Universidad Autónoma de Querétaro

Facultad de Ingeniería

UNIVERSIDAD AUTÓNOMA DE QUERÉTARO

December 1, 2023

Author:*Lic. Ijtsi Dzaya Ramos Morales*

Machine Learning, Dr. Marco Antonio Aceves Fernández

Final Exam: Classification of Dementia using text and audio features

Abstract

Dementia, a progressive cognitive syndrome impacting the elderly, lacks effective treatment, emphasizing the importance of early detection [1]. This study investigates a non-invasive approach, utilizing a classifier trained on acoustic and text features extracted through transformer-based methods and time series information transformation [2]. Models including K-Nearest Neighbors (KNN), Random Forest, Linear Support Vector Machines (SVM), Polynomial SVM, Precomputed SVM, and Radial Basis Function (RBF) SVM are evaluated. Notably, acoustic features exhibit minimal correlation with dementia classification, resulting in lower accuracy. Conversely, text features demonstrate promising results, warranting further investigation. This study contributes valuable insights to dementia detection methodologies, highlighting the nuanced performance of distinct feature modalities.

Introduction

The escalating prevalence of Dementia necessitates advanced screening methods for efficient detection. In response to this imperative, computational approaches have emerged. Notably, natural language processing (NLP) applied to patient conversations and spoken medical tests offers a non-invasive means of detecting early symptoms of dementia, exploiting the observable deterioration in communication skills as the disease progresses [**3**]. Previous studies have identified linguistic features such as telegraphic speech, repetitiveness, and misspelling as prognostic indicators of dementia [**4**] [**5**]. Moreover, tasks like the Cookie Theft picture description test have proven effective in evaluating speech impairments associated with dementia [**6**].

While some studies emphasize image-based approaches, employing convolutional neural networks (CNN) and deep learning architectures like VGG16, ResNet-152, and DenseNet-121 [**7**], others explore linguistic and acoustic features [**8**][**9**]. The text-based classifier developed by Yamanki et al. achieved promising accuracies [**8**], whereas Fraser et al. emphasized the importance of feature selection in high-dimensional spaces, yielding an accuracy of 81% when focusing on semantic, acoustic, syntactic, and information impairment features [**10**]. This underscores the need for a nuanced exploration of both acoustic and linguistic features for robust dementia classification.

In this context, this study delves into the binary classification of dementia, focusing on feature extraction through advanced methods such as transformer-based approaches for text and time series information transformation for acoustics. Leveraging a diverse set of classifiers, including KNN, Random Forest, and various SVM variants, the results obtained highlight unexpected disparities in the contribution of acoustic and text features to dementia classification.

Theoretical Foundation

Dementia

Dementia is a progressive cerebral disorder marked by the deterioration of cognitive functions, accompanied by manifestations such as apathy, social behavior decline, aggressiveness, delusions, and hallucinations [**11**]. Various screening methods aim to identify early signs, including spontaneous speech tasks that analyze language skills and speech patterns [**12**].

Dementia Screening

Conversation/Interview Speech This screening involves extracting features from casual speech, focusing on natural language and biological aspects to identify early signs of Alzheimer's disease (AD) [**12**].

Picture Description In picture description tasks, subjects orally narrate a story based on a sequence of images within a restricted time. The renowned Cookie Theft picture test is a prominent example, requiring patients to describe a kitchen scene involving individuals and activities [**13**].

Artificial Intelligence

Natural Language Processing

Natural Language Processing (NLP), situated at the intersection of Artificial Intelligence and Linguistics, has evolved from text information retrieval algorithms to extract semantics by discerning relationships within text [**14**]. NLP encompasses two subcategories: Natural Language Understanding (NLU) and Natural Language Generation (NLG). NLU focuses on reading and understanding text, incorporating phonology, morphology, lexical, and semantic elements [**14**, **15**].

Recent advancements in NLP involve Convolutional Neural Networks (CNNs) and Recurrent Neural Networks (RNNs), with attention mechanisms and transformer models playing crucial roles in enhancing performance [**16**, **17**, **18**].

Acoustic and Linguistic Features

Acoustic Features

Acoustic features, extracted from audio segments, play a vital role in training classifiers. Time-dependent features include Mel-frequency cepstral coefficient (MFCC), Gammatone cepstral coefficient (GTCC), delta-MFCC, delta-GTCC, Log-Energy, Formants, and Fundamental frequency. Time-independent features comprise Jitter, Shimmer, and Pitch [**19**].

Linguistic Features

Linguistic features can be categorized into lexical, syntactic, and semantic aspects. Lexical features involve word truncation, vowel repetition, vocabulary length, and lexical diversity. Syntactic features identify grammatical and sentence structural errors, while semantic features focus on extracting the main ideas of the text through keyword searches and similarity calculations [**20**].

Support Vector Machines (SVM)

Support Vector Machines (SVM) is a supervised machine learning algorithm used for classification and regression tasks. In the context of dementia classification based on acoustic and text features, SVM aims to find a hyperplane that best separates the data into different classes. It works by mapping the input data into a high-dimensional space and finding the optimal hyperplane that maximizes the margin between different classes. SVM is effective in handling high-dimensional data and is known for its ability to generalize well to unseen data [**21**].

The decision function for a linear SVM can be expressed as:

$$f(\mathbf{x}) = \mathbf{w} \cdot \mathbf{x} + b \tag{1}$$

where $\mathbf{w}$ is the weight vector, $\mathbf{x}$ is the input feature vector, and b is the bias term.

$$y(\mathbf{x}) = \text{sign}\left(f(\mathbf{x})\right) \tag{2}$$

The optimization problem for SVM involves minimizing:

$$\frac{1}{2}\|\mathbf{w}\|^2 + C\sum_{i=1}^{N}\xi_i \tag{3}$$

subject to constraints $y_i(\mathbf{w} \cdot \mathbf{x}_i + b) \geq 1 - \xi_i$ and $\xi_i \geq 0$, where C is the regularization parameter and ξ_i are slack variables.

Random Forest

Random Forest is an ensemble learning algorithm that combines multiple decision trees to improve the overall performance and robustness of the model. In the context of dementia classification, Random Forest can handle both text and acoustic features effectively. It works by training multiple decision trees on different subsets of the data and combining their predictions through a voting mechanism. Random Forest is known for its ability to handle noisy and high-dimensional data, making it suitable for complex classification tasks [**22**].

The decision function for a Random Forest can be expressed as:

$$y(\mathbf{x}) = \text{mode}\left(\text{tree}_1(\mathbf{x}), \text{tree}_2(\mathbf{x}), \ldots, \text{tree}_N(\mathbf{x})\right) \tag{4}$$

where mode is the function that returns the most frequent class prediction among the decision trees.

k-Nearest Neighbors (kNN)

k-Nearest Neighbors (kNN) is a simple and intuitive classification algorithm based on the idea that similar data points should belong to the same class. In the context of dementia classification, kNN works by measuring the distance between a test data point and its k-nearest neighbors in the feature space. The majority class among the k-nearest neighbors determines the class of the test point. kNN is non-parametric and does not assume any underlying distribution of the data, making it versatile for different types of features [**23**].

The decision function for kNN can be expressed as:

$$y(\mathbf{x}) = \text{majority}\left(\text{class}(\mathbf{N}_1), \text{class}(\mathbf{N}_2), \ldots, \text{class}(\mathbf{N}_k)\right) \tag{5}$$

where majority returns the most frequent class among the k-nearest neighbors.

Materials and Methods

1. **Data Set:** The dataset utilized in this study was sourced from the Pitt Corpus from DementiaBank, a collection of longitudinal neuropsychological assessments gathered between 1983 and 1988 for the Alzheimer Research Program at the University of Pittsburgh. The dataset includes 104 control subjects, 208 subjects with

diagnosed dementia, and 85 subjects with an unknown diagnosis. The data is available at the following link: `https://sla.talkbank.org/TBB/dementia/English/Pitt`.

2. **Software:** The analysis was conducted using Python, a versatile programming language. The following libraries were employed for numerical computations, data manipulation, and graphical analysis: `numpy`, `pandas`, `matplotlib`, `sklearn`, `seaborn`, `random`, and `torch`.

3. **Methods:**

   - **Preprocessing:** The transcriptions underwent preprocessing using regular expressions to clean the text. The audio data was cleaned by removing noise through spectral subtraction. Additionally, acoustic features were transformed into tabular data to facilitate model training.

   - **Text Feature Extraction:** To obtain text features from transcriptions, a `Sberbank Transformer` was employed. This transformer is designed to capture meaningful features from textual data.

   - **Acoustic Feature Extraction:** For extracting acoustic features from audio recordings, the `librosa` library was utilized. This library enables the extraction of various acoustic features from audio signals.

   - **Metrics:** The evaluation metrics used for assessing model performance included accuracy, precision, recall, and F1 score. These metrics provide a comprehensive understanding of the model's classification effectiveness.

4. **Methodology Visualization:**

   Figure 2 illustrates the methodology for text classification, showcasing the steps involved in processing transcriptions and extracting text features.

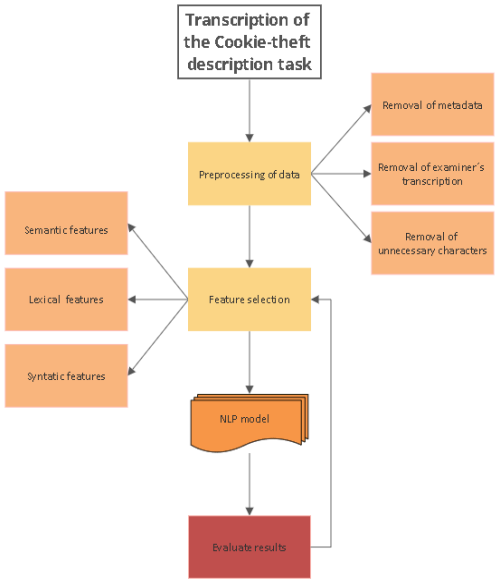

Figure 1: Methodology for Text Classification

Figure 3 presents a diagram of the methodology for audio classification, depicting the stages from audio data input to the extraction of acoustic features.

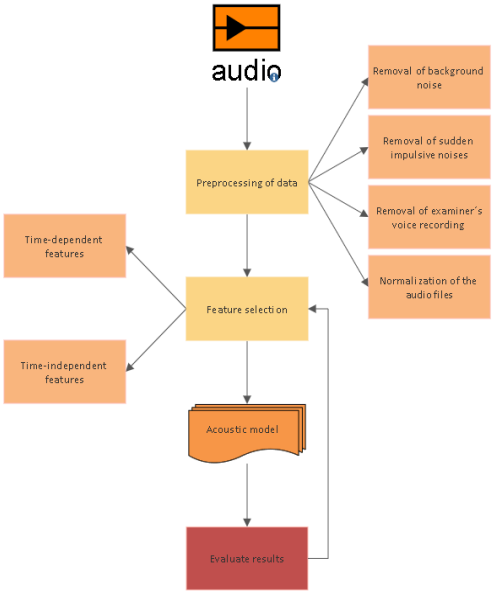

Figure 2: Methodology for Audio Classification

Results

Figure 3 depicts the results of the audio classification, providing a comparison of all the methods used. The results include all metrics except the f1-score, which can be observed in table 1.

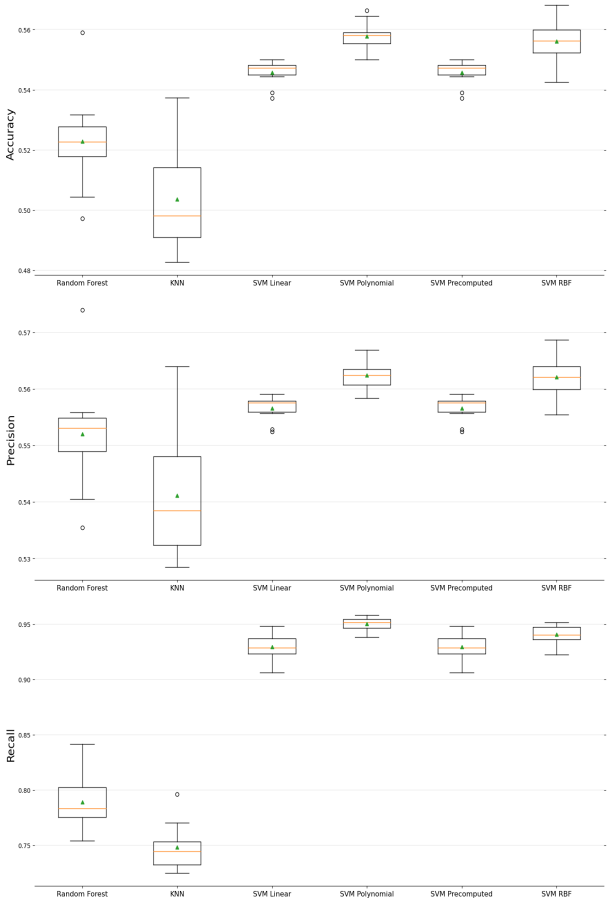

Figure 3: Results for Audio Classification

For a more concise presentation, Figure 4 summarizes the results. In the table, the indexes 0, 1, 2, 3, 4, and 5 correspond to the results of Random Forest, K-Nearest Neighbors (KNN), Linear Support Vector Machines (SVM), Polynomial SVM, Precomputed SVM, and Radial Basis Function (RBF) SVM, respectively.

	Accuracy	Precision	Recall	F1 Score
0	0.517370	0.548666	0.783763	0.644531
1	0.510097	0.546592	0.744624	0.629314
2	0.548149	0.557949	0.932151	0.697830
3	0.549968	0.558362	0.938387	0.699877
4	0.548149	0.557949	0.932151	0.697830
5	0.557175	0.562672	0.945054	0.705175

Figure 4: Results for Audio Classification

Figure 5 depicts the results of the text classification, providing a comparison of all the methods used. The results include all metrics except the f1-score, which can be observed in table 2.

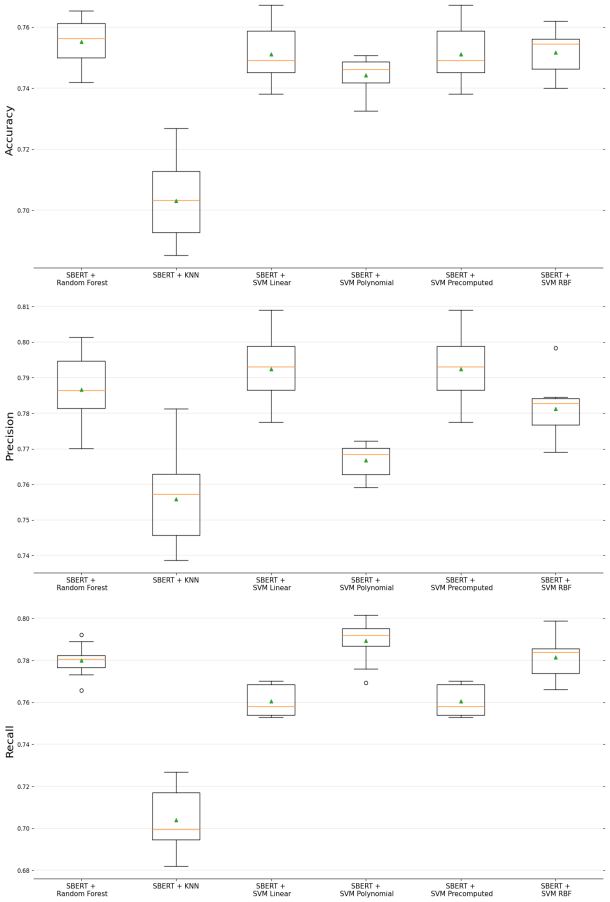

Figure 5: Results for Text Classification

Figure 6 summarizes the results. In the table, the indexes 0, 1, 2, 3, 4, and 5 correspond to the results of Random Forest, K-Nearest Neighbors (KNN), Linear Support Vector Machines (SVM), Polynomial SVM, Precomputed SVM, and Radial Basis Function (RBF) SVM, respectively.

	Accuracy	Precision	Recall	F1 Score
0	0.747306	0.770024	0.789032	0.778073
1	0.686734	0.738670	0.691935	0.712134
2	0.760067	0.799727	0.769462	0.782221
3	0.749091	0.770638	0.795269	0.781088
4	0.760067	0.799727	0.769462	0.782221
5	0.754579	0.782759	0.785699	0.783230

Figure 6: Results for Text Classification

Discussion

For audio classification, the obtained results showed an accuracy of around 50% for all models, which is not significantly better than a random classifier. This suggests that the models might be predicting ran-

domly. However, it's important to note the high values of recall. This could indicate that the model is primarily learning to predict one class and may be biased. The unbalanced nature of the dataset might contribute to this behavior. Data augmentation, a common technique to address class imbalance, was considered but ultimately discarded due to the complicated nature of the dataset. The results underscore that tabular acoustic features alone might not be optimal for dementia classification, aligning with the current state of the art.
On the contrary, the results for text classification were promising and comparable to those reported in the state of the art. The accuracy ranged from 70% to 80% for all models on average, suggesting that the text features obtained are valuable for classifying dementia. This indicates that the information extracted from transcriptions, likely capturing linguistic patterns and deficits, serves as a useful basis for distinguishing between subjects with and without dementia.
Further investigations could explore the combination of both text and acoustic features to enhance the classification performance. Additionally, refining data augmentation techniques or exploring other approaches to handle the class imbalance in the audio dataset may improve model performance.

Conclusion

The utilization of natural language processing (NLP) and acoustic features for dementia classification was explored for the task of Classification of Dementia. The Pitt Corpus from DementiaBank, a collection of longitudinal neuropsychological assessments, provided the basis for the analysis of both text and audio data. For audio classification, challenges were encountered in achieving high accuracy, potentially influenced by dataset imbalance and the intrinsic complexity of acoustic features. The models exhibited a tendency for biased predictions, and limitations in employing data augmentation underscored the intricacies of the audio dataset.
On the contrary, promising outcomes were observed in text classification, with accuracy ranging from 70% to 80%, aligning with state-of-the-art results. The success in text classification suggests that linguistic features extracted from transcriptions, using an SBERN transformer, play a significant role in effectively distinguishing between subjects with and without dementia.
Future work could involve the integration of both text and acoustic features to capitalize on the complementary information from these modalities. Additionally, exploring advanced techniques for handling class imbalance in the audio dataset and further refining preprocessing steps may enhance overall model performance. The findings underscore the potential of linguistic patterns in text as a valuable resource for distinguishing between cognitive states, while also acknowledging the complexities associated with acoustic features in the audio domain.

References

[1] C. C. Aggarwal and C. K. Reddy, *Data clustering: Algorithms and applications*. Chapman and Hall/CRC, 2014. (Book)

[2] P. Berkhin, "A survey of clustering data mining techniques," in *Grouping multidimensional data*, pp. 25-71, Springer, 2006. (Book Chapter)

[3] D. S. Rosas, S. T. Arriaga, and M. A. A. Fernandez, "Search for Dementia Patterns in Transcribed Conversations using Natural Language Processing," in 2019 16th International Conference on Electrical Engineering, Computing Science and Automatic Control (CCE), Mexico City, Mexico, Sep. 2019, pp. 1–6. doi: 10.1109/ICEEE.2019.8884572.

[4] E. Vuorinen, M. Laine, and J. Rinne, "Common Pattern of Language Impairment in Vascular Dementia and in Alzheimer Disease:," Alzheimer Disease and Associated Disorders, vol. 14, no. 2, pp. 81–86, Apr. 2000, doi: 10.1097/00002093-200004000-00005.

[5] Y. C. Wong, H. C. Kwan, W. A. MacKay, and J. T. Murphy, "Spatial organization of precentral cortex in awake primates. I. Somatosensory inputs," J Neurophysiol, vol. 41, no. 5, pp. 1107–1119, Sep. 1978, doi: 10.1152/jn.1978.41.5.1107.

[6] S. Boada-Rovira, M. Pérez-García, R. Marqués-Recuerda, and R. Arnedo-Montoro, "Comparison of Two Types of Picture Description Tasks for Dementia Detection: A Case-Control Study," Journal of Alzheimer's Disease, vol. 58, no. 3, pp. 731-740, 2017. DOI: 10.3233/JAD-170061.

[7] J. Chen, J. Zhu, and J. Ye, "An Attention-Based Hybrid Network for Automatic Detection of Alzheimer's Disease from Narrative Speech," in Interspeech 2019, Sep. 2019, pp. 4085–4089. doi: 10.21437/Interspeech.2019-2872.

[8] Y. Santander-Cruz, S. Salazar-Colores, W. J. Paredes-García, H. Guendulain-Arenas, and S. Tovar-Arriaga, "Semantic feature extraction using SBERT for dementia detection," Brain Sciences, vol. 12, no. 2, p. 270, 2022.

[9] R. Chakraborty, M. Pandharipande, C. Bhat, and S. K. Kopparapu, "Identification of Dementia Using Audio Biomarkers." arXiv, Feb. 27, 2020. Accessed: Dec. 07, 2022. [Online]. Available: http://arxiv.org/abs/2002.12788.

[10] K. C. Fraser, J. A. Meltzer, and F. Rudzicz, "Linguistic Features Identify Alzheimer's Disease in Narrative Speech," JAD, vol. 49, no. 2, pp. 407–422, Nov. 2015, doi: 10.3233/JAD-150520.

[11] M. Maj and N. Sartorius, *Dementia*, 2nd ed. Chichester; Hoboken, NJ: Wiley, 2002.

[12] J. Casarella, *Types of Dementia Explained*. WebMD. Available online: https://www.webmd.com/alzheimers/guide/alzheimersdementia (accessed on 20 September 2022).

[13] L. Cummings, "Describing the Cookie Theft picture: Sources of breakdown in Alzheimer's dementia," PS, vol. 10, no. 2, pp. 153–176, Jul. 2019, doi: 10.1075/ps.17011.cum.

[14] P. M. Nadkarni, L. Ohno-Machado, and W. W. Chapman, "Natural language processing: an introduction," J Am Med Inform Assoc, vol. 18, no. 5, pp. 544–551, Sep. 2011, doi: 10.1136/amiajnl-2011-000464.

[15] P. M. Nadkarni, L. Ohno-Machado, and W. W. Chapman, "Natural language processing: an introduction," J Am Med Inform Assoc, vol. 18, no. 5, pp. 544–551, Sep. 2011, doi: 10.1136/amiajnl-2011-000464.

[16] W. Wang and J. Gang, "Application of Convolutional Neural Network in Natural Language Processing," in 2018 International Conference on Information Systems and Computer Aided Education (ICISCAE), Changchun, China, Jul. 2018, pp. 64–70. doi: 10.1109/ICISCAE.2018.8666928.

[17] K. Greff, R. K. Srivastava, J. Koutnik, B. R. Steunebrink, and J. Schmidhuber, "LSTM: A Search Space Odyssey," IEEE Trans. Neural Netw. Learning Syst., vol. 28, no. 10, pp. 2222–2232, Oct. 2017, doi: 10.1109/TNNLS.2016.2582924.

[18] D. Bahdanau, K. Cho, and Y. Bengio, "Neural Machine Translation by Jointly Learning to Align and Translate." arXiv, May 19, 2016. Accessed: Jan. 20, 2023. [Online]. Available: http://arxiv.org/abs/1409.0473.

[19] M. R. Kumar, S. Vekkot, S. Lalitha, D. Gupta, V. J. Govindraj, K. Shaukat, Y. A. Alotaibi, and M. Zakariah, "Dementia detection from speech using machine learning and Deep Learning Architectures," Sensors, vol. 22, no. 23, p. 9311, 2022.

[20] A. H. Alkenani, Y. Li, Y. Xu, and Q. Zhang, "Predicting Prodromal Dementia Using Linguistic Patterns and Deficits," IEEE Access, vol. 8, pp. 193856–193873, 2020, doi: 10.1109/ACCESS.2020.3029907.

[21] C. Cortes and V. Vapnik, "Support-vector networks," Machine Learning, vol. 20, no. 3, pp. 273–297, Sep. 1995, doi: 10.1007/BF00994018.

[22] L. Breiman, "Random forests," Machine Learning, vol. 45, no. 1, pp. 5–32, Oct. 2001, doi: 10.1023/A:1010933404324.

[23] T. M. Cover and P. E. Hart, "Nearest neighbor pattern classification," IEEE Transactions on Information Theory, vol. 13, no. 1, pp. 21–27, Jan. 1967, doi: 10.1109/TIT.1967.1053964.