Exploiting Linguistic Information from Nepali Transcripts for Early Detection of Alzheimer's Disease using the State-of-the-art Techniques of Machine Learning and Natural Language Processing

Surabhi Adhikari[1], Surendrabikram Thapa[1,2], Usman Naseem[3], Priyanka Singh[4], Angela Huo[4], Gnana Bharathy[5], Mukesh Prasad[4]

[1] Department of Computer Science and Engineering, Delhi Technological University, Delhi, India
[2] Department of Computer Science, Virginia Tech, Blacksburg, Virginia
[3] School of Computer Science, The University of Sydney, Sydney, Australia
[4] School of Computer Science, FEIT, University of Technology Sydney, Sydney, Australia
[5] School of Information, Systems and Modelling, University of Technology Sydney, Sydney, Australia

Abstract: Alzheimer's disease (AD) is considered as progressing brain disease, which can be slowed down with the early detection and proper treatment by identifying the early symptoms. Language change serves as an early sign that a patient's cognitive functions have been impacted, potentially leading to early detection. The effects of language changes are being studied thoroughly in the English language to analyze the linguistic patterns in AD patients using Natural Language Processing (NLP). However, it has not been much explored in local languages and low-resourced languages like Nepali. In this paper, we have created a novel dataset on low resources language, i.e., Nepali, consisting of transcripts of the AD patients and control normal subjects. We have also presented baselines by applying various machine learning (ML) and deep learning (DL) algorithms on a novel dataset for the early detection of AD. The proposed work incorporates the speech decline of AD patients in order to classify them as control subjects or AD patients. This study makes an effective conclusion that the difficulty in processing information of AD patients reflects in their speech narratives of patients while describing a picture. The dataset is made publicly available.

Keywords: Alzheimer's Disease, Deep Learning, Natural Language Processing, Machine Learning, Nepali Language, Low Resourced Language

1. Introduction

Alzheimer's disease (AD) is a progressive neurodegenerative condition affecting more than 50 million population across the globe. With someone developing the disease every three seconds, AD renders to be the most common form of dementia [1]. According to the 2018 World Alzheimer Report [2], the number of patients suffering from AD will cross the mark of 150 million by 2050, and the cost of treatment of Alzheimer's Disease is expected to cross 2 trillion US dollars by 2030. Currently, there aren't any approved drugs that can cure or completely stop how AD progresses [3]. However, there are some drugs and medications that can aid patients who are diagnosed in the earlier stages of AD. The early diagnosis of AD thus also helps in better management of the disease for both patients and caretakers. Hence, it is extremely necessary to find out the methods for the early diagnosis of AD for our aging society. AD patients show a wide range of symptoms due to the changes in the cortical anatomy [4]. One of the essential early indications of AD is cognitive impairment. Such cognitive impairments are mostly due to biological factors like atrophies in the various regions of the brain [5]. For example, atrophies in the left anterior temporal lobe impair naming tasks, such as picture description problems [6]. Such atrophies in the brain regions can be detected only by imaging techniques such as Magnetic Resonance Imaging (MRI) or Computed Tomography (CT) scans of the brain. Analyzing such imaging modalities would help us to classify the AD patients from the CN subjects, but analyzing them should be highly mediated by medical personnel. On the other hand, the patients with cognitive impairments show some visible symptoms like aphasia or limited ability in producing and understanding speech even for day-to-day tasks [7]. Such cognitive impairment is also often characterized by semantic memory deficits

and is mostly evidenced by naming impairment and the use of substitution words [8]. AD patients tend to reduce the amount of information, and such impaired subjects tend to use reduced working vocabulary. These impairments become noticeably evident with the progression of AD. Faber-Langendoen et al. [9], in the study of aphasia in AD patients, found out that 100% of the AD patients and 36% of the patients with mild cognitive impairment (MCI) had problems aphasia whose severity increased with increased severity of dementia. Such anomalies in linguistic features of speech produced by AD patients can be leveraged in building intelligent predictive systems for the diagnosis of AD in earlier stages.

Ahmed et al. [10] found that more than two-thirds of the participants showed significant changes in speech production way earlier before the medical diagnosis of AD. The speech patterns were significant as early as one year before the diagnosis of AD. Thus, speech can be a simple yet most prominent feature that can be used to build a powerful model for AD diagnosis. Kirshner et al. [11] found that all the participants had naming impairments despite absolutely normal speech in other respects. So, picture description tasks that heavily involve naming and identifying objects can be useful for learning the problems in speech. Also, thematic coherence, the ability of the speaker to maintain flow or theme in their speech, is heavily impaired in AD patients [12]. Their discourse lacked coherence as compared to CN (Control Normal) individuals. Currently, there are various neuropsychological tests available to assess the cognitive abilities of patients with AD. Some of the most widely used tests are Mini-mental State Examination (MMSE) [13], Rowland Universal Dementia Assessment Scale (RUDAS) [14], Alzheimer's Disease Assessment Scale-Cognitive (ADAS-Cog) [15], etc. The neuropsychological tests are mostly general, and since the memory impairment cannot be assessed with narrow criteria, the questions in such tests cannot assess cognitive abilities effectively. The same set of questions does not fit all the patients because the questions in which one patient may excel can be found difficult by other patients [16]. Also, most neuropsychological tests are used for an extended period and require psychologists or trained personnel to intervene throughout the assessment process. Similarly, ethics over the collection of personal information in neuropsychological assessment is also a problem to be looked upon. Moreover, AD diagnosis becomes difficult in times when the patients keep the symptoms to themselves only [2]. In such scenarios, the assessment of cognitive impairment can be done using self-generated speech on problems like picture description tasks [6].

When psychologists try to use naturally spoken language for the analysis of dementia or AD, it takes much time because of different linguistic patterns for different individuals. When computational linguistics is used for this purpose, the learning models trained on a large corpus of speech transcripts can show promising results as such models can be instrumental in learning the pattern in speech narratives of the subjects. Natural Language Processing (NLP) can hence be an alternative as well as a more appropriate technique for analyzing and interpreting the AD patient's speech. With computers becoming faster and faster, the speech narratives of subjects under study can be processed using NLP in real-time for the detection of AD. With the prospects of NLP being explored for mental illnesses like depression or schizophrenia, NLP can thus be significantly useful in improving the care delivery system for AD as well [17]. Apart from the difficulties in carrying out daily activities due to the severe symptoms of AD, it also poses an unprecedented burden and stigma upon those diagnosed with the disease. This study is also in the direction of lessening the stigma around AD by leveraging NLP tools. Due to the very limited amount of work done in the South Asian region and especially in low resources languages, the study employs a work-around for procuring data and underlying NLP experiments for the purpose. The dataset we have created is a translation from the pre-existing DementiaBank dataset in the English language. The

motivation behind this work is majorly the use of automation and mainly NLP for detection of AD in the low-resource language, as the advanced computational tools are still lagging in this region. This way of detecting AD is fast, cost-effective, and very accurate. If this approach can be demonstrated, it would provide an economic augmentation to both traditional assessments and primary data collection in AD detection on several under-resourced languages in various regions of the world.

The main contributions of the paper are:

- A novel manually annotated Alzheimer's disease dataset for low resource language, i.e., Nepalese, consisting of 168 Alzheimer's disease patients and 98 Control normal subjects, is presented. The dataset is made publicly available to the research community.
- An NLP-based framework is presented for the early detection of AD patients using Nepali transcripts and developed a visualization of content present in textual data. In addition to this, a word cloud of the most common words is presented to give qualitative analysis.
- The performance of different state-of-the-art machine learning-based textual classification mechanisms are presented with the baseline results.

Section 2 of the paper describes the works that have been done to detect AD from the linguistic features of the speech. The literature includes the work done using speech and transcripts for early detection of AD. Section 3 describes the methodology that has been used in this paper. The experimental results are discussed in section 4, and section 5 is the conclusion section that summarizes the findings of the paper, along with the future works that need to be done.

2. Related Works

In recent times, there have been various research going around in the task of the early diagnosis of AD using speech narratives of the subjects under study. In the last decade or so, much research is being conducted to figure out ways for the detection of AD using speech and linguistic features as the impairment of speech is one of the earliest symptoms of AD or Mild Cognitive Impairment (MCI). Thus, various machine learning (ML) methods are being used to detect anomalies in the speech narratives of subjects under study. Orimaye et al. [18] took syntactic, lexical, and n-gram based features for building the diagnostic model. The n-gram models had improved performance as compared to those models which used syntactic and lexical features alone. Using the top 1000 n-gram features, the model gave the Area Under Curve (AUC) value of 0.930, which was estimated using the Leave-Pair-Out Cross-Validation (LPOCV) technique. Also, Vincze et al. [19] used transcripts to classify patients with MCI and AD. The importance of morphological and speech-based features was highlighted in the research. Using only statistically significant features, Support Vector Machines (SVM) provided accuracy as high as 75%. These previously mentioned works used machine learning models. ML techniques require hand-crafted features. Such hand-crafted features vary extensively because of the different levels of expertise of the researchers in the diagnosis of AD. Also, hand-picked features are very easily outdated as the culture and language keep evolving continuously.

To overcome this drawback of using ML methods for the diagnosis of AD using transcripts of speech, some of the recent works have used intelligent deep learning models which can learn the intrinsic complexities of speech transcripts to automatically identify the linguistic features that reflect in narratives of AD patients with multiple levels of abstraction. Fritsch et al. [20] used a

neural network language model (NNLM) with Long Short-Term Memory (LSTM) cells to enhance the statistical approach of n-gram language models. The model was evaluated by measuring its perplexity. The scripts were evaluated by the model in a Leave-One-Out Cross-Validation (LOOCV) scheme. The perplexity values showed that the model could classify the AD and CN subjects with an accuracy of 85.6%. This suggests that the AD patients described the picture in an unexpected manner leading to unpredictable language structures that resulted in higher perplexity values. Chen et al. [21] proposed an attention-based hybrid network for automatic detection of AD. The hybrid model of attention-based Convolutional Neural Networks (CNN) and attention-based Bidirectional Gated Recurrent Unit (BiGRU) categorized the transcripts with an accuracy of 97.4%. The paper suggests that including attention mechanisms allowed the network to emphasize the decisive features of the subjects. Much work has been done in the English language, and some of the experiments have resulted in state-of-the-art (SOTA) models. The linguistic components of the English language, which affect the classification of CN vs. MCI vs. AD, are well explored. On the other hand, the research for the early detection of AD using linguistic features in languages other than English is not well explored. According to WHO, among the total number of dementia patients worldwide, 58% of the patients are from low and middle-income generating countries [22]. Building models only in the English language would leave a considerable fraction of the population without diagnostic tools that use NLP.

Much work has been done in major languages like Mandarin Chinese, German [23], Hungarian [24], etc. For instance, Liu et al. [25] used a dependency network approach to examine syntactic impairments of Chinese AD patients. Most of the AD patients showed regular syntactic impairments, which is evidence that there is language deterioration. Apart from Chinese, linguistic and acoustic features have been explored in various other languages like German [23], Hungarian [24], etc. There have also been researches on how spontaneous speech in various languages can be used in the analysis of AD through speech. Weiner et al. [23] used spontaneous conversational speeches in the German language to build models. Using Linear Discriminant Analysis (LDA) classifier with singular value decomposition, the researchers could get an F1-score of 0.800. They, however, used models for three classes classification viz. Control Normal (CN), Aging-associated Cognitive Decline (AACD), and Alzheimer's Disease (AD). Similarly, low resource language researchers have researched this domain using telephonic conversations also. Khodabakhsh et al. [26] used 10 minutes of telephonic conversations recorded using microphones for Turkish speakers. The texts were manually transcribed, and the learning algorithms were used. With conversational recording transcripts of 20 AD patients and 20 healthy individuals, the models were built. The features like hesitation and puzzlement features, Part of Speech (POS) based features, unintelligible word rate, complexity features like phonemes per word, etc., were used. They used algorithms like Support Vector Machine (SVM), LDA, and decision trees. With the LOOCV scheme, the researchers were able to get accuracy as high as 90%. It can be seen that there have been many initiatives to build models in multiple languages. However, there has not been any research in this domain in the South Asian regional languages. For a low-resource language like Nepali, where there is very limited research in NLP, this work in the detection of AD using speech narratives by exploiting linguistic cues is the first of its kind.

3. Methodology

The proposed framework for the experiment is as shown in Fig. 1. The process starts with data collection, which involves extracting the transcripts from the dementia bank and translating them into the Nepali language. The text is further pre-processed, and features are extracted. Similarly,

the models are trained and tested using a 10-fold stratified cross-validation scheme. After that, the various performance measures like precision, recall, accuracy, and F1-score are calculated. The components of the framework are explained below in great detail.

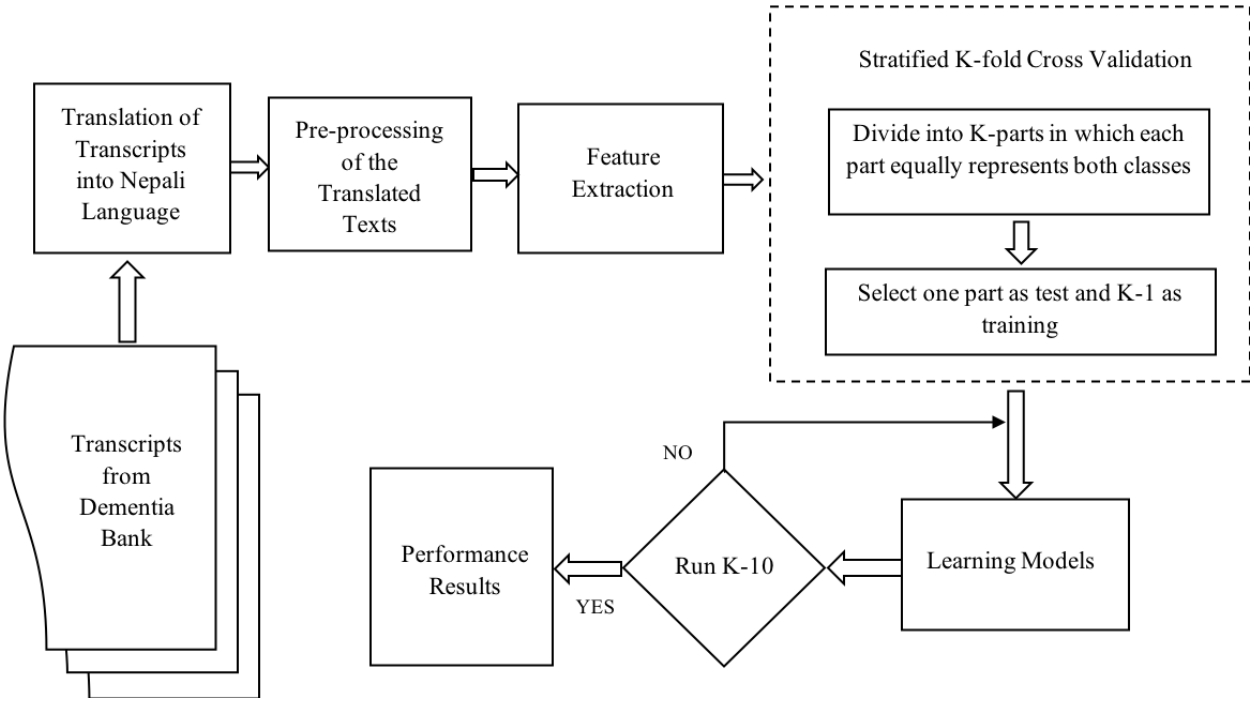

Fig. 1 Flow diagram of the Framework Used in Experiment

3.1 Data Collection

In this study, DementiaBank's Pitt Corpus has been utilized. The DementiaBank is one of the largest available datasets of audio recordings and transcripts of subjects who participated in the research conducted by Becker et al. [27] of the University of Pittsburgh School of Medicine. The recordings were manually transcribed by the CHAT protocol [28]. CHAT stands for (Codes for the Human Analysis of Transcripts), which was the format used for the CHILDES (Child Language Data Exchange System).

This study uses the transcripts of audio recordings for the Cookie Theft picture description task explicitly. The cookie theft description task was first used by the Boston Diagnostic Aphasia Examination protocol and hence mentioned as the Boston Cookie Theft picture description task in literature [29]. In this task, participants were asked to describe the kitchen scene, as shown in Fig. 2. There were 292 participants in this study conducted by the University of Pittsburgh School of Medicine. Among 292 participants, 194 had at least one sort of dementia. Some of the participants had several recording sessions. Thus, the dementia category consists of 309 transcripts. Since this study deals with AD diagnosis, only the transcripts of patients with AD are taken. So, 255 CHAT transcripts belonging to 168 AD patients were taken. Similarly, there were 244 transcripts from 98 CN participants used in this experiment. Furthermore, all the participants of Becker's [27] study were over the age of 44 and had a minimum MMSE score of 10. The demographic information about the subjects under study can be shown in Table 1. The mean of attributes, along with their

standard deviation, can also be found in Table 1.

Table 1. Demographic information of subjects from Dementia Bank

Attributes	Control Normal (CN)	Alzheimer's Disease (AD)
No. of participants	98 CN subjects	168 AD patients
Gender	31M / 67F	55M / 113F
Age	64.7 (7.6)	71.2 (8.4)
Education	14.0 (2.3)	12.2 (2.6)
MMSE	29.1 (1.1)	19.9 (4.2)

Originally, the available transcripts of the recordings are available in the English language, which was translated into the Nepali language for this study. The translations were done by two native Nepali language speakers who had at least 13 years of formal education in the Nepali language. Translating the entire dataset took around eighty-four hours. After the entire translation, they were again sent for verification to an independent linguistic expert, who assessed and verified the adequacy of the translations. We chose manual translation, as this has a better chance of capturing cultural nuances required for the target language. Justification in the literature will be shown in the Discussion section. Here, we will urge the readers to take this at face value and move forward.

Fig. 2 Cookie Theft Picture

An example of the manual translation of each category is shown in Table 2(a). Words such as uhm, uhh, and other pause words were not removed and translated as it is. Such words are not removed from translation to retain more accurate transcripts of the actual recordings. Khodabakhsh [30]

suggested that such filter sounds like uhm, uhh, etc., are used more often by AD subjects and form a significant feature in their speech. AD patients tend to use longer pauses than cognitively normal individuals. The repetitions, linguistics, and syntactic errors and the words that depict confusion of the AD participants have been translated as they are to retain the originality. The transcripts have been translated in a way that maximum linguistic characteristics are preserved. However, the annotations in the transcripts, such as clears throats, laughs, etc., have not been included as they are not a part of the linguistic feature used by the subjects. The ten most frequently used words are given in Table 3, with their number of occurrences.

We have also developed a dataset by machine translation (google translate) for comparison. The manual and machine translations are being compared in Tables 2(a) and Tables 2(b). The translations from google translate did not show the accurate translation of the text. Moreover, in many of the NLP tasks that require annotations, machine translation fails to show accurate translation at par with human translation. Even though the translations involving deep learning methods provide substantial advantages, they still lack human performance on data that require cultural nuances to be preserved [31]. Hence, manual translation was incorporated into the study for better perseverance and assertion of the general tone of the texts. The word clouds of the English and Nepali texts are shown in Fig 3. Fig. 3 (a) shows the word cloud of the transcripts of CN individuals, and Fig. 3 (b) shows the word cloud of the transcripts of AD patients in the English language. Similarly, Fig. 3 (c) and Fig. 3 (d) represent the transcripts in the Nepali language by CN and AD subjects, respectively.

Table 2 (a). Examples of the Manual Translation of the DementiaBank

CHAT ID	English Sentence	Nepali Sentence
015-2.cha	you have two children and the boy is on a stool getting to the cookie jar. and the stool is tilting over and he's probably going to fall. his mother in the meantime is wiping dishes, looking out what is obviously the kitchen window. she has the water on in the sink and the sink is overflowing. there are two cups and one plate sitting on the sink. the little girl is laughing at the little boy who's getting into the cookie jar and is going to fall.	दुईजना बच्चाहरू छन् र केटो कुर्सीमा चढेर कुकी जारबाट कुकी लिन खोजिरहेको छ कुर्सी बाङ्गिएको छ र ऊ सायद लड्ने वाला छ उसकी आमा यो समयमा भाँडा पुस्दै छे ऊ झ्याल बाहिर हेरिरहेकी छे सिंकमा पानी छ र पानी भरिएको छ सिंकमा दुईवटा कप र एउटा प्लेट छ सानी केटी सानो केटो माथि हाँसिरहेकी छे ऊ कुकी लिँदै छ र लड्ने वाला छ
472-0.cha	the boy and the girl are playing and he's gonna fall down off the ladder. and the mother's washing the dishes and it's flying out over the sink down to the floor. what else do you want me to tell you whatever you see happening. yeah that's it.	केटो र केटी खेलिरहेका छन् अनि ऊ भऱ्याङ बाट लड्नेवाला छ अनि आमा भाँडा माझिरहेकी छे अनि त्यो सिंकबाट माथी उडिरहेको छ, भुइँमा पुगिरहेको छ अरू के चाहन्छौ कि म भनौँ भनेर.. यत्ति हो

Table 2 (b). Examples of the Google Translate Translations of the DementiaBank

CHAT ID	English Sentence	Nepali Sentence
015-2.cha	you have two children and the boy is on a stool getting to the cookie jar. and the stool is tilting over and he's probably going to fall. his mother in the meantime is wiping dishes, looking out what is obviously the kitchen window. she has the water on in the sink and the sink is overflowing. there are two cups and one plate sitting on the sink. the little girl is laughing at the little boy who's getting into the cookie jar and is going to fall.	तपाईंका दुई बच्चाहरू छन् र केटा स्टूलमा कुकी जारमा पुगिरहेको छ। र मल माथि झुकेको छ र ऊ सायद खस्दैछ। यस बीचमा उनकी आमा भान्साकोठा पुछिरहेकी छिन्, स्पष्ट रूपमा भान्साको झ्याल हेर्दै। उसको सिङ्कमा पानी छ र सिङ्क भरिएको छ। त्यहाँ सिङ्कमा दुई कप र एउटा प्लेट छ। सानो केटी कुकीको भाँडोमा पसेको सानो केटालाई देखेर हाँस्दै छ।
472-0.cha	the boy and the girl are playing and he's gonna fall down off the ladder. and the mother's washing the dishes and it's flying out over the sink down to the floor. what else do you want me to tell you whatever you see happening. yeah that's it.	केटा र केटी खेलिरहेका छन् र ऊ भर्याङबाट तल खस्नेछ। र आमाले भाँडा धुँदै हुनुहुन्छ र यो सिङ्क माथि भुइँमा उडिरहेको छ। अरु के चाहान्छौ म तिमीलाई जे भइरहेछ देख्छु। हो त्यो हो।

Table 3: top 10 most used words in the transcript

Words	Number of Appearances
कुकी (cookie)	1092
भाँडा (utensil)	588
पानी (water)	580
केटो (boy)	448
आमा (mother)	384
केटी (girl)	326
कुर्सी (chair)	276
बाहिर (outside)	252
सानी (small)	222
सायद (maybe)	213

(a) Word Cloud of CN English Text

(b) Word Cloud of AD English Text

(c) Word Cloud of CN Nepali Text

(d) Word Cloud of AD Nepali Text

Fig 3. The word clouds of the English and Nepali texts

3.2 Data Preprocessing

The preprocessing step usually includes the removal of filter words, unnecessary noise, and unwanted information that does not add any value to the true meaning of the text [32]. In the English language, a major preprocessing step would be to make all the words uppercase or lowercase. The Nepali language is a case insensitive language. Hence, it does not require any such conversion. In this study, as a preprocessing step, the punctuation marks like commas, semicolons, etc., that do not add any semantic meaning to the text are removed. Another general practice in classification tasks using NLP involves removing stop words, which usually helps improve performance metrics. Since AD vs. CN is also a text classification task, anyone would think of proceeding with preprocessing the text by removing the stop words. The domain knowledge of AD helps tackle the ways to preprocess the CHAT transcripts used in this study. The AD patients tend to repeat stop words like ‘and, ‘therefore,’ etc. more often, and in this experiment, stop words are not removed since they preserve the linguistic characteristics of AD patients [30].

3.3 Feature Extraction

Text feature extraction is the process of extracting a list of words and creating a vocabulary from the text data [33]. These words are transformed into a feature set that a classifier can use. In this experiment, word statistics-based feature extraction techniques have been used. Vectorization techniques are used to transform the words into vectors. They give positional weights to the words used in the text data. Similarly, word embeddings are a way of transforming words into vectors by capturing the similarity between words. Words with similar meanings appear in the same feature space. The various feature extraction techniques used in this experiment have been discussed below:

3.3.1 Vectorization Methods

The experiment uses two popular vectorization methods, namely CountVectorizer and Term Frequency Inverse Document Frequency (TF-IDF). CountVectorizer is used to build a dictionary of known words from the test dataset. It is also used to encode the new documents using the vocabulary [34]. CountVectorizer tokenizes and creates a respective vector representation of each word fed to a machine learning model. TF-IDF is another popular vectorization technique for generating vector representations of the text [35]. TF-IDF represents the importance of a word to a document. It does so by being able to count the number of occurrences. TF-IDF punishes the words that are used very often in the documents, hence being able to give more weightage to the words that are more relevant and important to a particular document.

3.3.2 Word embeddings

After the text is preprocessed, real-valued vectors are assigned to words or phrases using word embeddings. Word embeddings are based on the idea that if features have similar meanings, it is useful to represent the features to depict this similarity [36]. Bengio et al. [37] proposed a probabilistic neural model where the words in the vocabulary were mapped to a distributed word feature vector. The feature vector represents several aspects of the word. These features are smaller than the size of the vocabulary. This study makes use of the two most efficient word embeddings viz. Word2Vec and fastText. Both the pre-trained and domain-specific Word2Vec [38] and fastText [39] models are trained to produce embeddings.

Domain-specific Embeddings: Domain-specific embeddings are trained on the dataset being used. It has been found that the pre-trained word embeddings perform very well in a large text corpus, but in sparse and specialized texts, the pre-trained word embeddings generally fail to produce appropriate vectors [40]. In this study, 300-dimensional embeddings have been used for both Word2Vec and fastText embeddings, and the maximum length is set to 270. Gensim [41] library is used to generate Word2Vec and fastText models from the text used in the study.

Pre-trained Embeddings: The pre-trained Nepali Word2Vec model created by Lamsal [42] is used in the study. This pre-trained Word2Vec model has 300-dimensional vectors for more than 0.5 million Nepali words and phrases. The embedding dimension is 300, and continuous bag-of-words (CBOW) architecture was used to create the given Word2Vec model. Similarly, for the pre-trained fastText embeddings, the pre-trained word vectors trained on Common Crawl and Wikipedia using fastText were used [39]. The model was trained by using CBOW with position-weights, in dimension 300, with character n-grams of length 5, a window of size 5, and 10 negatives.

3.3 Learning Models (Classifiers)

For the classification of the transcripts of CN and AD patients, some learning models should be

used. In this paper, both machine learning models and deep learning models were used to find the better model that would classify the transcripts with greater accuracy.

3.3.1 Machine Learning Baselines

Since the work of such classification in the Nepali language is the first of its kind, machine learning baselines are taken to evaluate the performance of machine learning algorithms in the delineation of the transcripts of AD patients from that of CN subjects. The machine learning algorithms like Decision Tree (DT) [43], K-Nearest Neighbors (KNN) [44], Support Vector Machines (SVM) [45], and Naïve Bayes (NB) [46] were used. Also, ensemble learners like Random Forest (RF) 43], AdaBoost [47], and XGBoost (XGB) [48] were used. Apart from the traditional vectorization techniques, such as CountVectorizer and TF-IDF vectorizer, word embeddings were also used for vectorizing the input text to feed them to the machine learning model.

3.3.2 Deep Learning Models

The deep learning models have recently shown very promising results in text classification, especially when the classification tasks deal with intrinsic and complex details of the linguistic features in the text. In our experimentation, three deep learning models have been used, viz, Convolutional Neural Network (CNN) [49], Bidirectional Long Short-Term Memory (BiLSTM) [50], and a combination of CNN and BiLSTM [51].

Convolutional Neural Network: Convolutional Neural Network (CNN) is a deep neural network architecture that uses layers with convolving filters. CNNs have been traditionally used for computer vision for identifying images. However, CNNs have also proven to be significantly useful for NLP. They have been used for various NLP tasks such as semantic parsing, search query retrieval, sentence modeling, and other traditional NLP tasks. The convolving filters, as well as applying max-pooling extract relevant n-gram features of the texts used. The input of the convolutional layer is the vector produced by word embeddings. A one-dimensional Convolutional Neural Network has been used in this study. As a 300-dimensional embedding with a maximum length of 270 has been used in this study, the input size is a matrix of size 270x300. Four convolutional layers with ReLu as the activation function have been used, and after every two layers, a max-pooling [52] of size three was done. For kernel regularization, L2 regularizers have been used. The optimizer used is Adam [53]. After flattening the convolutional layers, the output is connected to a fully connected dense layer. The softmax function is used in the output layer to predict the probabilities of the CN and AD categories. The number of epochs and batch size has been fixed to 20 and 50, respectively, for all embeddings.

Kim's Architecture: Apart from the CNN model mentioned above, the famous Kim's CNN architecture [54] has also been used. In Kim's architecture, after every convolutional layer, max-pooling is applied. In this experiment, three convolutional layers with tanh as the activation function have been used, and after each layer, a max pool of filter size three has been applied. After the last max pool filter, the flatten layer reshapes the input size, followed by the dropout layer with a rate of 0.5. The dropout layer randomly sets inputs to 0 and prevents overfitting. The output layer has softmax as the activation function that transforms the results into probabilities of each class. The number of epochs and batch size has been fixed to 20 and 50, respectively.

BiDirectional Long Short-Term Memory (BiLSTM): Unlike Long Short-Term Memory (LSTM) [55], in BiLSTMs, the signal propagates in both directions, i.e., backward and forward. BiLSTMs train first on the input sequence and then on the reversed input sequence. The forget, input, and

output gates and the cell states decide what information to throw away, update the cells, and then produce the output by carrying only the relevant information. In this work, four BiLSTM cells with 16, 8, 4, and 2 nodes subsequently and tanh as the activation function have been used. The input is the same as the convolutional layer, i.e., 270x300 dimensional vector of word embeddings. After the first BiLSTM layer, a dropout with a 0.5 rate has been used for regularization. After the three BiLSTM layers, again, a dropout of 0.25 has been used. The output of the BiLSTM cell has been connected to a dense layer with four nodes and ReLU as the activation function. The output layer has softmax as the activation function in order to predict probabilities for the two categories. To prevent overfitting, L2 regularizers have been used. Similarly, for optimization, an Adam optimizer has been used. The number of epochs and batch size has been fixed to 20 and 50, respectively, for all embeddings.

CNN with BiLSTM cells: CNNs learn the local features of the text, and RNNs learn long-term dependencies. Combining these architectures can better perform in various NLP tasks such as sentiment analysis and text classification [56]. In this experiment, four convolutional layers and two BiLSTM cells have been used. The word embeddings are fed to the convolutional layer. After every two convolutional layers, a max-pooling of size three has been applied. To prevent overfitting, L2 regularizers have been used in both networks. Tanh has been used as the activation function for the BiLSTM cells. After the first BiLSTM cell, batch normalization [57] has been done. Adam optimizer has been used. The output of the BiLSTM cell has been connected to a fully dense layer with ReLu as the activation function and twenty nodes. One more dense layer has been added with ReLU as the activation function with ten nodes. The softmax function in the output layer transforms the vectors to predict the category of the transcripts. The number of epochs and batch size has been fixed to 20 and 50, respectively.

Deep Learning Models with Attention Mechanisms: Attention mechanisms are used in encoder-decoder architectures to attend to the encoder and previous hidden states. With an input sentence and all the associated hidden states, attention layers decide what part of the input was most relevant and useful with each output instance. Attention preserves the context from beginning to end hence achieving great results on various NLP tasks such as machine translation [58], text summarization [59], text classification, etc. All the deep learning models used in this study have also been trained with an attention layer [60]. Apart from attending to the encoder and previous hidden states, attention can also be used to get a distribution over features, such as the word embeddings of a text [61]. The attention used in this study is the multiplicative self-attention layer because of its space efficiency and less operation time. Self-attention [62] is used to extract the relevant features by enabling it to attend to itself. The architecture of the models is the same as described above, with only an attention layer after the first layers in every model.

Deep Learning Models with Vectorization Techniques: Apart from pre-trained and domain-specific word embeddings, deep learning classifiers were also fed with the vectorized texts done by CountVectorizer and TF-IDF vectorizer as inputs. The input was a matrix of dimensions (499, 270). The remaining layers of the architecture of the deep learning models were kept the same as described above.

3.5 Performance Measures

In all the architectures afore-mentioned, binary cross-entropy has been used as the loss function. The validation has been done using stratified k-fold cross-validation. The stratified K-fold cross-validation is a variation of k-fold cross validation that returns stratified folds in which the

percentage of samples of each class is preserved. After stratified 10-fold cross-validation, the performance of the proposed architectures has been measured using four evaluation metrics viz. accuracy (acc), precision (pre), recall (rec), and F1-score as shown in equations (1)-(4).

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} \tag{1}$$

$$Precision = \frac{TP}{TP+FP} \tag{2}$$

$$Recall = \frac{TP}{TP+FN} \tag{3}$$

$$F1-score = 2 * \frac{Precision*Recall}{Precision+Recall} \tag{4}$$

where TP, TN, FP, and FN represent True Positive, True Negative, False Positive, and False Negative respectively.

4. Results and Discussion

The baseline is established with various machine learning algorithms. TF-IDF, CountVectorizer (CV), Word2Vec, and FastText were used to convert the text document into vectors for the experiment. The results of the machine learning baselines with the TF-IDF and CountVectorizer are shown in Table 4. For machine learning models, the Naive Bayes classifier performed the best for both the vectorization techniques. With CountVectorizer, the model had an F1-score of 0.940. With TF-IDF vectorization as well, the model was able to achieve an F1-score of 0.940. TF-IDF seemed to perform slightly better when vectorization methods are compared than CountVectorizer for different machine learning models. A possible explanation for this is that TF-IDF, instead of just representing words with vectors in terms of their number of appearances, balances the most frequent words by giving them less weightage. Rarer words common in a particular class would be scored higher, eventually leading to better performance of models.

Similarly, with domain-specific Word2Vec word embeddings, the decision tree performed the best with an F1-score of 0.937. As far as pre-trained word embeddings are concerned, pre-trained Word2Vec performed the best with the XGBoost algorithm giving an F1-score of 0.828. On the other hand, with pre-trained fastText, the SVM classifier outperformed other models with an F1-score of 0.934. It can be inferred from the comparison of vectorization techniques with word embeddings that vectorization techniques performed better than word embeddings with machine learning models. The reason behind this is that the data corpus was small to train the word embeddings. Hence, the similarity between words is not captured well. The results with Word2Vec and FastText embeddings with the machine learning models are shown in Table 5 and Table 6, respectively.

Table 4. ML Classifiers with CV and TF-IDF

ML Classifier	TF-IDF				Count Vectorizer (CV)			
	Accuracy	Precision	Recall	F1-score	Accuracy	Precision	Recall	F1-score
DT	0.900	0.903	0.900	0.900	0.854	0.855	0.854	0.854
KNN	0.910	0.918	0.910	0.909	0.870	0.892	0.870	0.868
SVM	0.936	0.939	0.936	0.936	0.902	0.907	0.902	0.902
NB	0.940	0.944	0.940	**0.940**	0.940	0.945	0.940	**0.940**

RF	0.934	0.937	0.934	0.934	0.936	0.939	0.936	0.936
ADB	0.886	0.889	0.886	0.886	0.904	0.907	0.904	0.904
XGB	0.926	0.929	0.926	0.926	0.920	0.924	0.920	0.920

Table 5. ML Classifiers with Word2Vec

ML Classifiers	Domain-Specific Word2Vec				Pre-trained Word2Vec			
	Accuracy	Precision	Recall	F1-score	Accuracy	Precision	Recall	F1-score
Decision Tree	0.938	0.949	0.938	**0.937**	0.695	0.698	0.695	0.694
KNN	0.718	0.780	0.716	0.694	0.547	0.564	0.547	0.464
SVM	0.940	0.947	0.940	0.931	0.764	0.771	0.764	0.760
Naïve Bayes (Gaussian)	0.891	0.930	0.892	0.890	0.529	0.582	0.529	0.414
Random Forest	0.902	0.912	0.902	0.901	0.788	0.797	0.782	0.785
AdaBoost	0.890	0.898	0.890	0.889	0.754	0.766	0.754	0.751
XGBoost	0.918	0.924	0.918	0.917	0.826	0.836	0.826	**0.828**

As far as the deep learning models are concerned, they seem to have performed better than the machine learning models in the experiments. The initial experiments with deep learning models showed that with domain-specific Word2Vec, Kim's Architecture had the best F1-score of 0.964. Similarly, with pre-trained word embeddings, pre-trained fastText with the BiLSTM model outperformed other models with pre-trained embeddings with an F1-score of 0.887. The deep learning models performed slightly better with domain-specific word embeddings than pre-trained embeddings. As domain-specific word embeddings are formed from the data corpus, it can capture the domain words well and perform better. The dataset is based on the cookie theft description task, and hence it contains words related explicitly to the problem than general words. The results of the deep learning models with domain-specific and pre-trained word embeddings are shown in Table 7.

Table 6. ML Classifiers with fastText

ML Classifiers	Domain-Specific fastText				Pre-trained fastText			
	Accuracy	Precision	Recall	F1-score	Accuracy	Precision	Recall	F1-score
Decision Tree	0.674	0.683	0.673	0.668	0.794	0.799	0.794	0.792
KNN	0.599	0.639	0.599	0.56	0.908	0.913	0.908	0.907
SVM	0.739	0.747	0.739	0.737	0.934	0.944	0.934	**0.932**
Naïve Bayes (Gaussian)	0.531	0.575	0.531	0.419	0.523	0.56	0.523	0.400
Random Forest	0.7876	0.797	0.788	0.7855	0.914	0.92	0.914	0.913
AdaBoost	0.755	0.762	0.755	0.753	0.918	0.925	0.918	0.917
XGBoost	0.826	0.837	0.826	**0.824**	0.914	0.924	0.914	0.912

Table 7. Deep Learning Models with word embeddings

Deep Learning Models		Word2Vec				fastText			
		Accuracy	Precision	Recall	F1-score	Accuracy	Precision	Recall	F1-score
Domain-Specific	Kim's CNN	0.964	0.965	0.964	**0.964**	0.936	0.937	0.936	**0.936**
	CNN	0.950	0.952	0.950	0.950	0.928	0.930	0.928	0.928
	BiLSTM	0.946	0.948	0.946	0.946	0.900	0.910	0.900	0.897
	CNN + BiLSTM	0.962	0.964	0.962	0.962	0.928	0.931	0.928	0.927
Pretrained	Kim's CNN	0.828	0.834	0.828	0.827	0.870	0.880	0.870	0.867
	CNN	0.861	0.865	0.861	0.861	0.866	0.877	0.866	0.864
	BiLSTM	0.872	0.884	0.872	0.869	0.888	0.892	0.888	**0.887**
	CNN + BiLSTM	0.872	0.880	0.872	**0.871**	0.756	0.839	0.756	0.727

Table 8. Deep Learning Models with vectorizers

Deep Learning Models	CountVectorizer				TF-IDF Vectorizer			
	Accuracy	Precision	Recall	F1-score	Accuracy	Precision	Recall	F1-score
Kim's Architecture	0.897	0.900	0.897	**0.897**	0.847	0.852	0.847	**0.846**
CNN	0.793	0.798	0.793	0.773	0.731	0.755	0.731	0.727
BiLSTM	0.735	0.738	0.735	0.733	0.738	0.743	0.738	0.726
CNN+BiLSTM	0.83	0.849	0.832	0.831	0.732	0.796	0.732	0.717

Apart from the initial experiments with word embeddings for the deep learning models, they were also trained with vectorization techniques. As done with machine learning models, CountVectorizer and TF-IDF were used to vectorize the words for deep learning models. In this case, CountVectorizer outperformed TF-IDF with Kim's CNN architecture. The model was able to achieve an F1-score of 0.897. Kim's CNN contains max-pool filters after each convolution operation. This potentially extracts just the relevant features with reducing dimensionality simultaneously. The performance of the models with vectorizers is shown in Table 8.

Attention mechanisms were also applied to deep learning models in the experiments. With attention mechanisms, CNN with Word2Vec showed the best performance with an F1-score of 0.968. From the results obtained, it can be seen that the attention mechanism gave the best results, implying that more weightage was given to those words which carried more importance in the sentence. Also, CNN outperformed the other models with attention giving an idea that the features captured by the model were just the relevant ones, and only they were attended to. With attention as well, domain-specific Word2Vec performed better than pre-trained word embeddings. The results of the deep learning models trained with word embeddings and attention are shown in Table 9.

Table 9. Attention with DL models and word embeddings

Deep Learning Models		Word2Vec				fastText			
		Accuracy	Precision	Recall	F1-score	Accuracy	Precision	Recall	F1-score
Domain specific	Kim's Architecture	0.954	0.955	0.954	0.954	0.924	0.927	0.923	0.923
	CNN	0.968	0.969	0.968	**0.968**	0.932	0.933	0.932	**0.932**
	BiLSTM	0.956	0.959	0.956	0.956	0.923	0.934	0.924	0.929
	CNN+ BiLSTM	0.962	0.962	0.962	0.962	0.922	0.928	0.922	0.921
Pre-Trained	Kim's Architecture	0.890	0.907	0.900	0.900	0.922	0.924	0.922	0.922
	CNN	0.902	0.901	0.902	0.901	0.902	0.904	0.907	0.902
	BiLSTM	0.913	0.918	0.913	0.913	0.764	0.825	0.784	0.767
	CNN+ BiLSTM	0.890	0.904	0.890	0.887	0.660	0.711	0.660	0.572

Table 10. Attention with vectorizer in DL models

Deep Learning Models	CountVectorizer				TF-IDF			
	Accuracy	Precision	Recall	F1-score	Accuracy	Precision	Recall	F1-score
Kim's Architecture	0.701	0.712	0.701	0.702	0.627	0.627	0.620	0.623
CNN	0.765	0.763	0.765	**0.766**	0.699	0.699	0.689	0.632
BiLSTM	0.629	0.634	0.629	0.616	0.723	0.723	0.717	**0.729**
CNN+BiLSTM	0.678	0.682	0.678	0.685	0.713	0.713	0.710	0.721

The word embeddings and deep learning models used with vectorization techniques were also trained by using an attention layer. The CNN model with CountVectorizer had the highest F1-score of 0.766 in this experiment. From the results obtained, it can be inferred that attention methods with vectorization did not perform well as they did with word embeddings. When attention layers are applied, attending to features with vectors as the representations of the number of appearances of words, does not perform well. Hence, with attention, the vectorization techniques had lower F1-scores overall. The results with attention-based deep learning models with vectorization techniques are shown in Table 10.

It can be seen from the results that the best performing model is the attention-based CNN with domain-specific Word2Vec. This model can be utilized in making a clinician-friendly application for helping them with identifying Alzheimer's disease in its earliest stages. A pipeline with a mechanism for speech synthesis could also be developed with the given methodologies for better detection of the disease.

5. Discussions and Future Works

5.1 Discussions

Computational methods are very significant for clinical research and enable healthcare professionals to make clinical decisions about disease identification. There are many emerging

success stories in NLP applications in the English language. This motivates researchers to bring about results in the clinical domain, other than the English language. Hence, to put another dimension, the dataset in the clinical research in the domain of Alzheimer's disease in the Nepali language is substantial for further research. The work presented in this paper contributes to the growing body of research in the area of clinical applications in AD, albeit a preliminary one.

The model provides a methodology for early detection of AD using translated corpus in a low resource language, which in this case is Nepali. With such a model, we would be able to label, identify, and provide an early detection system in low-resourced languages. For this project, we were able to reach a reasonable accuracy even without specifically collecting data in the local language. The approach makes use of an AD corpus in English and translates the corpus into Nepali using human translators. We chose manual translation, as this has a better chance of capturing cultural nuances required for the target language. As there is very little amount of text data in languages such as Nepali for NLP tasks, manual translation was an effective method for collecting data. Even in the best of times, manual data collection in the field requires a lot of resources. Now, it is even more challenging to manually collect the dataset from patients in Nepal, given the ongoing pandemic, and hence the available data were used. This might seem like a contradiction, and also raises the question of why not use machine learning to do the translation at all. After all, using machine translations would be a full machine learning approach to detecting Alzheimer's using natural language processing. However, the literature search did not support this approach with the amount of data available.

Earlier, Guzman et.al. [63] experimented with machine translations of the FLORES dataset, and the results were poor, indicating more annotations and preparatory work might be needed before embarking on machine translations. On the other hand, there have been hundreds of years of successful manual translations [64][65] even to low-resource languages such as Nepali. There are also previous studies that supported manual translation as a better option at least in the early stages of working with low resource languages. Mohammad et al. [31] did a sentiment analysis in the Arabic language utilizing the available English texts and showed competitive results with the manually translated Arabic data. Similarly, Balahur et al. [66] did a sentiment analysis on four different languages, namely, Italian, Spanish, German, and French, by translating English data. They considered the manual translation as the gold standard of their datasets. Some machine translations have been attempted between English and Nepali [67]. In terms of low-resource datasets, DARPA programs like LORELEI [68] and the Asian Language Treebank project [69] have collected and introduced translations on several low-resource languages. However, these are still in the early stages, the coverage is still low, and do not include Nepali. Also, such machine learning translation systems need improvements before they can be employed on their own. There are reports of about 68% accuracy on TDIL (Technology Development for Indian Languages). Some studies employed English-Nepali parallel corpus for machine translation. Therefore, it would be some time before we can use the machine-based approach entirely and eliminate manual translations.

The Nepali dataset thus derived by manual translation works well with early detection of Alzheimer, and is a good candidate for creating a baseline for detecting Alzheimer's disease in the Nepali language since there is no available data for use for the purpose. As mentioned earlier, the translations were later verified by a linguistic expert who is currently working at the University of Auckland. The expert corroborated that the translations preserved the emotions of the participants. Moreover, the overall intonation of the text has been maintained, and hence there were no cultural

inconsistencies. Therefore, the experiments were finally conducted with the assurance from the expert about the dataset. The expert review should not be surprising, as manual translators could address both the message as well as cultural meaning [70]. The reasons for better performance through translations could be postulated based on a number of factors. Firstly, English has come to be used and studied worldwide, and Nepali speech has a significant degree of code-switching that occurs. Code-mixing [71] is common in Nepal, and code-switching between English and Nepali is fairly common among urban and educated Nepali speakers [72]. Medical professionals tend to fall in these categories. So much so, that the Gurung [72] reports extensive code-switching and code-mixing, argues that Nepali-English mixed language has emerged as a dialect in the Nepali speech community through the recurrent use of the English elements in the Nepali conversation.

In addition, Nepali is one of the several languages spoken in Nepal and is *lingua franca*, i.e., a common language [73][74][75]. The multi-lingual nature of Nepal's landscape, along with code-mixing make the speakers familiar with or have evolved, cultural insertions from English. The machine translation of the AD corpus has an inherent limitation, as the cultural nuances are harder to replicate algorithmically. Without a significantly annotated corpus, the machine translations will not capture cultural and linguistic nuances native to the target language. This will lower the accuracy of AD detection. However, in this particular case, the translations were carried out by native Nepali speakers with 13 years of formal education in that language and verified by a linguistic expert. This is the strength of the research. There are some syntactical differences between English and Nepali, especially in the order and placement of elements. For example, English follows the default word order of subject-verb-object (SVO). Whereas in Nepali the default word order is subject-object-verb (SOV). That is, in Nepali, the verb occurs at the end of a sentence. In English, the object complements the verb and occurs after the verb (to the right), while in Nepali, the object occurs to the left of the verb. Nepali nouns following numerals will be marked for plurality. While translating, considerations have been given to such structural differences between English and Nepali grammar. This can be relatively easily codified as described by researchers [76]. The features like hesitation and puzzlement, Part of Speech (POS) based features, unintelligible word rate, complexity features like phonemes per word, etc. were taken care of. The performance has been improved through human translators. However, using human translators and linguists does take time, although considerably lower than collecting primary data in Nepali and annotating them.

5.2 Future Works

The future work could include actually getting medical practitioners to verify and validate the translated corpus as representative of actual patient's language usage. The exercise would provide validation as well as promote understanding of the richness and appropriateness of the translation-based approach. In addition to medical experts, it is also possible to combine alternative approaches such as MR-based image recognition of neuroanatomy to build multimodal systems. The NLP-based model may still benefit from improvements in the form of injection of native features to strengthen the translated corpus. In the future, we will also assess, if any cultural or linguistic features are missing in translation, and accordingly, inject language and culturally specific features of Nepali (low-resource) language into the translated corpus (as part of translation and processing).

In addition, in the long term, primary data of corpus can be developed in the low resource language, in this case, Nepali. This native corpus can be compared with the translated corpus for similarities.

Also, developing a speech recognition system that helps to analyze the speech of people can be a direct method for early detection of Alzheimer's disease. Also, the work can be extended to other forms of dementia, such as Parkinson's among Nepalese patients. When collecting data as the primary source, the program should be well designed otherwise it would amplify the advantage of one sub-groups over others. Through this approach, we can plan where the gaps are and compensate for collecting data. This way, it would nullify the enforcement of social disadvantages caused by any normative biases and finally expand a language to improve the ML technology in the domain. While these improvements will make the solution more efficient, it is also advisable that in order to improve efficiency, a combined human-machine translation be explored.

6. Conclusion

Detecting Alzheimer's disease at its earliest stage is still a challenging task. Speech degeneration, being one of the most common and earliest symptoms in AD patients, should be leveraged to identify the disease. Since there is no clinical medicine or method to cure the disease completely, the only practical way would be to identify it in its early stage to stop the progression of the disease. Hence, the study aims to detect AD early for the people who speak the Nepali language. This is a step towards solving problems in identifying the disease and motivation for further researchers working in this field. The significant advantage of this automated system is that it takes significantly less time to predict the presence of AD. Also, the treatment costs are highly reduced and can be used over a large number of cycles for many people. The further improvement in the study can include acoustic features such as the duration of pause a person takes while speaking, how confused his words sound, etc. Also, developing a speech recognition system that helps to analyze the speech of people can be a direct method for early detection of Alzheimer's disease. Also, the work can be extended to other forms of dementia, such as Parkinson's among Nepalese patients. It is especially vital because healthcare service is not very useful in the country. Thus, it can help the health specialist in their decision-making and reduce the time and cost associated with the identification of the disease.

Conflicts of Interest: The authors declare no conflicts of interest.

References

[1] X. Zhou and J. W. Ashford, "Advances in screening instruments for Alzheimer's disease," *Aging Med.*, vol. 2, no. 2, pp. 88–93, 2019, doi: 10.1002/agm2.12069.

[2] P. Benefits, "2018 ALZHEIMER'S DISEASE FACTS AND FIGURES Includes a Special Report on the Financial and Personal Benefits of Early Diagnosis," 2018.

[3] H. Liu-Seifert *et al.*, "Disease Modification in Alzheimer's Disease: Current Thinking," *Ther. Innov. Regul. Sci.*, vol. 54, no. 2, pp. 396–403, 2020, doi: 10.1007/s43441-019-00068-4.

[4] B. C. Dickerson *et al.*, "The cortical signature of Alzheimer's disease: Regionally specific cortical thinning relates to symptom severity in very mild to mild AD dementia and is detectable in asymptomatic amyloid-positive individuals," *Cereb. Cortex*, vol. 19, no. 3, pp. 497–510, 2009, doi: 10.1093/cercor/bhn113.

[5] S. Thapa, P. Singh, D. K. Jain, N. Bharill, A. Gupta, and M. Prasad, "Data-Driven Approach based on Feature Selection Technique for Early Diagnosis of Alzheimer's Disease," *Proc. Int. Jt. Conf. Neural Networks*, 2020, doi: 10.1109/IJCNN48605.2020.9207359.

[6] K. Domoto-Reilly, D. Sapolsky, M. Brickhouse, and B. C. Dickerson, "Naming impairment in Alzheimer's disease is associated with left anterior temporal lobe atrophy," *Neuroimage*, vol. 63, no. 1, pp. 348–355, 2012, doi: 10.1016/j.neuroimage.2012.06.018.

[7] G. McKhann, D. Drachman, M. Folstein, R. Katzman, D. Price, and E. M. Stadlan, “Clinical diagnosis of alzheimer’s disease: Report of the NINCDS-ADRDA work group⋆ under the auspices of department of health and human services task force on alzheimer’s disease,” *Neurology*, vol. 34, no. 7, pp. 939–944, 1984, doi: 10.1212/wnl.34.7.939.

[8] K. Fraser and G. Hirst, “Detecting semantic changes in Alzheimer’s disease with vector space models,” *Proc. Lr. 2016 Work. Resour. Process. Linguist. Extra-Linguistic Data from People with Var. Forms Cogn. Impair.*, no. May, pp. 1–8, 2016.

[9] K. Faber-Langendoen, J. C. Morris, J. W. Knesevich, E. LaBarge, J. P. Miller, and L. Berg, “Aphasia in senile dementia of the alzheimer type,” *Ann. Neurol.*, vol. 23, no. 4, pp. 365–370, 1988, doi: 10.1002/ana.410230409.

[10] S. Ahmed, A. M. F. Haigh, C. A. de Jager, and P. Garrard, “Connected speech as a marker of disease progression in autopsy-proven Alzheimer’s disease.,” *Brain*, vol. 136, no. Pt 12, pp. 3727–3737, 2013, doi: 10.1093/brain/awt269.

[11] H. S. Kirshner, W. G. Webb, and M. P. Kelly, “The naming disorder of dementia,” *Neuropsychologia*, vol. 22, no. 1, pp. 23–30, 1984, doi: 10.1016/0028-3932(84)90004-6.

[12] G. Glosser, “Patterns of Discourse Production among Neurological Patients with Fluent Language Disorders,” *Brain Lang.*, vol. 40, pp. 67–88, 1990.

[13] J. P. R. Dick, R. J. Guiloff, and A. Stewart, “Mini-mental state examination in neurological patients,” *J. Neurol. Neurosurg. Psychiatry*, vol. 47, no. 5, pp. 496–499, 1984, doi: 10.1136/jnnp.47.5.496.

[14] J. E. Storey, J. T. J. Rowland, D. A. Conforti, and H. G. Dickson, “The Rowland Universal Dementia Assessment Scale (RUDAS): A multicultural cognitive assessment scale,” *Int. Psychogeriatrics*, vol. 16, no. 1, pp. 13–31, 2004, doi: 10.1017/S1041610204000043.

[15] S. J. Cano *et al.*, “The ADAS-cog in Alzheimer ’ s Disease clinical trials : Psychometric evaluation of the sum and its parts To cite this version : HAL Id : hal-00580696,” 2011.

[16] R. W. Heinrichs, “Current and Emergent Applications of Neuropsychological Assessment: Problems of Validity and Utility,” *Prof. Psychol. Res. Pract.*, vol. 21, no. 3, pp. 171–176, 1990, doi: 10.1037/0735-7028.21.3.171.

[17] S. Velupillai, H. Suominen, M. Liakata, A. Roberts, and D. Anoop, “Europe PMC Funders Group Using clinical Natural Language Processing for health outcomes research : Overview and actionable suggestions for future advances,” pp. 11–19, 2020, doi: 10.1016/j.jbi.2018.10.005.Using.

[18] S. O. Orimaye, J. S. M. Wong, K. J. Golden, C. P. Wong, and I. N. Soyiri, “Predicting probable Alzheimer’s disease using linguistic deficits and biomarkers,” *BMC Bioinformatics*, vol. 18, no. 1, pp. 1–13, 2017, doi: 10.1186/s12859-016-1456-0.

[19] V. Vincze *et al.*, “Detecting mild cognitive impairment by exploiting linguistic information from transcripts,” *54th Annu. Meet. Assoc. Comput. Linguist. ACL 2016 - Short Pap.*, no. August, pp. 181–187, 2016, doi: 10.18653/v1/p16-2030.

[20] J. Fritsch, S. Wankerl, N. Elmar, E. Polytechnique, and F. De Lausanne, “AUTOMATIC DIAGNOSIS OF ALZHEIMER ’ S DISEASE USING NEURAL NETWORK LANGUAGE MODELS Friedrich-Alexander-University Erlangen-Nuremberg , Germany,” pp. 5841–5845, 2019.

[21] J. Chen, J. Zhu, and J. Ye, “An attention-based hybrid network for automatic detection of Alzheimer’s disease from narrative speech,” *Proc. Annu. Conf. Int. Speech Commun. Assoc. INTERSPEECH*, vol. 2019-Septe, pp. 4085–4089, 2019, doi: 10.21437/Interspeech.2019-2872.

[22] J. H. Chen *et al.*, “Dementia-related functional disability in moderate to advanced parkinson’s disease: Assessment using the world health organization disability assessment schedule 2.0,” *Int. J. Environ. Res. Public Health*, vol. 16, no. 12, 2019, doi: 10.3390/ijerph16122230.

[23] J. Weiner, C. Herff, and T. Schultz, “Speech-based detection of Alzheimer’s disease in conversational German,” *Proc. Annu. Conf. Int. Speech Commun. Assoc. INTERSPEECH*, vol. 08-12-Sept, pp. 1938–1942, 2016, doi: 10.21437/Interspeech.2016-100.

[24] F. Rudzicz, L. C. Currie, A. Danks, T. Mehta, and S. Zhao, “Automatically identifying trouble-indicating speech behaviors in Alzheimer’s disease,” *ASSETS14 - Proc. 16th Int. ACM SIGACCESS Conf. Comput. Access.*, pp. 241–242, 2014, doi: 10.1145/2661334.2661382.

[25] J. Liu, J. Zhao, and X. Bai, “Syntactic Impairments of Chinese Alzheimer’s Disease Patients from a Language Dependency Network Perspective,” *J. Quant. Linguist.*, vol. 28, no. 3, pp. 253–281, 2021, doi: 10.1080/09296174.2019.1703485.

[26] A. Khodabakhsh, F. Yesil, E. Guner, and C. Demiroglu, “Evaluation of linguistic and prosodic features for detection of Alzheimer’s disease in Turkish conversational speech,” *Eurasip J. Audio, Speech, Music Process.*, vol. 2015, no. 1, 2015, doi: 10.1186/s13636-015-0052-y.

[27] K. I. James T. Becker, Francois Boller, Oscar L. Lopez, Judith Saxton, “The natural History of Alzheimer’s Disease,” *Sight and Sound*, vol. 26, no. 12, pp. 16–19, 2016, doi: 10.1177/1097184x09352181.

[28] B. MacWhinney and Carnegie Mellon University, “The CHILDES Project: Tools for Analyzing Talk. Part 1: The CHAT Transcription Format,” no. 2000, 2000.

[29] E. Giles, K. Patterson, and J. R. Hodges, “Performance on the Boston Cookie Theft picture description task in patients with early dementia of the Alzheimer’s type: Missing information,” *Aphasiology*, vol. 10, no. 4, pp. 395–408, 1996, doi: 10.1080/02687039608248419.

[30] A. Khodabakhsh, S. Kusxuoglu, and C. Demiroglu, “Natural language features for detection of Alzheimer’s disease in conversational speech,” *2014 IEEE-EMBS Int. Conf. Biomed. Heal. Informatics, BHI 2014*, pp. 581–584, 2014, doi: 10.1109/BHI.2014.6864431.

[31] S. M. Mohammad, M. Salameh, and S. Kiritchenko, “How translation alters sentiment,” *J. Artif. Intell. Res.*, vol. 55, pp. 95–130, 2016, doi: 10.1613/jair.4787.

[32] U. Naseem and K. Musial, “DICE: Deep intelligent contextual embedding for twitter sentiment analysis,” *Proc. Int. Conf. Doc. Anal. Recognition, ICDAR*, pp. 953–958, 2019, doi: 10.1109/ICDAR.2019.00157.

[33] U. Naseem, I. Razzak, and I. A. Hameed, “Deep Context-Aware Embedding for Abusive and Hate Speech detection on Twitter,” *J. Chem. Inf. Model.*, vol. 53, no. 9, pp. 1689–1699, 2019.

[34] A. Kulkarni and A. Shivananda, “Natural Language Processing Recipes,” *Nat. Lang. Process. Recipes*, pp. 67–96, 2019, doi: 10.1007/978-1-4842-4267-4.

[35] D. Isa, L. H. Lee, V. P. Kallimani, and R. Rajkumar, “Text document preprocessing with the bayes formula for classification using the support vector machine,” *IEEE Trans. Knowl. Data Eng.*, vol. 20, no. 9, pp. 1264–1272, 2008, doi: 10.1109/TKDE.2008.76.

[36] U. Naseem, K. Musial, P. Eklund, and M. Prasad, “Biomedical Named-Entity Recognition by Hierarchically Fusing BioBERT Representations and Deep Contextual-Level Word-Embedding,” *Proc. Int. Jt. Conf. Neural Networks*, 2020, doi: 10.1109/IJCNN48605.2020.9206808.

[37] Y. Bengio, R. Ducharme, and P. Vincent, “A neural probabilistic language model,” *Adv. Neural Inf. Process. Syst.*, no. July, 2001.

[38] T. Mikolov, I. Sutskever, K. Chen, G. Corrado, and J. Dean, “Distributed representations ofwords and phrases and their compositionality,” *Adv. Neural Inf. Process. Syst.*, no. October, 2013.

[39] E. Grave, P. Bojanowski, P. Gupta, A. Joulin, and T. Mikolov, “Learning word vectors for 157 languages,” *Lr. 2018 - 11th Int. Conf. Lang. Resour. Eval.*, pp. 3483–3487, 2019.

[40] A. Roy, Y. Park, and Sh. Pan, “Learning Domain-Specific Word Embeddings from Sparse Cybersecurity Texts,” 2017, [Online]. Available: http://arxiv.org/abs/1709.07470.

[41] B. Srinivasa-Desikan, “Natural Language Processing and Computational Linguistics: A practical guide to text analysis with Python, Gensim, spaCy, and Keras,” *Packt Publ.*, 2018.

[42] J. Bhatta, D. Shrestha, S. Nepal, S. Pandey, and S. Koirala, “Efficient Estimation of Nepali Word Representations in Vector Space,” *J. Innov. Eng. Educ.*, vol. 3, no. 1, pp. 71–77, 2020, doi: 10.3126/jiee.v3i1.34327.

[43] S. R. Safavian and D. Landgrebe, “A Survey of Decision Tree Classifier Methodology,” *IEEE Trans. Syst. Man Cybern.*, vol. 21, no. 3, pp. 660–674, 1991, doi: 10.1109/21.97458.

[44] S. Rajora *et al.*, “A Comparative Study of Machine Learning Techniques for Credit Card Fraud Detection Based on Time Variance,” *Proc. 2018 IEEE Symp. Ser. Comput. Intell. SSCI 2018*, pp. 1958–1963, 2019, doi: 10.1109/SSCI.2018.8628930.

[45] C. CORTES and V. VAPNIK, “Support-Vector Networks,” *Mach. Lang.*, vol. 7, no. 2, pp. 142–147, 1995, doi: 10.1111/j.1747-0285.2009.00840.x.

[46] I. Rish, “An Empirical Study of the Naïve Bayes Classifier An empirical study of the naive Bayes classifier,” *Cc.Gatech.Edu*, no. January 2001, pp. 41–46, 2014, [Online]. Available: https://www.cc.gatech.edu/~isbell/reading/papers/Rish.pdf.

[47] S. Adhikari, S. Thapa, and B. K. Shah, “Oversampling based Classifiers for Categorization of Radar Returns from the Ionosphere,” *Proc. Int. Conf. Electron. Sustain. Commun. Syst. ICESC 2020*, no. Icesc, pp. 975–978, 2020, doi: 10.1109/ICESC48915.2020.9155833.

[48] T. Chen and C. Guestrin, “XGBoost: A scalable tree boosting system,” *Proc. ACM SIGKDD Int. Conf. Knowl. Discov. Data Min.*, vol. 13-17-August-2016, pp. 785–794, 2016, doi: 10.1145/2939672.2939785.

[49] A. Krizhevsky, I. Sutskever, and G. E. Hinton, “ImageNet Classification with Deep Convolutional Neural Networks,” *Commun. ACM*, vol. 60, no. 6, pp. 84–90, 2012.

[50] A. Graves and J. Schmidhuber, “Framewise phoneme classification with bidirectional LSTM and other neural network architectures,” *Neural Networks*, vol. 18, no. 5–6, pp. 602–610, 2005, doi: 10.1016/j.neunet.2005.06.042.

[51] M. Rhanoui, M. Mikram, S. Yousfi, and S. Barzali, “A CNN-BiLSTM Model for Document-Level Sentiment Analysis,” *Mach. Learn. Knowl. Extr.*, vol. 1, no. 3, pp. 832–847, 2019, doi: 10.3390/make1030048.

[52] D. Scherer, A. Müller, and S. Behnke, “Evaluation of pooling operations in convolutional architectures for object recognition,” *Lect. Notes Comput. Sci. (including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics)*, vol. 6354 LNCS, no. PART 3, pp. 92–101, 2010, doi: 10.1007/978-3-642-15825-4_10.

[53] D. P. Kingma and J. L. Ba, “Adam: A method for stochastic optimization,” *3rd Int. Conf. Learn. Represent. ICLR 2015 - Conf. Track Proc.*, pp. 1–15, 2015.

[54] Y. Kim, “Convolutional neural networks for sentence classification,” *EMNLP 2014 - 2014 Conf. Empir. Methods Nat. Lang. Process. Proc. Conf.*, pp. 1746–1751, 2014, doi: 10.3115/v1/d14-1181.

[55] S. Hochreiter and J. Schmidhuber, “Long Short-Term Memory,” *Neural Comput.*, vol. 9, no. 8, pp. 1735–1780, 1997, doi: 10.1162/neco.1997.9.8.1735.

[56] X. Wang, W. Jiang, and Z. Luo, “Combination of convolutional and recurrent neural network for sentiment analysis of short texts,” *COLING 2016 - 26th Int. Conf. Comput. Linguist. Proc. COLING 2016 Tech. Pap.*, pp. 2428–2437, 2016.

[57] S. Ioffe, “Batch Renormalization: Towards reducing minibatch dependence in batch-normalized models,” *Adv. Neural Inf. Process. Syst.*, vol. 2017-Decem, no. Nips, pp. 1946–1954, 2017.

[58] H. Choi, K. Cho, and Y. Bengio, “Fine-grained attention mechanism for neural machine translation,” *Neurocomputing*, vol. 284, pp. 171–176, 2018, doi: 10.1016/j.neucom.2018.01.007.

[59] Y. Diao *et al.*, “CRHASum: extractive text summarization with contextualized-representation hierarchical-attention summarization network,” *Neural Comput. Appl.*, vol. 32, no. 15, pp. 11491–11503, 2020, doi: 10.1007/s00521-019-04638-3.

[60] A. Vaswani *et al.*, “Attention is all you need,” *Adv. Neural Inf. Process. Syst.*, vol. 2017-Decem, no. Nips, pp. 5999–6009, 2017.

[61] R. Kadlec, M. Schmid, O. Bajgar, and J. Kleindienst, “Text understanding with the attention sum reader network,” *54th Annu. Meet. Assoc. Comput. Linguist. ACL 2016 - Long Pap.*, vol. 2, pp. 908–918, 2016, doi: 10.18653/v1/p16-1086.

[62] J. Salazar, K. Kirchhoff, and Z. Huang, “Self-attention Networks for Connectionist Temporal Classification in Speech Recognition,” *ICASSP, IEEE Int. Conf. Acoust. Speech Signal Process. - Proc.*, vol. 2019-May, pp. 7115–7119, 2019, doi: 10.1109/ICASSP.2019.8682539.

[63] F. Guzmán *et al.*, “The Flores evaluation datasets for low-resource machine translation: Nepali-English and Sinhala-English,” *EMNLP-IJCNLP 2019 - 2019 Conf. Empir. Methods Nat. Lang. Process. 9th Int. Jt. Conf. Nat. Lang. Process. Proc. Conf.*, pp. 6098–6111, 2020, doi: 10.18653/v1/d19-1632.

[64] D. Wild, A. Grove, S. Eremenco, S. McElroy, A. Verjee-Lorenz, and P. Erikson, “Wild2005_Value in Health 8(2)_94-104.pdf,” *Value Heal.*, vol. 8, no. 2, pp. 94–104, 2005.

[65] S. McKown *et al.*, “Good practices for the translation, cultural adaptation, and linguistic validation of clinician-reported outcome, observer-reported outcome, and performance outcome measures,” *J. Patient-Reported Outcomes*, vol. 4, no. 1, 2020, doi: 10.1186/s41687-020-00248-z.

[66] A. Balahur and M. Turchi, “Improving sentiment analysis in twitter using multilingual machine translated data,” *Int. Conf. Recent Adv. Nat. Lang. Process. RANLP*, no. September, pp. 49–55, 2013.

[67] A. Paul and B. S. Purkayastha, “English to Nepali Statistical Machine Translation System,” *Lect. Notes Networks Syst.*, vol. 24, pp. 423–431, 2018, doi: 10.1007/978-981-10-6890-4_41.

[68] S. Strassel and J. Tracey, “LORELEI language packs: Data, tools, and resources for technology development in low resource languages,” *Proc. 10th Int. Conf. Lang. Resour. Eval. Lr. 2016*, pp. 3273–3280, 2016.

[69] H. Riza *et al.*, “Introduction of the Asian Language Treebank,” *2016 Conf. Orient. Chapter Int. Comm. Coord. Stand. Speech Databases Assess. Tech. O-COCOSDA 2016*, no. October, pp. 1–6, 2017, doi: 10.1109/ICSDA.2016.7918974.

[70] A. Riccardi, “Translation Studies: Perspectives on an Emerging Discipline,” *South. African Linguist. Appl. Lang. Stud.*, vol. 24, no. 1, pp. 129–132, 2006, doi: 10.2989/16073610609486411.

[71] B. C. Myers-scotton, “SIL Electronic Book Reviews 2006-006 Contact Linguistics : Bilingual encounters and grammatical outcomes,” *Oxford Univ. Press. 2002. Pp. 356. Pap.*, 2006, doi: 10.1093/acprof:oso/9780198299530.001.0001.

[72] D. Gurung, “Nepali-English code-switching in the conversations of Nepalese people Nepali-English Code-switching in the Conversations of Nepalese People: A Sociolinguistic Study,” 2018, [Online]. Available: https://pure.roehampton.ac.uk/ws/portalfiles/portal/1284973/.

[73] P. Trudgill, “A Glossary of Sociolinguistics,” *Edinburgh Univ. Press*, pp. 234–235, 2003, doi: 10.1590/S0102-44502003000100014.

[74] R. K. Dahal, “Language Politics in Nepal,” *J. Polit. Sci.*, vol. 1, no. 1, 1998, doi: https://doi.org/10.3126/jps.v1i1.1685.

[75] C. Genetti, *How languages work: An introduction to language and linguistics*. Cambridge University Press and Assessment, 2014.

[76] L. Wei and M. G. Moyer, “The Blackwell Guide to Research Methods in Bilingualism and

Multilingualism," *Blackwell Publ. Ltd*, pp. 1–403, 2009, doi: 10.1002/9781444301120.