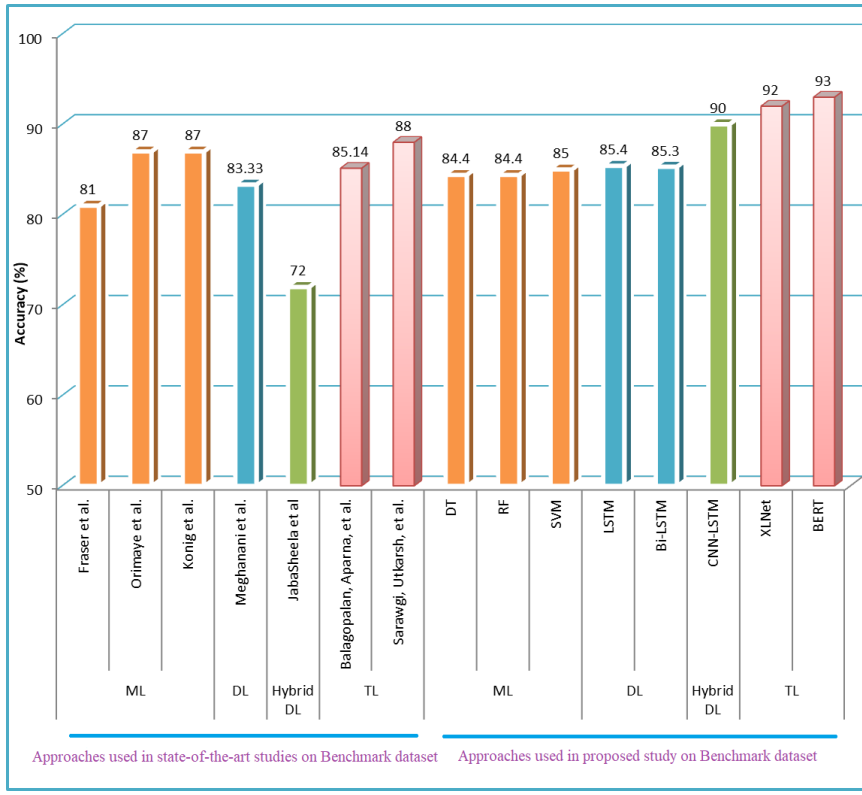


Figure 16. Performance comparison of machine learning, deep learning, and transfer learning approaches used in the current study with state-of-the-art studies

Data Availability

The Python code of this article will be shared on request to the corresponding author. The data underlying this article is available at URL: `https://dementia.talkbank.org/access/English/Pitt.html`.

Acknowledgement

The authors are thankful to Dr. Ankush Sharma, MD, DM (Neurology), Senior Consultant, Shri Mata Vaishno Devi Narayana Multispecialty Clinic, Jammu and Kashmir (India), for providing the consistent guidance related to Alzheimer's disease and also helping to affirm the important risk factors and symptoms that lead to cognitive and linguistic disabilities in Alzheimer's patient.

REFERENCES

[1] World Health Organization: Dementia. 2021, `https://www.who.int/news-room/fact-sheets/detail/dementia` [accessed April 2021].

[2] National Health Service: About Dementia. 2021, `https://www.nhs.uk/conditions/dementia/about/` [accessed April 2021].

[3] Stutzmann, G. E.: The Pathogenesis of Alzheimers Disease – Is It a Lifelong "Calciumopathy"? The Neuroscientist, Vol. 13, 2007, No. 5, pp. 546–559, doi: 10.1177/1073858407299730.

[4] Poulakis, K.—Pereira, J. B.—Mecocci, P.—Vellas, B.—Tsolaki, M.—Kłoszewska, I.—Soininen, H.—Lovestone, S.—Simmons, A.—Wahlund, L. O.—Westman, E.: Heterogeneous Patterns of Brain Atrophy in Alzheimer's Disease. Neurobiology of Aging, Vol. 65, 2018, pp. 98–108, doi: 10.1016/j.neurobiolaging.2018.01.009.

[5] Tunnard, C.—Whitehead, D.—Hurt, C.—Wahlund, L. O.—Mecocci, P.—Tsolaki, M.—Vellas, B.—Spenger, C.—Kłoszewska, I.—Soininen, H.—Lovestone, S.—Simmons, A.: Apathy and Cortical Atrophy in Alzheimer's Disease. International Journal of Geriatric Psychiatry, Vol. 26, 2011, No. 7, pp. 741–748, doi: 10.1002/gps.2603.

[6] Meilan, J. J. G.—Martinez-Sanchez, F.—Carro, J.—Carcavilla, N.—Ivanova, O.: Voice Markers of Lexical Access in Mild Cognitive Impairment and Alzheimer's Disease. Current Alzheimer Research, Vol. 15, 2018, No. 2, pp. 111–119, doi: 10.2174/1567205014666170829112439.

[7] Fox, N. C.—Warrington, E. K.—Seiffer, A. L.—Agnew, S. K.—Rossor, M. N.: Presymptomatic Cognitive Deficits in Individuals at Risk of Familial Alzheimer's Disease. A Longitudinal Prospective Study. Brain, Vol. 121, 1998, No. 9, pp. 1631–1639, doi: 10.1093/brain/121.9.1631.

[8] de Lira, J. O.—Minett, T. S. C.—Bertolucci, P. H. F.—Ortiz, K. Z.: Analysis of Word Number and Content in Discourse of Patients with Mild to Moderate Alzheimer's Disease. Dementia and Neuropsychologia, Vol. 8, 2014, No. 3, pp. 260–265, doi: 10.1590/S1980-57642014DN83000010.

[9] Smith, J. A.—Knight, R. G.: Memory Processing in Alzheimer's Disease. Neuropsychologia, Vol. 40, 2002, No. 6, pp. 666–682, doi: 10.1016/S0028-3932(01)00137-3.

[10] Hailstone, J. C.—Ridgway, G. R.—Bartlett, J. W.—Goll, J. C.—Buckley, A. H.—Crutch, S. J.—Warren, J. D.: Voice Processing in Dementia: A Neuropsychological and Neuroanatomical Analysis. Brain, Vol. 134, 2011, No. 9, pp. 2535–2547, doi: 10.1093/brain/awr205.

[11] König, A.—Satt, A.—Sorin, A.—Hoory, R.—Toledo-Ronen, O.—Derreumaux, A.—Manera, V.—Verhey, F.—Aalten, P.—Robert, P. H.—David, R.: Automatic Speech Analysis for the Assessment of Patients with Predementia and Alzheimer's Disease. Alzheimer's and Dementia: Diagnosis, Assessment and Disease Monitoring, Vol. 1, 2015, No. 1, pp. 112–124, doi: 10.1016/j.dadm.2014.11.012.

[12] Srivastava, S. K.—Singh, S. K.—Suri, J. S.: Chapter 16 – A Healthcare Text Classification System and Its Performance Evaluation: A Source of Better Intelligence by Characterizing Healthcare Text. In: Sinha, G., Suri, J. S. (Eds.): Cognitive Informatics, Computer Modelling, and Cognitive Science. Volume 2: Application to Neural Engineering, Robotics, and STEM. Elsevier, 2020, pp. 319–369, doi: 10.1016/B978-0-12-819445-4.00016-3.

[13] Nigam, K.—McCallum, A. K.—Thrun, S.—Mitchell, T.: Text Classification from Labeled and Unlabeled Documents Using EM. Machine Learning, Vol. 39, 2000, No. 2, pp. 103–134, doi: 10.1023/A:1007692713085.

[14] Maqsood, M.—Nazir, F.—Khan, U.—Aadil, F.—Jamal, H.—Mehmood, I.—Song, O. Y.: Transfer Learning Assisted Classification and Detection of Alzheimer's Disease Stages Using 3D MRI Scans. Sensors, Vol. 19, 2019, No. 11, Art. No. 2645, doi: 10.3390/s19112645.

[15] Zhou, K.—He, W.—Xu, Y.—Xiong, G.—Cai, J.: Feature Selection and Transfer Learning for Alzheimer's Disease Clinical Diagnosis. Applied Sciences, Vol. 8, 2018, No. 8, Art. No. 1372, doi: 10.3390/app8081372.

[16] Li, B.—Hsu, Y. T.—Rudzicz, F.: Detecting Dementia in Mandarin Chinese Using Transfer Learning from a Parallel Corpus. 2019, doi: 10.48550/arXiv.1903.00933.

[17] Cook, B. L.—Progovac, A. M.—Chen, P.—Mullin, B.—Hou, S.—Baca-Garcia, E.: Novel Use of Natural Language Processing (NLP) to Predict Suicidal Ideation and Psychiatric Symptoms in a Text-Based Mental Health Intervention in Madrid. Computational and Mathematical Methods in Medicine, Vol. 2016, 2016, Art. No. 8708434, doi: 10.1155/2016/8708434.

[18] Wang, Y.—Wang, Z.—Li, C.—Zhang, Y.—Wang, H.: A Multitask Deep Learning Approach for User Depression Detection on Sina Weibo. 2020, doi: 10.48550/arXiv.2008.11708.

[19] Do, B. H.—Wu, A.—Biswal, S.—Kamaya, A.—Rubin, D. L.: Informatics in Radiology: RADTF: A Semantic Search-Enabled, Natural Language Processor-Generated Radiology Teaching File. RadioGraphics, Vol. 30, 2010, No. 7, pp. 2039–2048, doi: 10.1148/rg.307105083.

[20] Orimaye, S. O.—Wong, J. S. M.—Golden, K. J.—Wong, C. P.—Soyiri, I. N.: Predicting Probable Alzheimer's Disease Using Linguistic Deficits and Biomarkers. BMC Bioinformatics, Vol. 18, 2017, No. 1, Art. No. 34, doi: 10.1186/s12859-016-1456-0.

[21] Fraser, K. C.—Meltzer, J. A.—Rudzicz, F.: Linguistic Features Identify Alzheimer's Disease in Narrative Speech. Journal of Alzheimer's Disease, Vol. 49, 2016, No. 2, pp. 407–422, doi: 10.3233/JAD-150520.

[22] Clark, D. G.—McLaughlin, P. M.—Woo, E.—Hwang, K.—Hurtz, S.—Ramirez, L.—Eastman, J.—Dukes, R. M.—Kapur, P.—DeRamus, T. P.—Apostolova, D. G.: Novel Verbal Fluency Scores and Structural Brain Imaging for Prediction of Cognitive Outcome in Mild Cognitive Impairment. Alzheimer's and Dementia: Diagnosis, Assessment and Disease Monitoring, Vol. 2, 2016, No. 1, pp. 113–122, doi: 10.1016/j.dadm.2016.02.001.

[23] Karlekar, S.—Niu, T.—Bansal, M.: Detecting Linguistic Characteris-

tics of Alzheimer's Dementia by Interpreting Neural Models. 2018, doi: 10.48550/arXiv.1804.06440.

[24] Galaverna, F.—Bueno, A. M.—Morra, C. A.—Roca, M.—Torralva, T.: Analysis of Errors in Verbal Fluency Tasks in Patients with Chronic Schizophrenia. The European Journal of Psychiatry, Vol. 30, 2016, No. 4, pp. 305–320.

[25] Meghanani, A.—Anoop, C. S.—Ramakrishnan, A. G.: Recognition of Alzheimer's Dementia from the Transcriptions of Spontaneous Speech Using fastText and CNN Models. Frontiers in Computer Science, Vol. 3, 2021, Art. No. 624558, doi: 10.3389/fcomp.2021.624558.

[26] JabaSheela, L.—Vasudevan, S.—Yazhini, V. R.: A Hybrid Model for Detecting Linguistic Cues in Alzheimer's Disease Patients. Journal of Information and Computational Science, Vol. 10, 2020, No. 1, pp. 85–90.

[27] Sarawgi, U.—Zulfikar, W.—Soliman, N.—Maes, P.: Multimodal Inductive Transfer Learning for Detection of Alzheimer's Dementia and Its Severity. 2020, doi: 10.48550/arXiv.2009.00700.

[28] Balagopalan, A.—Eyre, B.—Robin, J.—Rudzicz, F.—Novikova, J.: Comparing Pre-Trained and Feature-Based Models for Prediction of Alzheimer's Disease Based on Speech. Frontiers in Aging Neuroscience, Vol. 13, 2021, Art. No. 635945, doi: 10.3389/fnagi.2021.635945.

[29] Subba, B.—Gupta, P.: A Tfidfvectorizer and Singular Value Decomposition Based Host Intrusion Detection System Framework for Detecting Anomalous System Processes. Computers and Security, Vol. 100, 2021, Art. No. 102084, doi: 10.1016/j.cose.2020.102084.

[30] Vidhya, R.—Gopalakrishnan, P.—Vallamkondu, N. K.: Sentiment Analysis Using Machine Learning Classifiers: Evaluation of Performance. Proceedings of the First International Conference on Computing, Communication and Control System (I3CAC 2021), EAI, 2021, doi: 10.4108/eai.7-6-2021.2308565.

[31] Patel, H. H.—Prajapati, P.: Study and Analysis of Decision Tree Based Classification Algorithms. International Journal of Computer Sciences and Engineering, Vol. 6, 2018, No. 10, pp. 74–78, doi: 10.26438/ijcse/v6i10.7478.

[32] Charbuty, B.—Abdulazeez, A.: Classification Based on Decision Tree Algorithm for Machine Learning. Journal of Applied Science and Technology Trends, Vol. 2, 2021, No. 1, pp. 20–28.

[33] Biau, G.—Scornet, E.: A Random Forest Guided Tour. TEST, Vol. 25, 2016, No. 2, pp. 197–227, doi: 10.1007/s11749-016-0481-7.

[34] Goudjil, M.—Koudil, M.—Bedda, M.—Ghoggali, N.: A Novel Active Learning Method Using SVM for Text Classification. International Journal of Automation and Computing, Vol. 15, 2018, No. 3, pp. 290–298, doi: 10.1007/s11633-015-0912-z.

[35] Searle, T.—Ibrahim, Z.—Dobson, R.: Comparing Natural Language Processing Techniques for Alzheimer's Dementia Prediction in Spontaneous Speech. 2020, doi: 10.48550/arXiv.2006.07358.

[36] Altinel, B.—Ganiz, M. C.—Diri, B.: A Corpus-Based Semantic Kernel for Text Classification by Using Meaning Values of Terms. Engineering Applications of Artificial Intelligence, Vol. 43, 2015, pp. 54–66, doi: 10.1016/j.engappai.2015.03.015.

[37] Van Phan, T.—Nakagawa, M.: Text/Non-Text Classification in Online Handwritten Documents with Recurrent Neural Networks. 2014 14th International Conference on Frontiers in Handwriting Recognition, 2014, pp. 23–28, doi: 10.1109/ICFHR.2014.12.

[38] Edo-Osagie, O.—Lake, I.—Edeghere, O.—De La Iglesia, B.: Attention-Based Recurrent Neural Networks (RNNs) for Short Text Classification: An Application in Public Health Monitoring. In: Rojas, I., Joya, G., Catala, A. (Eds.): Advances in Computational Intelligence (IWANN 2019). Springer, Cham, Lecture Notes in Computer Science, Vol. 11506, 2019, pp. 895–911, doi: 10.1007/978-3-030-20521-8_73.

[39] Yao, L.—Guan, Y.: An Improved LSTM Structure for Natural Language Processing. 2018 IEEE International Conference of Safety Produce Informatization (IICSPI), 2018, pp. 565–569, doi: 10.1109/IICSPI.2018.8690387.

[40] Cummins, N.—Pan, Y.—Ren, Z.—Fritsch, J.—Nallanthighal, V. S.—Christensen, H.—Blackburn, D.—Schuller, B. W.—Magimai-Doss, M.—Strik, H.—Härmä, A.: A Comparison of Acoustic and Linguistics Methodologies for Alzheimer's Dementia Recognition. Proceedings of Interspeech 2020, ISCA – International Speech Communication Association, 2020, pp. 2182–2186.

[41] Rohanian, M.—Hough, J.—Purver, M.: Multi-Modal Fusion with Gating Using Audio, Lexical and Disfluency Features for Alzheimer's Dementia Recognition from Spontaneous Speech. 2021, doi: 10.48550/arXiv.2106.09668.

[42] Rekha, R.—Indhuja, M.—Nivethitha, S.—Shophiya, K.—Sundarrajan, V.: Alzheimer's Disease Detection Using Speech Dataset. In: Mandal, J. K., De, D. (Eds.): Advanced Techniques for IoT Applications (EAIT 2021). Springer, Singapore, Lecture Notes in Networks and Systems, Vol. 292, 2022, pp. 183–194, doi: 10.1007/978-981-16-4435-1_19.

[43] Ruder, S.—Peters, M. E.—Swayamdipta, S.—Wolf, T.: Transfer Learning in Natural Language Processing. In: Sarkar, A., Strube, M. (Eds.): Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Tutorials. ACL, 2019, pp. 15–18, doi: 10.18653/v1/N19-5004.

[44] Yang, Z.—Dai, Z.—Yang, Y.—Carbonell, J.—Salakhutdinov, R. R.—Le, Q. V.: XLNet: Generalized Autoregressive Pretraining for Language Understanding. In: Wallach, H., Larochelle, H., Beygelzimer, A., d'Alché-Buc, F., Fox, E., Garnett, R. (Eds.): Advances in Neural Information Processing Systems 32 (NeurIPS 2019). Curran Associates, Inc., 2019, pp. 5753–5763.

[45] Özçift, A.—Akarsu, K.—Yumuk, F.—Söylemez, C.: Advancing Natural Language Processing (NLP) Applications of Morphologically Rich Languages with Bidirectional Encoder Representations from Transformers (BERT): An Empirical Case Study for Turkish. Automatika, Vol. 62, 2021, No. 2, pp. 226–238, doi: 10.1080/00051144.2021.1922150.

[46] Bhat, P. C.—Prosper, H. B.—Sekmen, S.—Stewart, C.: Optimizing Event Selection with the Random Grid Search. Computer Physics Communications, Vol. 228, 2018, pp. 245–257, doi: 10.1016/j.cpc.2018.02.018.

[47] Luque, A.—Carrasco, A.—Martín, A.—de las Heras, A.: The Impact of Class Imbalance in Classification Performance Metrics Based on the Bi-

nary Confusion Matrix. Pattern Recognition, Vol. 91, 2019, pp. 216–231, doi: 10.1016/j.patcog.2019.02.023.

[48] Sokolova, M.—Lapalme, G.: A Systematic Analysis of Performance Measures for Classification Tasks. Information Processing and Management, Vol. 45, 2009, No. 4, pp. 427–437, doi: 10.1016/j.ipm.2009.03.002.

[49] Kamarudin, A. N.—Cox, T.—Kolamunnage-Dona, R.: Time-Dependent ROC Curve Analysis in Medical Research: Current Methods and Applications. BMC Medical Research Methodology, Vol. 17, 2017, No. 1, Art. No. 53, doi: 10.1186/s12874-017-0332-6.

[50] Hajian-Tilaki, K.: Receiver Operating Characteristic (ROC) Curve Analysis for Medical Diagnostic Test Evaluation. Caspian Journal of Internal Medicine, Vol. 4, 2013, No. 2, pp. 627–635.

Yusera Farooq Khan received B.Tech. degree in computer science and engineering from the BGSB University, Jammu and Kashmir in 2012 and her Master degree in computer science and engineering from NIMS University, Rajasthan in 2014. She has cleared GATE. From 2014 to 2018, she was Assistant Professor at BGSB University, Jammu and Kashmir, India. She has 4 years of teaching and research experience. Her areas of interest are machine learning, deep learning, transfer learning, NeuroImage image analysis and computer vision.

Baijnath Kaushik received B.Eng. in computer science and engineering from the Nagpur University in 1997, Master of Technology from the University School of Information Technology, GGSIPU, New Delhi in 2009 and his Ph.D. in computer science and engineering from IIT Dhanbad in 2016. He has more than 21 years of teaching and research experience. Presently, he is Associate Professor in the School of Computer Science and Engineering, SMVDU, Katra, Jammu and Kashmir, India. His research interest includes machine learning, deep learning, nature inspired algorithms, soft computing, and parallel algorithms.

Bilal Ahmad Mir received his Bachelor degree in computer science and engineering from the Dr. K. N. Modi University, Rajasthan, India in 2017; his Master degree in intellectual information engineering from the University of Toyama, Japan in 2019. His research interest is focused on artificial intelligence, machine learning, image processing, computer vision, robotics technology.